Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Areeb Ahmad
Roll Number: 180135
Date: November 15, 2022

QUESTION 1

We are approximating $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ with Monte-Carlo's approximation $\hat{f} = \frac{1}{S}\sum_{s=1}^{S} f(z^{(s)})$, where S samples are $z^{(s)} \sim p(\boldsymbol{z})$ as iid.

Proof1 :

$$\mathbb{E}[f] = \mathbb{E}\left[\hat{f}\right] \tag{1}$$

$$\begin{aligned}
\mathbb{E}\left[\hat{f}\right] &= \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z^{(s)})\right] \\
&= \frac{1}{S}\mathbb{E}\left[\sum_{s=1}^{S} f(z^{(s)})\right], \textit{by Linearity property} \\
&= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}[f] \\
&= \mathbb{E}[f]
\end{aligned} \tag{2}$$

Proof2 :

$$var[\hat{f}] = \frac{1}{S}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right] \tag{3}$$

$$\begin{aligned}
var[\hat{f}] &= \mathbb{E}\left[\hat{f}^2\right] - \mathbb{E}\left[\hat{f}\right]^2 \\
&= \mathbb{E}\left[\left(\frac{1}{S}\sum_{s=1}^{S} f(z^{(s)})\right)^2\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z^{(s)})\right]^2 \\
&= \mathbb{E}\left[\frac{1}{S^2}\sum_{s=1}^{S} f^2(z^{(s)}) + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S} f(z^{(s)})f(z^{(k)})\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z^{(s)})\right]^2 \\
&= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}\left[f^2(z^{(s)})\right] + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S}\mathbb{E}\left[f(z^{(s)})f(z^{(k)})\right] - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z)\right]^2 \\
&= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}\left[f^2(\boldsymbol{z})\right] + \frac{1}{S^2}\sum_{s=1}^{S}\sum_{k=1}^{S}\mathbb{E}[f(\boldsymbol{z})]^2 - \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z)\right]^2 \\
&= \frac{1}{S^2}S\mathbb{E}\left[f^2(\boldsymbol{z})\right] + \frac{1}{S^2}S(S-1)\mathbb{E}[f(\boldsymbol{z})]^2 - \mathbb{E}[f(\boldsymbol{z})]^2 \\
&= \frac{1}{S}\mathbb{E}\left[f^2(\boldsymbol{z})\right] - \frac{1}{S}\mathbb{E}[f(\boldsymbol{z})]^2 \\
&= \frac{1}{S}var[f]
\end{aligned} \tag{4}$$

hence,

$$var[\hat{f}] = \frac{1}{S}(\mathbb{E}\left[f^2\right] - \mathbb{E}\left[f\right]^2) \tag{5}$$

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Areeb Ahmad
Roll Number: 180135
Date: November 15, 2022

QUESTION 2

My solution to problem 2
The joint distribution is given by:

$$
\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{w}, \alpha, \beta | \boldsymbol{X}) &= p(\boldsymbol{y}|\boldsymbol{w}, \beta, \boldsymbol{X}) p(\boldsymbol{w}|\alpha) p(\alpha) p(\beta) \\
&= \prod_{n=1}^{N} p(y_n|\boldsymbol{w}, \beta, x_n) p(\boldsymbol{w}|\alpha) \prod_{d=1}^{D} p(\alpha_d) p(\beta) \\
&= \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^T x_n, \beta^{-1}) \mathcal{N}(\boldsymbol{w}|0, diag(\alpha_1^{-1} ... \alpha_D^{-1})) \prod_{d=1}^{D} Gamma(\alpha_d|e_0, f_0) Gamma(\beta|a_0, b_0) \\
\log p(\boldsymbol{y}, \boldsymbol{w}, \alpha, \beta | \boldsymbol{X}) &= \sum_{n=1}^{N} \log p(y_n|\boldsymbol{w}, \beta, x_n) + \log p(\boldsymbol{w}|\alpha) + \sum_{d=1}^{D} \log p(\alpha_d) + \log p(\beta) \\
&= \sum_{n=1}^{N} \log \left[\sqrt{\frac{\beta}{2\pi}} exp\left(\frac{-\beta}{2} (y_n - \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x_n})^2 \right) \right] + \log \left[\sqrt{\frac{\prod_{d=1}^{D} \alpha_d}{(2\pi)^D}} exp\left(\frac{-\boldsymbol{w}^T diag(\alpha) \boldsymbol{w}}{2} \right) \right] \\
&+ \sum_{d=1}^{D} \log \left[\frac{f_0^{e_0}}{\Gamma(e_0)} \alpha_d^{e_0 - 1} exp(-f_0 \alpha_d) \right] + \log \left[\frac{b_0^{a_0}}{\Gamma(a_0)} \beta^{a_0 - 1} exp(-b_0 \beta) \right] \\
&= -\frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^T x_n)^2 + \frac{N}{2} \log \beta - \frac{1}{2} \boldsymbol{w}^T diag(\alpha) \boldsymbol{w} + \sum_{d=1}^{D} \left((\frac{1}{2} + e_0 - 1) \log \alpha_d - f_0 \alpha_d \right) \\
&+ (a_0 - 1) \log \beta - b_0 \beta + constant
\end{aligned}
\tag{6}
$$

Mean field VI updates:

- For β keeping other fixed

$$
\begin{aligned}
\log q_\beta^*(\beta) &= E_{\boldsymbol{w}, \alpha}[\frac{N}{2} \log \beta - \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^T x_n)^2 + (a_0 - 1) \log \beta - b_0 \beta] \\
&= \left(\frac{N}{2} + a_0 - 1 \right) \log \beta - \beta \left[\frac{1}{2} \sum_{n=1}^{N} \mathbb{E}\left[(y_n - \boldsymbol{w}^T x_n)^2 \right] + b_0 \right] \qquad (7) \\
\beta &\sim Gamma\left(\frac{N}{2} + a_0, \frac{1}{2} \sum_{n=1}^{N} \mathbb{E}\left[(y_n - \boldsymbol{w}^T x_n)^2 \right] + b_0 \right)
\end{aligned}
$$

- For $\boldsymbol{w}$ keeping others fixed

$$
\begin{aligned}
\log q^*_{\boldsymbol{w}}(\boldsymbol{w}) &= E_{\alpha,\beta}\left[p(\boldsymbol{y},\boldsymbol{w},\alpha,\beta|\boldsymbol{X})\right] \\
&= -\frac{1}{2}\left[\boldsymbol{w^T}\left(\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}x_n x_n^T + diag(\mathbb{E}\left[\alpha_1\right]...\mathbb{E}\left[\alpha_d\right])\right)\boldsymbol{w} - 2\boldsymbol{w^T}\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}y_n x_n\right] \\
\boldsymbol{w} &\sim \mathcal{N}\left(\boldsymbol{U}^{-1}\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}y_n x_n, \boldsymbol{U}^{-1}\right) \\
\boldsymbol{U} &= \left(\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}x_n x_n^T + diag(\mathbb{E}\left[\alpha_1\right]...\mathbb{E}\left[\alpha_d\right])\right)
\end{aligned}
\tag{8}
$$

- For $\alpha_d \forall$ d keeping others fixed

$$
\begin{aligned}
\log q^*_{\alpha_d} &= E_{\boldsymbol{w},\alpha_{-d},\beta}\left[\left(\frac{1}{2}+e_0-1\right)\log\alpha_d - \alpha_d\left(f_0+\frac{\boldsymbol{w}_d^2}{2}\right)\right] \\
&= \left(\frac{1}{2}+e_0-1\right)\log\alpha_d - \alpha_d\left(f_0+\frac{\mathbb{E}\left[\boldsymbol{w}_d^2\right]}{2}\right) \\
\alpha_d &\sim Gamma\left(\frac{1}{2}+e_0, f_0+\frac{\mathbb{E}\left[\boldsymbol{w_d}^2\right]}{2}\right)
\end{aligned}
\tag{9}
$$

We have,

$$
\begin{aligned}
w &\sim \mathcal{N}(\boldsymbol{\mu},\Sigma) \\
\boldsymbol{\mu} &= \Sigma\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}y_n x_n \\
\Sigma &= \left(\mathbb{E}\left[\beta\right]\sum_{n=1}^{N}x_n x_n^T + diag(\mathbb{E}\left[\alpha_1\right]...\mathbb{E}\left[\alpha_d\right])\right)^{-1}
\end{aligned}
\tag{10}
$$

$$
\beta \sim Gamma(a_\beta, b_\beta)
$$

$$
\begin{aligned}
a_\beta &= \frac{a}{2}+a_0 \\
b_\beta &= \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[(y_n-\boldsymbol{w}^T x_n)^2\right]+b_0
\end{aligned}
\tag{11}
$$

$$
\alpha_d \sim Gamma(c_{\alpha_d}, g_{\alpha_d})
$$

$$
\begin{aligned}
c_{\alpha_d} &= \frac{1}{2}+e_0 \\
g_{\alpha_d} &= f_0+\frac{\mathbb{E}\left[\boldsymbol{w_d}^2\right]}{2}
\end{aligned}
\tag{12}
$$

Therefore we have:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{w}\right] &= \boldsymbol{\mu} \\
\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^{\boldsymbol{T}}\right] &= \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}} + \Sigma \\
\mathbb{E}\left[\beta\right] &= \frac{a_\beta}{b_\beta} \\
\mathbb{E}\left[\alpha_d\right] &= \frac{c_{\alpha_d}}{g_{\alpha_d}} \\
\mathbb{E}\left[w_d^2\right] &= \Sigma_{dd} + \mu_d^2
\end{aligned} \tag{13}$$

Mean Field Inference VI algorithm

- Let $c_{\alpha_d} = \frac{1}{2} + e_0$ and $a_\beta = \frac{N}{2} + a_0$ and initialize $c_{\alpha_d} \forall$ d and b_β.
- Compute $\mathbb{E}\left[\alpha_d\right]^{(0)} \forall$ d and $\mathbb{E}\left[\beta\right]^{(0)}$
- For t=1,2...T or until convergence

$$\begin{aligned}
\Sigma^{(t)} &= \left(\mathbb{E}\left[\beta\right]^{(t-1)} \sum_{n=1}^{N} x_n x_n^T + diag(\mathbb{E}\left[\alpha_1\right]^{(t-1)} ... \mathbb{E}\left[\alpha_d\right]^{(t-1)})\right)^{-1} \\
\boldsymbol{\mu}^{(t)} &= \Sigma^{(t)} \mathbb{E}\left[\beta\right]^{(t-1)} \sum_{n=1}^{N} y_n x_n \\
b_\beta^{(t)} &= \frac{1}{2}\mathbb{E}\left[(y_n - \boldsymbol{w}^T x_n)^2\right] + b_0 \\
\mathbb{E}\left[\boldsymbol{w}\right] &= \boldsymbol{\mu}^{(t)} \\
\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^{\boldsymbol{T}}\right] &= \boldsymbol{\mu}^{(\boldsymbol{t})}\boldsymbol{\mu}^{\boldsymbol{T}(\boldsymbol{t})} + \Sigma^{(t)} \\
\mathbb{E}\left[\beta\right]^{(t)} &= \frac{a_\beta}{b_\beta^{(t)}} \\
\mathbb{E}\left[\alpha_d\right] &= \frac{c_{\alpha_d}}{g_{\alpha_d}^{(t)}} \forall d \\
g_{\alpha_d}^{(t)} &= f_0 + \frac{\mathbb{E}\left[\boldsymbol{w_d}^{2(t)}\right]}{2}
\end{aligned} \tag{14}$$

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Areeb Ahmad
Roll Number: 180135
Date: November 15, 2022

QUESTION 3

Conditional posterior of α

$$\begin{aligned} p(\alpha|\mathbf{X}, \boldsymbol{\lambda}, \beta) &\propto p(\boldsymbol{\lambda}|\alpha, \beta)p(\alpha) \\ &\propto \prod_{n=1}^{N} Gamma(\lambda_n|\alpha, \beta) * Gamma(\alpha|a, b) \\ &\propto \left[\frac{\beta^{N\alpha}(\prod_{n=1}^{N}\lambda_n)^{\alpha-1}}{\Gamma(\alpha)^N}\right] * \alpha^{a-1}exp(-\alpha b) \end{aligned} \tag{15}$$

We don't have closed form expression. so we will need a sampling method to draw samples from this distribution.

Conditional posterior of β

$$\begin{aligned} p(\beta|\mathbf{X}, \boldsymbol{\lambda}, \alpha) &\propto p(\boldsymbol{\lambda}|\alpha, \beta)p(\beta) \\ &\propto \prod_{n=1}^{N} Gamma(\lambda_n|\alpha, \beta) * Gamma(\beta|c, d) \\ &\propto \beta^{N\alpha+c-1} * \alpha^{a-1}exp(-\beta(d + \sum_{n=1}^{N}\lambda_n)) \\ \beta &\sim Gamma(N\alpha + c, d + \sum_{n=1}^{N}\lambda_n) \end{aligned} \tag{16}$$

Conditional posterior of λ_n

$$\begin{aligned} p(\lambda_n|\mathbf{X}, \lambda_{-n}, \alpha, \beta) &\propto p(x_n|\lambda_n)p(\lambda_n|\alpha, \beta) \\ &\propto Poisson(x_n|\lambda_n) * Gamma(\lambda_n|\alpha, \beta) \\ &\propto \lambda_n^{\alpha+x_n-1}exp(-(\beta+1)\lambda_n) \\ \lambda_n &\sim Gamma(\alpha + x_n, \beta + 1) \end{aligned} \tag{17}$$

Gibbs Sampling:

- $\alpha^{(0)}$,$\beta^{(0)}$
- for t = 1 . . . T :

$$\begin{aligned} \lambda_n^{(t)} &\sim Gamma(\alpha^{(t-1)} + x_n, \beta^{(t-1)} + 1) \\ \beta^{(t)} &\sim Gamma(N\alpha^{(t-1)} + c, d + \sum_{n=1}^{N}\lambda_n^{(t)}) \\ \alpha^{(t)} &\sim p(\alpha|\mathbf{X}, \boldsymbol{\lambda}^{(t)}, \beta^{(t)}) \;\; Using \;\; Sampling \end{aligned} \tag{18}$$

- Return $(\alpha^{(t)}, \beta^{(t)}, \boldsymbol{\lambda}^{(t)})_{t=0}^{T}$

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Areeb Ahmad
Roll Number: 180135
Date: November 15, 2022

QUESTION 4

- For Mean $r_{ij} = \mathbf{u}_i^T \mathbf{v}_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim \mathcal{N}(0, \beta^{-1})$

$$\begin{aligned} \mathbb{E}\left[r_{ij}\right] &= \mathbb{E}\left[\mathbf{u}_i^T \mathbf{v}_j + \epsilon_{ij}\right] \\ &= \mathbb{E}\left[\mathbf{u}_i^T \mathbf{v}_j\right] + \mathbb{E}\left[\epsilon_{ij}\right] \\ &= \frac{1}{S}\sum_{s=1}^{S} \mathbf{u}_i^{(s)T} \mathbf{v}_j^{(s)} + 0 \end{aligned} \tag{19}$$

- For variance:

$$\begin{aligned} \mathbb{E}\left[r_{ij}^2\right] &= \mathbb{E}\left[(\mathbf{u}_i^T \mathbf{v}_j)^2 + 2\epsilon_{ij}(\mathbf{u}_i^T \mathbf{v}_j) + \epsilon_{ij}^2\right] \\ &= \mathbb{E}\left[(\mathbf{u}_i^T \mathbf{v}_j)^2\right] + \mathbb{E}\left[\epsilon_{ij}^2\right] \\ &= \frac{1}{S}\sum_{s=1}^{S} (\mathbf{u}_i^{(s)T} \mathbf{v}_j^{(s)})^2 + \beta^{-1} \end{aligned} \tag{20}$$

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

$$var[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S} (\mathbf{u}_i^{(s)T} \mathbf{v}_j^{(s)})^2 + \beta^{-1} - \left(\frac{1}{S}\sum_{s=1}^{S} \mathbf{u}_i^{(s)T} \mathbf{v}_j^{(s)}\right)^2$$

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Areeb Ahmad
Roll Number: 180135
Date: November 15, 2022

QUESTION 5

Part I:

- Maximizing over pre-specified range, we will get $M \approx 6.5$
- Acceptance rate ≈ 0.458 .
- We know $Z_p = p(accept) * M$,using $p(accept) = 0.458$, we will get $Z_p \approx 3$.Hence the histogram and plot become similar.

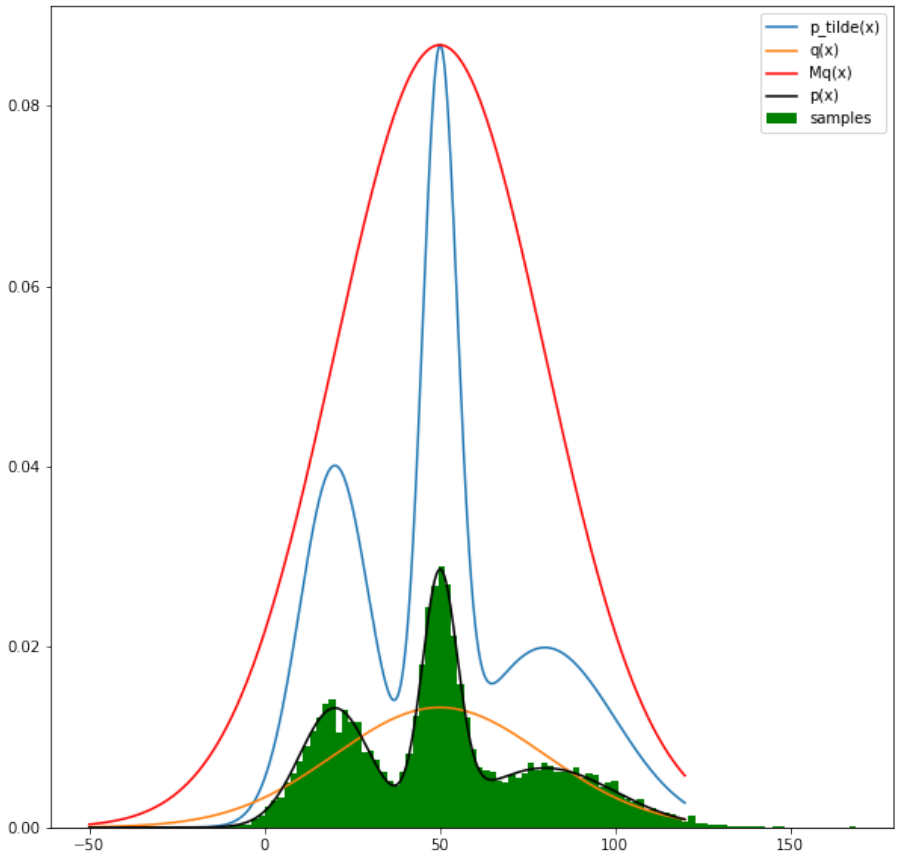
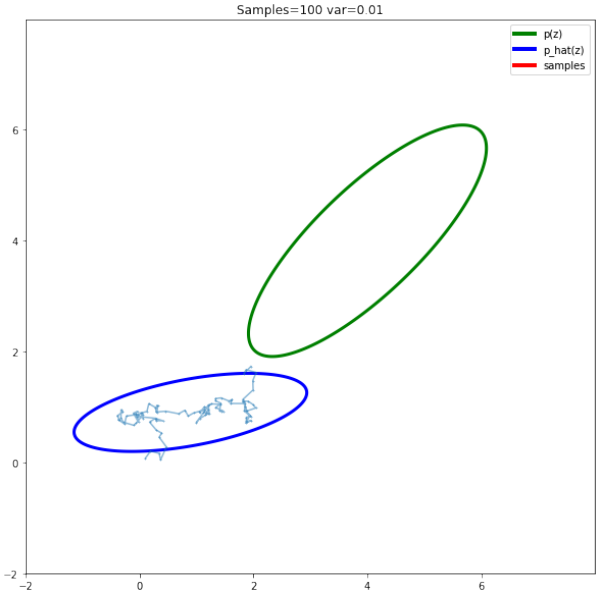


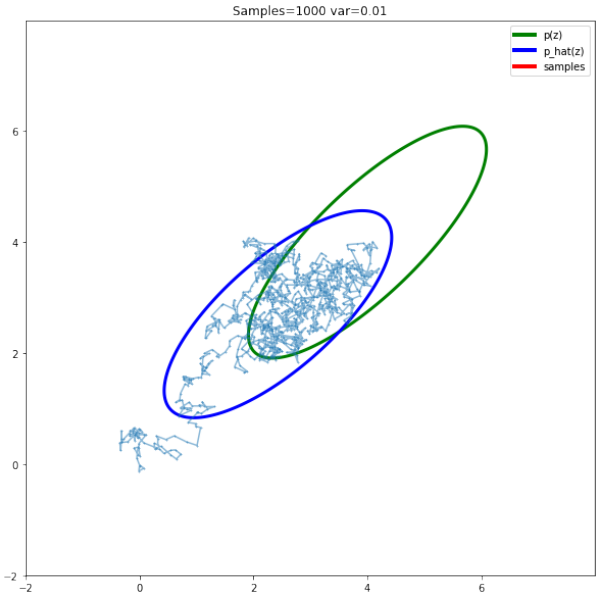
Figure 1: $\tilde{p}(x), q(x), Mq(x), p(x)$and samples

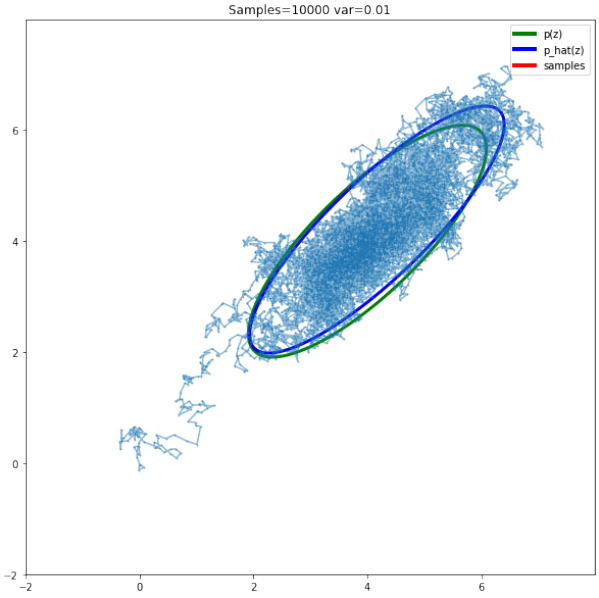
Part II:

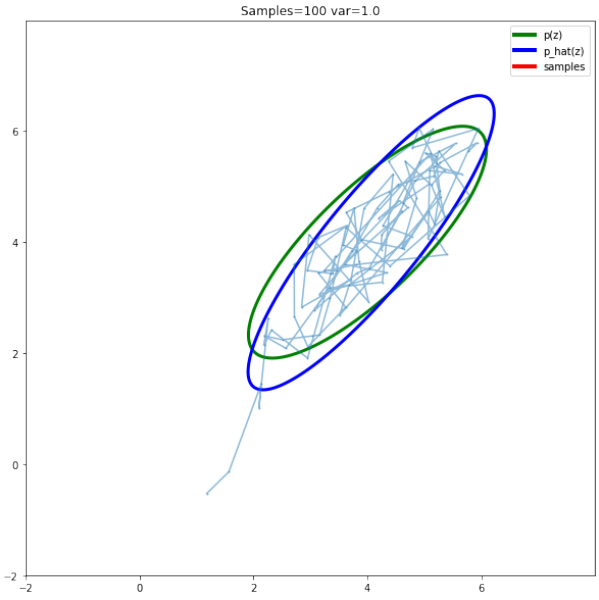
- For $\sigma^2 = 0.01$,Rejecting rate ≈ 0.088
- For $\sigma^2 = 1$,Rejecting rate ≈ 0.600
- For $\sigma^2 = 100$,Rejecting rate ≈ 0.988

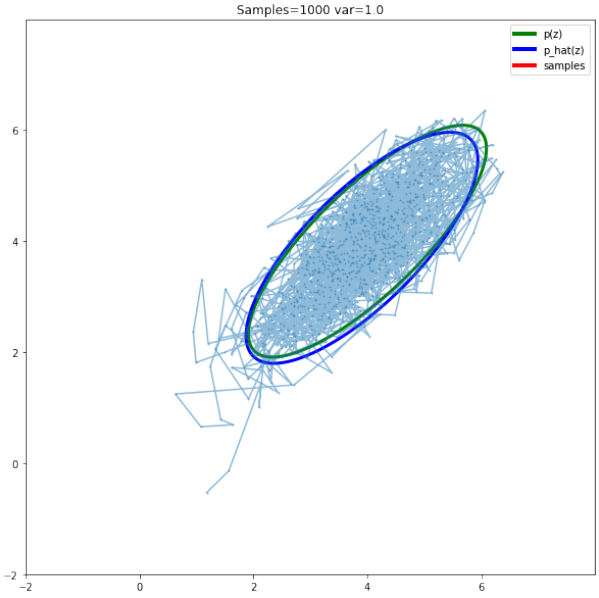
– Given 10k samples we can see from the plots that $\sigma^2 = 1$ works fairly well in terms of convergence. It is able to traverse the entire space relatively fast. We see that for $\sigma^2 = 0.01$ rejection rate is less but the traversal of space is poor. For $\sigma^2 = 100$ traversal of space is good but the convergence rate is very slow. It is for $\sigma^2 = 1$ that traversal is good and the convergence rate is not bad either. Therefore, $\sigma^2 = 1$ is the preferred choice.

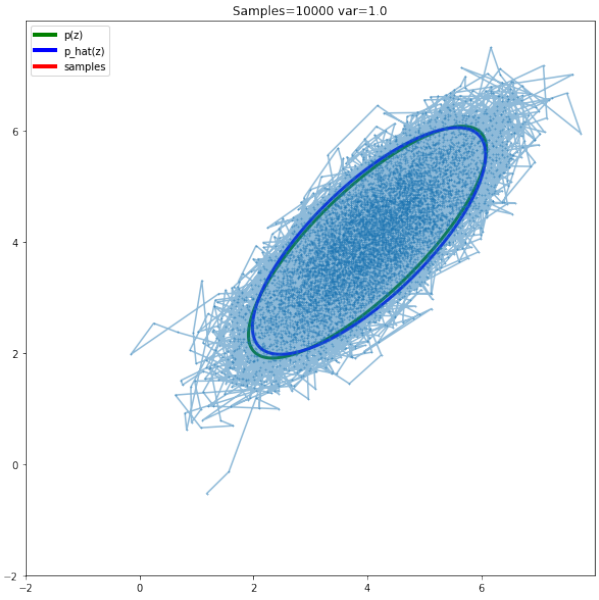


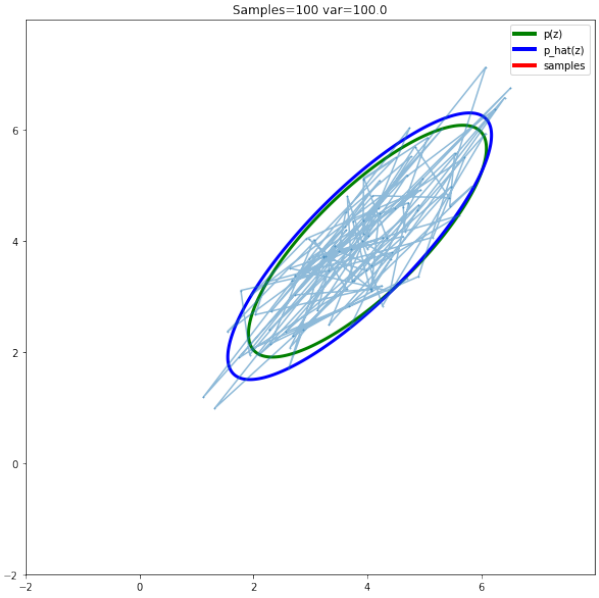


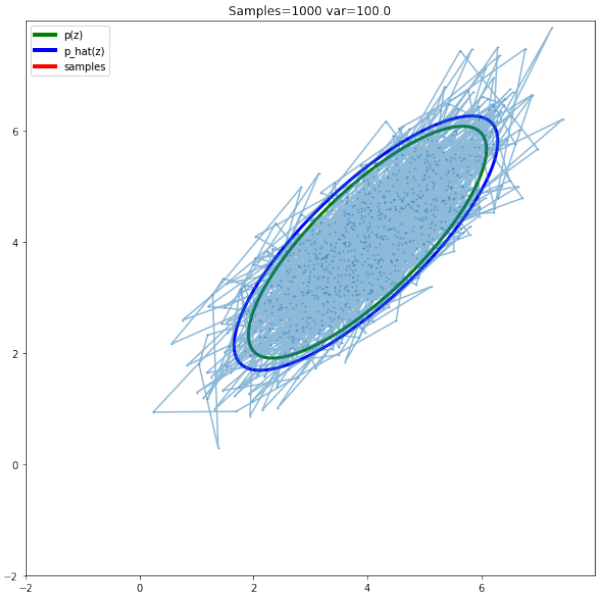

Samples=10000 var=0.01
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6

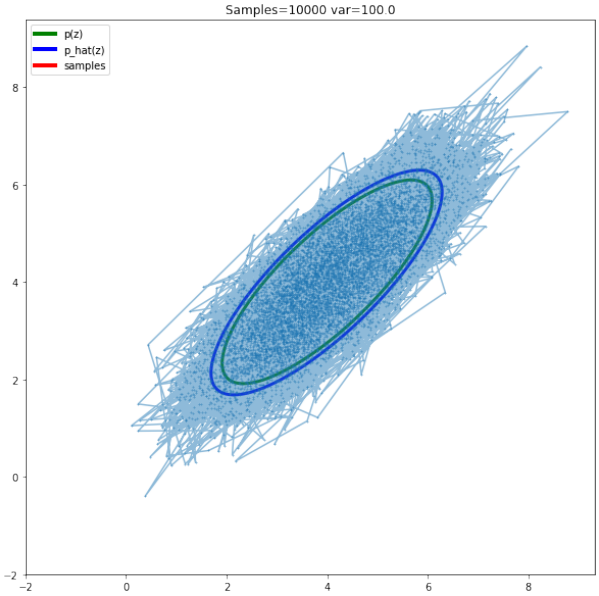

Samples=100 var=1.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6


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
6
4
2
0
−2
−2
0
2
4
6


Samples=100 var=100.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6


Samples=1000 var=100.0
p(z)
p_hat(z)
samples
6
4
2
0
−2
−2
0
2
4
6


Samples=10000 var=100.0
p(z)
p_hat(z)
samples
8
6
4
2
0
−2
−2
0
2
4
6
8