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

QUESTION 1

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 2, 2019

$\mathbb{E}\left[\hat{f}\right] = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\left[f\left(\mathbf{z}^{(s)}\right)\right]$, using the linearity of expectations idea.

$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\left[f\left(\mathbf{z}^{(s)}\right)\right] = \frac{S\times\mathbb{E}[f]}{S} = \mathbb{E}\left[f\right]$. Hence, the approximation is unbiased.

We know if random variables X and Y are independent, $var\left(aX+bY\right) = a^2 var\left(X\right) + b^2 var\left(Y\right)$. Here, we have S independent samples. So, we get

$$var\left[\hat{f}\right] = \frac{1}{S^2}\sum_{s=1}^{S} var\left[f\left(z^{(s)}\right)\right] = \frac{S\times\mathbb{E}\left[\left(f-\mathbb{E}\left[f\right]\right)^2\right]}{S^2} = \frac{\mathbb{E}\left[\left(f-\mathbb{E}\left[f\right]\right)^2\right]}{S}$$

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

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 2, 2019

QUESTION 2

$$p\left(\mathbf{w}, z_1, \cdots, z_n|\mathbf{X}, \mathbf{y}, \Theta\right) \propto p\left(\mathbf{y}|\mathbf{X}, \mathbf{w}, z_1, \cdots, z_n, \Theta\right) p\left(\mathbf{w}, \Theta\right) p\left(\mathbf{z}|\Theta\right)$$
$$= \left[\prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathbf{w}^T\mathbf{x}_n, \frac{\sigma^2}{z_n}\right) \Gamma\left(z_n|\frac{v}{2}, \frac{v}{2}\right)\right] \mathcal{N}\left(\mathbf{w}|0, \rho^2\mathbf{I}_D\right)$$

Taking logarithm on both sides, we get RHS as,

$$\frac{1}{2}\sum_{n=1}^{N}\log z_n - \frac{1}{2}\frac{\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^T\mathbf{Z}\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)}{\sigma^2} + \left(\frac{v}{2}-1\right)\left(\sum_{n=1}^{N}\log z_n\right) - \frac{v}{2}\left(\sum_{n=1}^{N} z_n\right) - \frac{1}{2}\frac{\mathbf{w}^T\mathbf{w}}{\rho^2}$$

where $\mathbf{Z}$ is a diagonal matrix with entries $z_1, z_2, \cdots, z_n$. Let us now derive the conditional posteriors. (Note that the terms that contain $\mathbf{w}$ would only be included in the CP. The others get cancelled in the numerator and denominator when we write posterior probability of $\mathbf{w}$ according to Bayes rule).

$$\log p\left(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{z}, \Theta\right) \propto -\frac{1}{2}\frac{\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^T\mathbf{Z}\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)}{\sigma^2} - \frac{1}{2}\frac{\mathbf{w}^T\mathbf{w}}{\rho^2} = -\frac{1}{2}\begin{bmatrix}\mathbf{y} & \mathbf{w}\end{bmatrix}^T \begin{bmatrix}\frac{\mathbf{Z}}{\sigma^2} & -\frac{\mathbf{Z}\mathbf{X}}{\sigma^2} \\ -\frac{\mathbf{X}^T\mathbf{Z}}{\sigma^2} & \frac{\mathbf{X}^T\mathbf{Z}\mathbf{X}}{\sigma^2} + \frac{\mathbf{I}_D}{\rho^2}\end{bmatrix} \begin{bmatrix}\mathbf{y} & \mathbf{w}\end{bmatrix}$$

Using Gaussian conditional properties, we get, $\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}} = \Lambda_{\mathbf{ww}}^{-1} = \left(\frac{\mathbf{X}^T\mathbf{Z}\mathbf{X}}{\sigma^2} + \frac{\mathbf{I}_D}{\rho^2}\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}} = \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}\frac{\mathbf{X}^T\mathbf{Z}}{\sigma^2}\mathbf{y}$. Thus $p\left(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{z}, \Theta\right) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}\right)$.

$$\log p\left(z_n|\mathbf{y}, \mathbf{X}, \mathbf{z}_{-n}, \Theta\right) \propto -\frac{z_n}{2}\frac{\left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2}{\sigma^2} + \left(\frac{v+1}{2} - 1\right)\left(\log z_n\right) - \frac{v}{2}\left(z_n\right)$$

This is similar to log of Gamma distribution with parameters $\boldsymbol{\alpha} = \frac{v+1}{2}$ and $\boldsymbol{\beta} = \frac{v}{2} + \frac{\left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2}{2\sigma^2}$. Thus, $p\left(z_n|\mathbf{y}, \mathbf{X}, \mathbf{z}_{-n}, \Theta\right) = \Gamma\left(z_n|\frac{v+1}{2}, \frac{v}{2} + \frac{\left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2}{2\sigma^2}\right)$.

The Gibbs Sampler would be as follows

1. Draw $\mathbf{w}^{(0)} \sim \mathcal{N}\left(\mathbf{w}|0, \rho^2\mathbf{I}_D\right)$. Set t = 1.
2. Draw $z_n^{(t)} \sim \Gamma\left(z_n|\frac{v+1}{2}, \frac{v}{2} + \frac{\left(y_n - \left(\mathbf{w}^{(t-1)}\right)^T\mathbf{x}_n\right)^2}{2}\right)$ for n = 1, 2, ..., N
3. Draw $\mathbf{w}^{(t)} \sim \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}}^{(t)}, \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}^{(t)}\right)$ (Note that $\mathbf{Z}^{(t)}$ is used here)
4. t = t + 1; Go to step-2 if t less than T.

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 2, 2019

QUESTION 3

Let $\mathbf{Z}, \mathbf{W}$ represent all the latent varibles and words respectively. Let $\mathbf{Z}_{-dn}, \mathbf{W}_{-dn}$ be $\mathbf{Z}, \mathbf{W}$ with all but dn$^{(th)}$ position known. Then we have,

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

$$p\left(z_{dn}=k|\mathbf{Z}_{-dn}\right) = \int p\left(z_{dn}=k|\theta_d\right)p\left(\theta_d|\mathbf{Z}_{-dn}\right)d\theta_d = \int \theta_{dk}p\left(\theta_d|\mathbf{Z}_{-dn}\right)d\theta_d$$

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

$$\implies p\left(z_{dn}=k|\mathbf{Z}_{-dn}\right) = \mathbb{E}\left[\theta_{dk}\right] = \frac{\alpha+\sum_{l=1,l\neq n}^{N_d}\mathbb{I}\left[z_{dl}=k\right]}{K\alpha+N_d-1} = \frac{\alpha+N_{dk,-n}}{K\alpha+N_d-1}$$

where $N_{dk,-n}$ is the number of words in document d assigned to topic k, not including nth word.

$$p\left(w_{dn}|z_{dn}=k,\mathbf{Z}_{-dn},\mathbf{W}_{-dn}\right) = \int p\left(w_{dn}|\phi_k\right)p\left(\phi_k|\mathbf{Z}_{-dn},\mathbf{W}_{-dn}\right)d\phi_k = \int \phi_{k,w_{dn}}p\left(\phi_k|\mathbf{Z}_{-dn},\mathbf{W}_{-dn}\right)d\phi_k$$

$$p\left(\phi_k|\mathbf{Z}_{-dn},\mathbf{W}_{-dn}\right) \propto p\left(\mathbf{W}_{-dn}|\mathbf{Z}_{-dn},\phi_k\right)p\left(\phi_k\right) = \text{Dir}\left(\left\{\eta+N_{kv,-dn}\right\}_{v=1}^{V}\right)$$

where $N_{kv,-dn} = \sum_{t=1}^{D}\sum_{l=1}^{N_t}\mathbb{I}\left[z_{tl}=k\right]\mathbb{I}\left[w_{tl}=v\right]$ excluding $(t,l)=(d,n)$, as in, the number of words equal to v belonging to topic k, excluding the dn$^{(th)}$ word. Since we know $\mathbf{Z}_{-dn}$, we know the topics that each word of $\mathbf{W}_{-dn}$ belongs to. Since we are conditioning on ϕ_k, we only care about the words that belong to topic k.

$$p\left(w_{dn}|z_{dn}=k,\mathbf{Z}_{-dn},\mathbf{W}_{-dn}\right) = \mathbb{E}\left[\phi_{k,w_{dn}}\right] = \frac{\eta+N_{kw_{dn},-dn}}{V\eta+N_{k,-dn}}$$

where $N_{k,-dn} = \sum_{t=1}^{D}\sum_{l=1}^{N_t}\mathbb{I}\left[z_{tl}=k\right]$, excluding $(t,l)=(d,n)$, as in, the number of words belonging to topic k, not including the dn$^{(th)}$ word. Thus we get

$$p\left(z_{dn}=k|\mathbf{Z}_{-dn},\mathbf{W}\right) \propto \frac{\eta+N_{kw_{dn},-dn}}{V\eta+N_{k,-dn}}\frac{\alpha+N_{dk,-n}}{K\alpha+N_d-1}$$

which can be normalized (ie, sum numberator over all k to obtain denominator) giving us the exact conditional probability.

The intuitive idea is that, the probability of the word w_{dn} belonging to topic k depends on proportion of the number of times the word w_{dn} **across the corpus** belonged to topic k (excluding the current occurence), and the proportion of the number of times the words **across the document** belonged to topic k (excluding current occurence). We are looking across the corpus for word w_{dn} because it depends on topic vectors which are for the entire corpus. On the other hand, z_{dn} which is drawn from θ_d depends on the document d, so we look across the document d.

Sketch of Gibbs Sampler is as follows -

1. Initialize the latent variable matrix $\mathbf{Z} = \mathbf{Z}^{(0)}$ randomly. Each z_{dn} can take any value from 1 to K. Set t = 1

2. Compute the following for all d, n cyclically (ie, keep updating $\mathbf{Z}^{(t-1)}$ as you draw the samples z_{dn}.)

$$\pi_k^{(t)} = p\left(z_{dn}^{(t)} = k|\mathbf{Z}_{-dn}^{(t-1)}, \mathbf{W}\right) \propto \frac{\eta + N_{kw_{dn},-dn}^{(t-1)}}{V\eta + N_{k,-dn}^{(t-1)}} \frac{\alpha + N_{dk,-n}^{(t-1)}}{K\alpha + N_d - 1}$$

$$z_{dn}^{(t)} \sim \text{Multinoulli}\left(\pi^{(t)}\right)$$

3. t = t + 1; Go to step-2 if t less than T.

Basically, we are sampling the $\mathbf{Z}$ matrix repeatedly. Using S samples of $\mathbf{Z}$, we can compute the expected values of θ_d and ϕ_k applying Monte-Carlo approximation.

$$\mathbb{E}\left[\theta_{dk}\right] = \frac{1}{S}\sum_{s=1}^{S} \frac{\alpha + \sum_{l=1}^{N_d} \mathbb{I}\left[z_{dl}^{(s)} = k\right]}{K\alpha + N_d} = \frac{1}{S}\sum_{s=1}^{S} \frac{\alpha + N_{dk}^{(s)}}{K\alpha + N_d}$$

where $N_{dk}^{(s)}$ is the number of words in document d assigned to topic k based on sample $\mathbf{Z}^{(s)}$. Note that $\mathbf{Z}^{(s)}$ gives us information of which topic a word belongs to. Repeating the same for all k gives us $\mathbb{E}\left[\theta_d\right]$ vector. This makes intuitive sense because expected value depends upon the frequency of words in document d being assigned to k (assuming that apriori α out of $K\alpha$ words in document d were assigned to topic k - a uniform Dirichlet prior).

$$\mathbb{E}\left[\phi_{kv}\right] = \frac{1}{S}\sum_{s=1}^{S} \frac{\eta + N_{kv}^{(s)}}{V\eta + N_k^{(s)}}$$

where $N_{kv}^{(s)} = \sum_{t=1}^{D}\sum_{l=1}^{N_t} \mathbb{I}\left[z_{tl}^{(s)} = k\right] \mathbb{I}\left[w_{tl} = v\right]$, ie, the number of times the word v belonged to topic k in the entire corpus, and $N_k^{(s)} = \sum_{t=1}^{D}\sum_{l=1}^{N_t} \mathbb{I}\left[z_{tl}^{(s)} = k\right]$, ie, the number of words belonging to topic k across the corpus, both wrt sample $\mathbf{Z}^{(s)}$. Compute this for all v, and we get the $\mathbb{E}\left[\phi_k\right]$ vector. The expressoin makes intuitive sense because the expected value, ie, how much we expect word v to belong to topic k, depends upon the frequency of the word v being assigned to topic k (while using a uniform Dirichlet prior, ie, before experimenting - the word v belonged to topic k η times out of $V\eta$).

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

QUESTION 4

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 2, 2019

Let us write $X_{nm} = \sum_{k=1}^{K} X_{nmk}$, where $X_{nmk} \sim Pois\left(u_{nk}v_{mk}\right)$. Suppose the generative story is as follows -

1. Generate u_{nk} and v_{mk} for all n, m, k from $\Gamma\left(a, b\right)$ and $\Gamma\left(c, d\right)$ respectively.
2. Generate latent variables $X_{nmk} \sim Pois\left(u_{nk}v_{mk}\right)$.
3. $X_{nm} = \sum_{k=1}^{K} X_{nmk}$

Now we construct the Gibbs Sampler including these latent variables (like Problem 2). This means we'd have to infer / sample X_{nmk}'s as well. Let $\mathbf{u}_{-nk}$ be $\{\mathbf{u}_n\}_{n=1}^{N}$ with kth position of $\mathbf{u}_n$ unknown. Let $\mathbf{X}_{nmk}$ represent all the latent variables. Then,

$$\begin{aligned}
p\left(u_{nk}|\mathbf{u}_{-nk}, \mathbf{v}, \mathbf{X}, \mathbf{X}_{nmk}, \Theta\right) &\propto \prod_{m=1}^{M} p\left(X_{nmk}|u_{nk}, v_{mk}, \Theta\right) p\left(u_{nk}|\mathbf{u}_{-nk}, \mathbf{v}, \Theta\right) \\
&= \prod_{m=1}^{M} Pois\left(X_{nmk}|u_{nk}v_{mk}\right) \Gamma\left(u_{nk}|a, b\right) \\
&\propto \left(u_{nk}\right)^{\left(\sum_{m=1}^{M} X_{nmk} + a - 1\right)} \exp\left[-u_{nk}\left(\sum_{m=1}^{M} v_{mk} + b\right)\right]
\end{aligned}$$

This is similar to a Gamma distribution with parameters given by $\boldsymbol{\alpha} = \sum_{m=1}^{M} X_{nmk} + a$ and $\boldsymbol{\beta} = \sum_{m=1}^{M} v_{mk} + b$. Thus $u_{nk} \sim \Gamma\left(\sum_{m=1}^{M} X_{nmk} + a, \sum_{m=1}^{M} v_{mk} + b\right)$.

Following similar ideas, we get $v_{mk} \sim \Gamma\left(\sum_{n=1}^{N} X_{nmk} + c, \sum_{n=1}^{N} u_{nk} + d\right)$.

Now, we need to infer the latent variables. Note the following property of poisson random variables. Suppose $X_i \sim Pois\left(\lambda_i\right)$ for i = 1 to N, are N random variables. Let $Y = \sum_{i=1}^{N} X_i$. Then $Y \sim Pois\left(\lambda\right)$ where $\lambda = \sum_{i=1}^{N} \lambda_i$. Then,

$$p\left(X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N | Y = y\right) = \frac{\prod_{i=1}^{N} Pois\left(X_i = x_i|\lambda_i\right)}{Pois\left(Y = y|\lambda\right)} = \frac{y!}{x_1!x_2!\cdots x_N!} \prod_{i=1}^{N} \left(\frac{\lambda_i}{\lambda}\right)^{x_i}$$

which is a multinomial distribution $Mult\left(y; \frac{\lambda_1}{\lambda}, \cdots, \frac{\lambda_N}{\lambda}\right)$. Using this idea, we can write the CP of latent variables as follows -

$$p\left(X_{nm1}, X_{nm2}, \cdots, X_{nmK}|\mathbf{X}, \mathbf{u}, \mathbf{v}, \Theta\right) = Mult\left(X_{nm}; \frac{u_{n1}v_{m1}}{\mathbf{u}_n^T\mathbf{v}_m}, \cdots, \frac{u_{nK}v_{mK}}{\mathbf{u}_n^T\mathbf{v}_m}\right)$$

The Gibbs Sampler works as follows -

1. $u_{nk}^{(0)} \sim \Gamma\left(a, b\right)$ and $v_{mk}^{(0)} \sim \Gamma\left(c, d\right) \quad \forall n, m, k$. Set t = 1.
2. Draw $\left\{X_{nmk}^{(t)}\right\}_{k=1}^{K} \sim Mult\left(X_{nm}; \left(\frac{u_{n1}v_{m1}}{\mathbf{u}_n^T\mathbf{v}_m}\right)^{(t-1)}, \cdots, \left(\frac{u_{nK}v_{mK}}{\mathbf{u}_n^T\mathbf{v}_m}\right)^{(t-1)}\right) \quad \forall n, m$
3. Draw $u_{nk}^{(t)} \sim \Gamma\left(\sum_{m=1}^{M} X_{nmk}^{(t)} + a, \sum_{m=1}^{M} v_{mk}^{(t-1)} + b\right) \quad \forall n, k$
4. Draw $v_{mk}^{(t)} \sim \Gamma\left(\sum_{n=1}^{N} X_{nmk}^{(t)} + c, \sum_{n=1}^{N} u_{nk}^{(t)} + d\right) \quad \forall m, k$
5. t = t + 1; Go to step - 2 if t less than T.

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

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 2, 2019

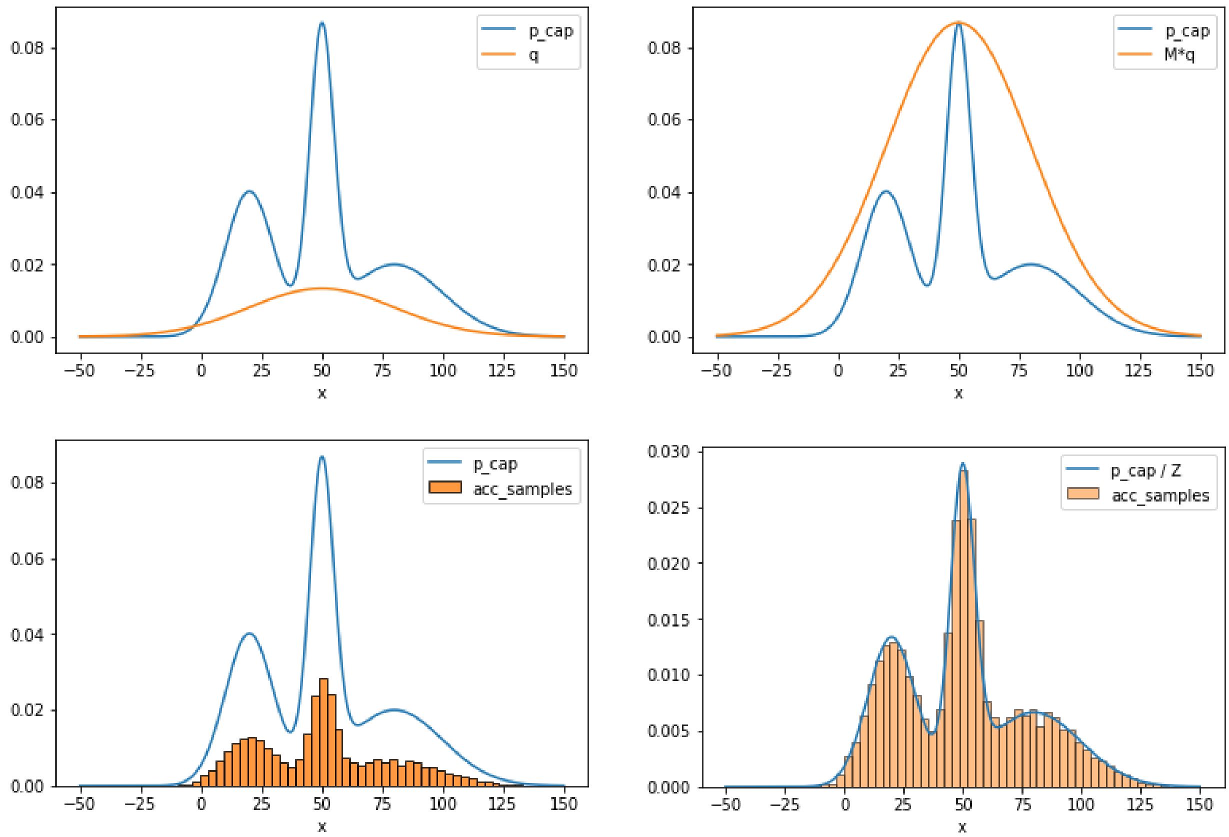
QUESTION

5

1 Part 1 - Rejection Sampling

- Obtained a value of M = 6.522 (approx). Its clear from figure that $Mq\,(x) \geq \hat{p}\,(x)$.
- Acceptance rate is found to be 0.4623. We see that $p_cap\,(z)$ occupies around half the area of $Mq\,(z)$, so the acceptance rate kind of makes sense.
- Note that p(accept) is approximately equal to acceptance rate (frequentist idea).

  p(accept) = Z / M. So, Z = 3.015 (approx). Plotting $\frac{\hat{p}(x)}{Z}$ along with the histogram of accepted samples (acc_samples), we see that they overlap very well. Hence, the acceptance rate makes sense.



2 Part 2 - MH Sampling

- The contours are plotted for probability = 0.05, as in, the level of contour = 0.05.
- The red contour represents the original distribution $p(z)$, and the yellow contour represents the approximated distribution $\hat{p}(z)$. The approx. normal distribution is obtained by computing the sample mean and the sample covariance of the samples collected until then.
- The rejection rates and time taken taken to obtain 10000 samples are given below. (Time taken may differ for other systems).

σ^2	Rejection Rate	Time (in sec)
0.01	0.0809	29.574
1	0.5961	85.127
100	0.9888	2486

Table 1: Rejection rate for various value of σ^2

- For $\sigma^2 = 0.01$, convergence is fast but the chain gets stuck in local maximum at times and doesn't explore well.
- For $\sigma^2 = 1$, convergence is not as fast as the previous case, but we reach the required region quickly (low burn-in), and explore the region well.
- For $\sigma^2 = 100$, convergence is way too slow. Its not practical to run it for so long. This is understood because the chain wanders a lot due to high variance.
- Hence, $\sigma^2 = 1$ seems to be the best choice for the proposal distribution. However if we want very quick convergence we can go with $\sigma^2 = 0.01$.
- The figures are shown in the next page.

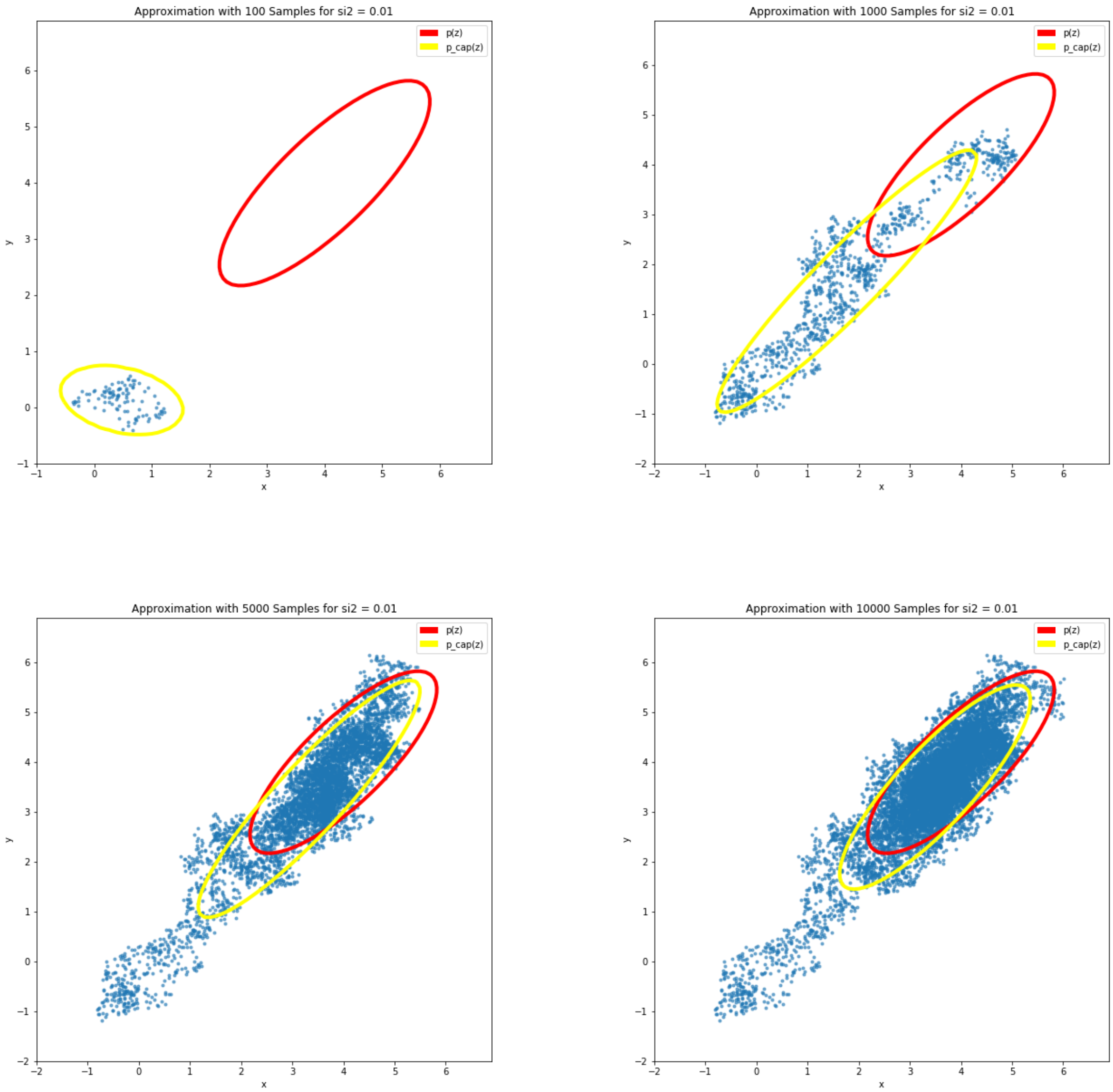


Table 2: Sampling with $\sigma^2 = 0.01$

Note that this sampler takes a lot of time in the burn-in period. Even after 100 samples haven been collected, we have still not reached the original distribution. The rejection rate is low, which is understood because we are taking small steps in a region which the sampler thinks is the local maximum.

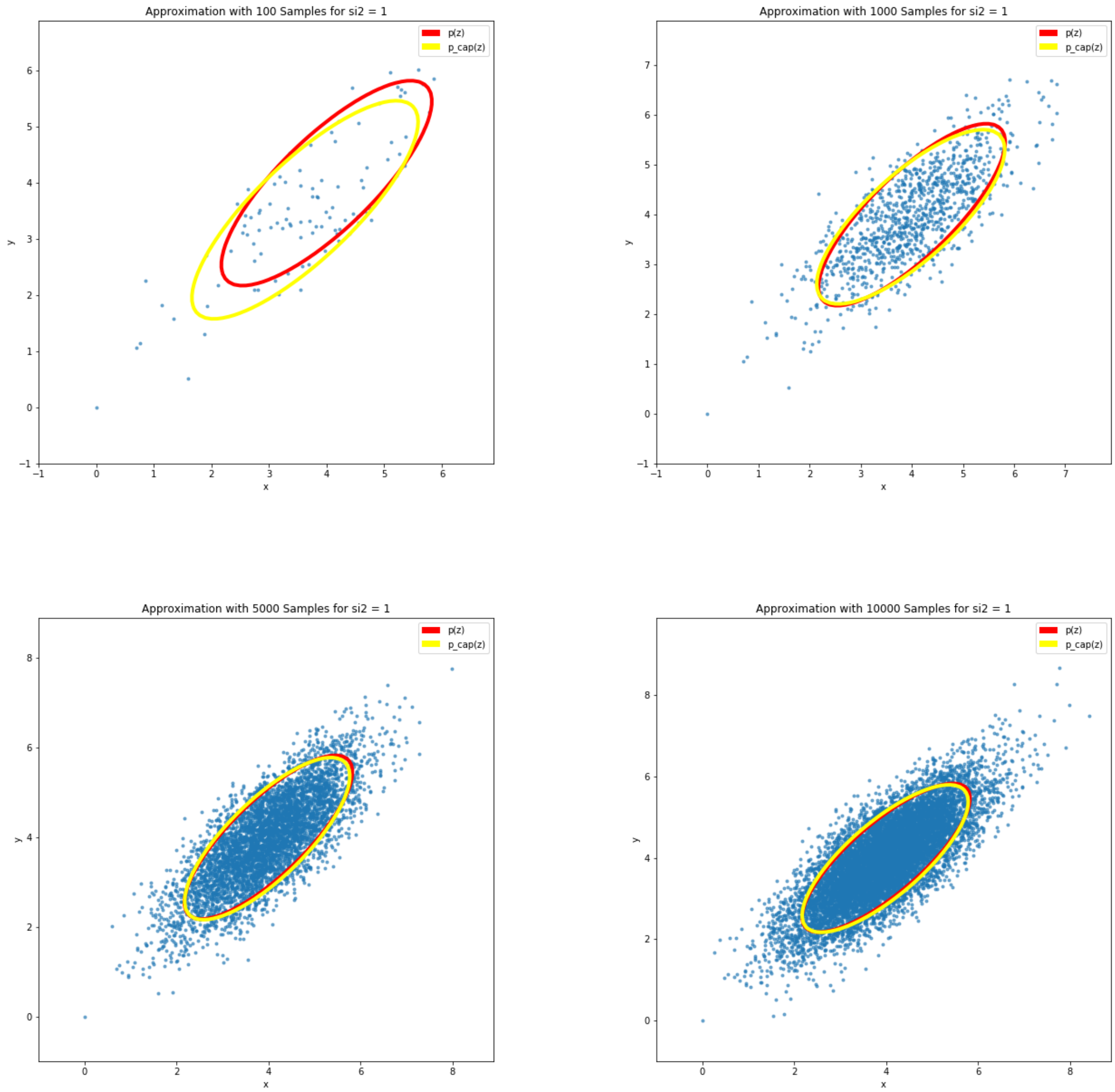


Table 3: Sampling with $\sigma^2 = 1$

The rejection rate is higher than the previous case but the burn-in is low. We reach the original distribution with 100 samples, and explore the required region, without getting stuck in any local maxima.

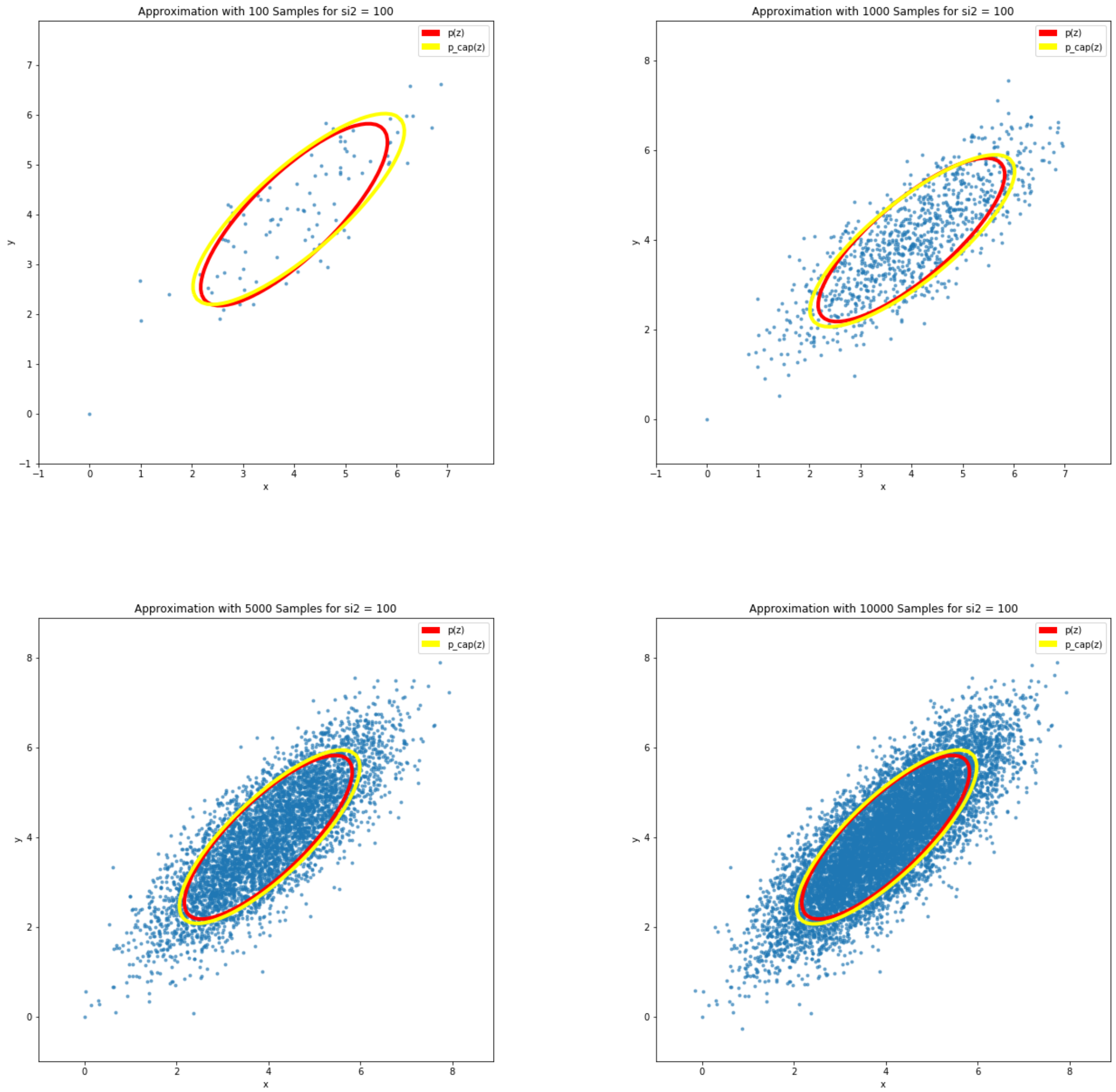


Table 4: Sampling with $\sigma^2 = 100$

The rejection rate is very high, and the time taken to convergence is approximately 45 minutes, which is too high.