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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: March 24, 2019

QUESTION 1

Monte-Carlo Estimates

Given,

$$\bar{f} = \frac{1}{S}\sum_{s=1}^{S} f(z^s)$$

Assumption : samples are independent of each other.

Expectation

We can say that approximation is unbiased if expected values remains same in limit, using the fact that samples are independently distributed we can say that :

$$\begin{aligned}\mathbb{E}(\bar{f}) &= \mathbb{E}(\frac{1}{S}\sum_{s=1}^{S} f(z^s)) \\ &= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}(f(z^s)) \\ &= \frac{S \times \mathbb{E}(f)}{S} \\ &= \mathbb{E}(f)\end{aligned}$$

So it's an unbiased estimate

Variance

we can write the variance of this estimate as follows:

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

As we can note that the variance in this estimate goes down as sample size increases. So Monte carlo are very good approximation in for large sample size.

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

QUESTION 2

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: March 24, 2019

Gaussian Scale Mixture

Given,
we have a regression model as follows as:

$$\begin{aligned} p(y_n|x_n, w, \sigma^2, \nu) &= \tau(y|w^\intercal x, \sigma^2, \nu) \\ &= \int \mathcal{N}(y|w^\intercal x, \frac{\sigma^2}{z}) Gamma(z|\frac{\nu}{2}, \frac{\nu}{2}) dz \\ p(w) &= \mathcal{N}(w|0, \rho^2 I_d) \end{aligned}$$

Considering z as a latent variable, let's write down the Joint distribution first,

$$\begin{aligned} p(Y, Z, w|X, \sigma^2, \nu, \rho^2) &= p(Y, Z|w, X, \sigma^2, \nu) p(w|\rho^2) \\ &= \prod_{n=1}^{N} \left[\mathcal{N}(y_n|w^\intercal x_n, \frac{\sigma^2}{z_n}) Gamma(z_n|\frac{\nu}{2}, \frac{\nu}{2}) \right] \mathcal{N}(w|0, \rho^2 I_d) \end{aligned}$$

Conditional Posteriors

For w ignoring terms not containing w and using linear gaussian models property we can write

$$\begin{aligned} p(w|Z, X, Y) &\propto p(Y, Z, w|X, \sigma^2, \rho^2, \nu) \\ &\propto \mathcal{N}(Y|Xw, \sigma^2 \Lambda_z^{-1}) \mathcal{N}(w|0, \rho^2 I_d) \\ &= \mathcal{N}(w|(\frac{\sigma^2}{\rho^2} I_d + X^T \Lambda_z X)^{-1} X^T \Lambda_z Y, (\frac{1}{\rho^2} I_d + \frac{1}{\sigma^2} X^T \Lambda_z X)^{-1}) \end{aligned}$$

where, $\Lambda_z^{nn} = z_n$ is a diagonal matrix.
Similarily for z_i we can write the conditional distribution as follows:

$$\begin{aligned} p(z_i|w, Z_{-i}, X, Y) &\propto p(y_i|z_i, w, x_i, \sigma^2) p(z_i|\nu) \\ &\propto \mathcal{N}(y_i|w^\intercal x_i, \frac{\sigma^2}{z_i}) Gamma(z_i|\frac{\nu}{2}, \frac{\nu}{2}) \\ &= Gamma\left(z_i|\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \end{aligned}$$

Gibbs Sampling

1. Initialize $w = w_0$ and for all $i \in [1, n], z_i = z_{i0}$
2. For t = 1:T
   (a) sample w^t from equation of $p(w|X, Y, Z)$ as given above.
   (b) sample z_i^t from equation of $p(z_i|X, Y, Z_{-i}, w)$ as given above for all values of i.
3. save $(w^t, z_1^t, ..., z_n^t)$ as one sample.

This concludes the answer to above question.

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

QUESTION 3

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: March 24, 2019

LDA as Gibbs Sampling

Given, The LDA model as

$$\begin{aligned}\phi_k &\sim Dir(\eta)\\ \theta_d &\sim Dir(\alpha)\\ z_{d,n} &\sim mult(\theta_d)\\ w_{d,n} &\sim mult(\phi_{z_{d,n}})\end{aligned}$$

We need to sample z_n, d so we can write the conditional posterior as product of likelihood and conditional prior as follows:

$$p(z_{n,d} = k|Z_{\neg n,d}, W, \alpha, \eta) \propto p(w_{n,d}|z_{n,d} = k, Z_{\neg n,d}, W_{\neg n,d}, \eta)p(z_{n,d} = k|Z_{\neg n,d}, \alpha) \tag{3.1}$$

Now we can write the likelihood part as :

$$p(w_{n,d}|z_{n,d} = k, Z_{\neg n,d}, W_{\neg n,d}, \eta) = \int p(w_{n,d}|z_{n,d} = k, \phi_k)p(\phi_k|Z_{\neg n,d}, W_{\neg n,d})d\phi_k \tag{3.2}$$

For above equation we need posterior over ϕ_k, so we can write:

$$\begin{aligned}p(\phi_k|Z_{\neg n,d}, W_{\neg n,d}) &\propto p(W_{\neg n,d}|\phi_k, Z_{\neg n,d})p(\phi_k) \\ &\propto \prod_{i\in\neg n,d} \phi_{kv}^{\mathbb{I}(z_i=k,w_i=v)} \prod_{v=1}^{V} \phi_{kv}^{\eta-1} \\ &\propto \prod_{v=1}^{V} \phi_{kv}^{N_{kv}+\eta-1} \\ &= Dir(\eta + N_{k1}, \eta + N_{k2}, ..., \eta + N_{kV})\end{aligned} \tag{3.3}$$

where , $N_{kv} = \sum_{i\in\neg n,d} I(z_i = k, w_i = v)$
Now we can write equation 3.2 by ignoring terms that do not depend on $w_{n,d}$ as follows:

$$\begin{aligned}p(w_{n,d}|z_{n,d} = k, Z_{\neg n,d}, W_{\neg n,d}, \eta) &= \frac{\prod_{v=1}^{V}\Gamma(\eta + N_{kv})}{\Gamma(\sum_{v=1}^{V}\eta + N_{kv})} \int \phi_{kw_{n,d}} \prod_{v=1}^{V} \phi_{kv}^{\eta+N_{kv}-1} d\phi_k \\ &\propto \frac{\Gamma(\eta + N_{kw_{n,d}} + 1)}{\Gamma(\eta + N_{kw_{n,d}})} \\ &\propto \eta + N_{kw_{n,d}} \\ &= \frac{\eta + N_{kw_{n,d}}}{\sum_v \eta + N_{kv}} \\ &= \frac{\eta + N_{kw_{n,d}}}{V\eta + N_k}\end{aligned} \tag{3.4}$$

Now moving our attention towards the second term in equation 3.1 the conditional prior as follows:

$$p(z_{n,d} = k|Z_{\neg n,d}, \alpha) = \int p(z_{n,d} = k|\theta_d)p(\theta_d|Z_{\neg n,d}, \alpha)d\theta_d \tag{3.5}$$

Again we need to find posterior here so we can write:

$$\begin{aligned} p(\theta_d|Z_{\neg n,d}, \alpha) &\propto p(Z_{\neg n|d}|\theta_d)p(\theta_d|\alpha) \\ &\propto [\prod_{i\in\neg n|d}\prod_{k=1}^{K}\theta_{dk}^{\mathbb{I}(z_i=k)}]\prod_{k=k}^{K}\theta_{dk}^{\alpha-1} \\ &\propto \prod_{k=1}^{K}\theta_{dk}^{N_{dk}+\alpha-1} \\ &= Dir(\alpha + N_{d1}, \alpha + N_{d2}, ..., \alpha + N_{dK}) \end{aligned} \tag{3.6}$$

where , $N_{dk} = \sum_{i\in\neg n|d}\mathbb{I}(z_i = k)$,and $\neg n|d$ means words in document d excluding n
Now we can write equation 3.5 as follows:

$$\begin{aligned} p(z_{n,d} = k|Z_{\neg n,d}, \alpha) &= \frac{\prod_{k=1}^{K}\Gamma(\alpha + N_{dk})}{\Gamma(\sum_{k=1}^{K}\alpha + N_{dk})}\int \theta_{dk}\prod_{i=1}^{K}\theta_{di}d\theta_d \\ &\propto \frac{\Gamma(\alpha + N_{dk} + 1)}{\Gamma(\alpha + N_{dk})} \\ &\propto \alpha + N_{dk} \\ &= \frac{\alpha + N_{dk}}{\sum_k \alpha + N_{dk}} \\ &= \frac{\alpha + N_{dk}}{K\alpha + N_d} \end{aligned} \tag{3.7}$$

Now we can write equation 3.1 as follows:

$$\begin{aligned} p(z_{n,d} = k|Z_{\neg n,d}, W, \alpha, \eta) &\propto \left(\frac{\eta + N_{kw_{n,d}}}{V\eta + N_k}\right)\left(\frac{\alpha + N_{dk}}{K\alpha + N_d}\right) \\ &= \frac{\left(\frac{\eta+N_{kw_{n,d}}}{V\eta+N_k}\right)\left(\frac{\alpha+N_{dk}}{K\alpha+N_d}\right)}{\sum_{k=1}^{K}\left(\frac{\eta+N_{kw_{n,d}}}{V\eta+N_k}\right)\left(\frac{\alpha+N_{dk}}{K\alpha+N_d}\right)} \end{aligned} \tag{3.1}$$

Why this makes sense ?
Well if you notice the first term in numerator is ratio representing total number of times word $w_{n,d}$ has appeared in topic k except the current word, so its the probability of that word being in that topic.

Similarly second term represents the probability of that topic being in current document, excluding the current word.

So it's like probability of word $w_{d,n}$ begin from topic k^{th} in the current document which makes perfect sense.

Posterior Expectations

We can now get certain expectations pretty easily, since we have already seen the posterior forms of these random variable in equations 3.3 and 3.6 as follows:

$$\mathbb{E}(\theta_{dk}) = \frac{\sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = k) + \alpha}{\sum_{i=1}^{K} \sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = i) + K\alpha}$$

Intuitive sense:
Similar to ratio of number of word assigned to topic k in given document to total number of words in document but with some injected prior. (i.e: probability of topic in document.)

$$\mathbb{E}(\phi_{kv}) = \frac{\eta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}(z_{nd} = k, w_{nd} = v)}{V\eta + \sum_{v=1}^{V} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}(z_{nd} = k, w_{nd} = v)}$$

Intuitive sense:
Similar to ratio of number of times the given word has been present in given topic to total number of words in the given topic with some injected prior. (i.e : probability of word in that topic)
They are essentially the expression we obtained in equations 3.4 and 3.7 but with the all the word included(not excluding any word).So Now we can write:

$$\mathbb{E}(\phi_k) = [\mathbb{E}(\phi_{k1}), \mathbb{E}(\phi_{k2}), ..., \mathbb{E}(\phi_{kV})]$$
$$\mathbb{E}(\theta_d) = [\mathbb{E}(\theta_{d1}), \mathbb{E}(\theta_{d2}), ..., \mathbb{E}(\theta_{dK})]$$

This concludes the answer to the above question.

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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: March 24, 2019

QUESTION 4

Gamma Poisson matrix factorization

Given Matrix Factorization setup,

$$\begin{aligned} X_{nm} &\sim Poisson(X_{nm}|u_n^T v_m) \\ u_{nk} &\sim Gamma(u_{nk}|a,b) \\ v_{mk} &\sim Gamma(v_{mk}|c,d) \end{aligned}$$

We will use the property that sum of poisson random variables is also a poisson distribution, so using the property of poisson distribution we can write our likelihood in a two stage process as:

$$X_{nmk} \sim Poisson(u_{nk}v_{mk})$$

and

$$X_{nm} = \sum_k X_{nmk}$$

which is same as sampling from original likelihood and so we can sample each individual part from a multinoulli distribution as sampling from a fixed budget with expected values of individual poisson model,

$$[X_{nm1}, X_{nm2}, .., X_{nmK}] \sim mult(X_{nm}|p_1, p_2, ..., p_k) \tag{4.1}$$

where,

$$p_i = \frac{u_{ni}v_{mi}}{\sum_k u_{nk}v_{mk}}$$

so that we can have sampling from multiple poisson distribution with same expected value.

Joint Distribution

Let's write down the joint distribution first, as:

$$\begin{aligned} p(X,U,V) &= p(X|U,V)p(U)p(V) \qquad (4.2) \\ p(X|U,V) &= \left[\prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K}\frac{1}{X_{nmk!}}(u_{nk}v_{mk})^{X_{nmk}}\exp(-u_{nk}v_{mk})\right] \\ p(U) &= \prod_{n=1}^{N}\prod_{k=1}^{K}\frac{b^a}{\Gamma(a)}u_{nk}^{a-1}\exp(-bu_{nk}) \\ p(V) &= \prod_{m=1}^{M}\prod_{k=1}^{K}\frac{d^c}{\Gamma(c)}v_{mk}^{c-1}\exp(-dv_{mk}) \end{aligned}$$

Now we can write the conditional posterior in terms of joint distribution ignoring the terms that do not contain random variables as follows:

$$
\begin{aligned}
p(u_{nk}|X, U_{\neg nk}, V) &\propto p(X|U, V)p(U) \qquad (4.3)\\
&\propto \left(\prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K}\frac{1}{X_{nmk!}}(u_{nk}v_{mk})^{X_{nmk}}\exp(-u_{nk}v_{mk})\right)\left(\prod_{k=1}^{K}u_{nk}^{a-1}\exp\left(-bu_{nk}\right)\right)\\
&\propto u_{nk}^{(\sum_m X_{nmk}+a-1)}\exp(-(b+\sum_m v_{mk}))\\
&= Gamma(a+\sum_m X_{nmk}, b+\sum_m v_{mk})
\end{aligned}
$$

Similarly we can find conditional posterior for v_{mk} as follows :

$$
\begin{aligned}
p(v_{mk}|X, U_{\neg nk}, V) &\propto p(X|U, V)p(V) \qquad (4.4)\\
&\propto \left(\prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K}\frac{1}{X_{nmk!}}(u_{nk}v_{mk})^{X_{nmk}}\exp(-u_{nk}v_{mk})\right)\left(\prod_{k=1}^{K}u_{nk}^{a-1}\exp\left(-bu_{nk}\right)\right)\\
&\propto v_{mk}^{(\sum_n X_{nmk}+c-1)}\exp(-(d+\sum_n u_{nk}))\\
&= Gamma(c+\sum_n X_{nmk}, d+\sum_n u_{nk})
\end{aligned}
$$

Gibbs Sampler

1. Initialize Sample some initial values for $U = U^0$, $V = V^0$ matrix from priors.
2. for t = 1:T
   (a) sample k X_{nmk}^t from equation 4.1
   (b) sample for all n and k u_{nk}^t from equation 4.3
   (c) sample for all m and k v_{mk}^t from equation 4.4
3. save sampled U^t, V^t matrices as current sample.

This concludes the answer for given question.

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

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: March 24, 2019

QUESTION 5

Sampling Methods

Rejection Sampling

Value of M obtained : **6.5203**
Acceptance rate obtained : **0.4668**

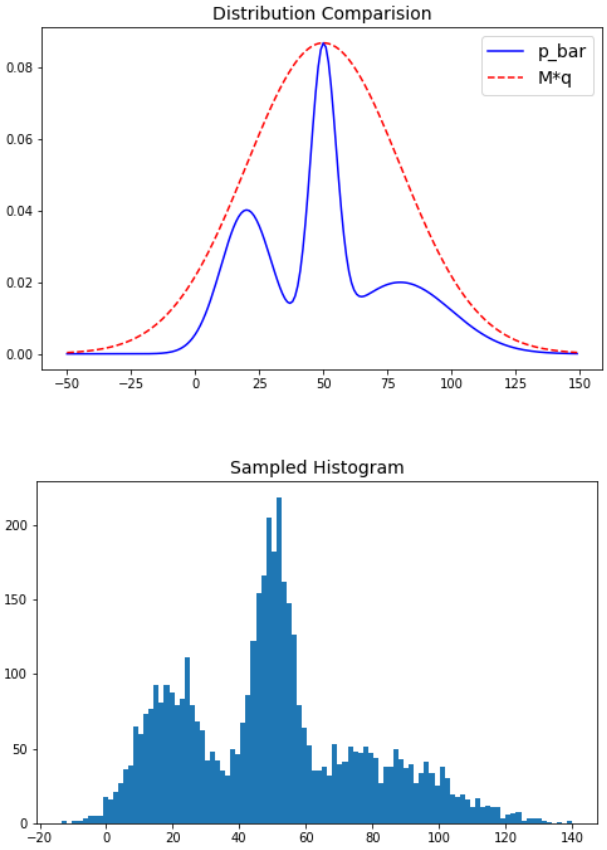

Figure 1: Plots of distributions and Histogram of samples

Yes , the acceptance rate makes sense based on M since the sum of value of $\bar{p}$ is 3 as its sum of 3 distributions and the sum for proposal distribution will be 6.52, so ratio of that will be near 0.46. so it makes very intuitive sense to have that as acceptance rate.

MH sampling

Notations
Red curve : Original Distribution
Green curve : Emperical Distribution
Observation : Higher varince results in slow convergence and large number of rejected samples.
Results
At Variance : **0.01**
Rejection Rate Obtained : **0.0752**

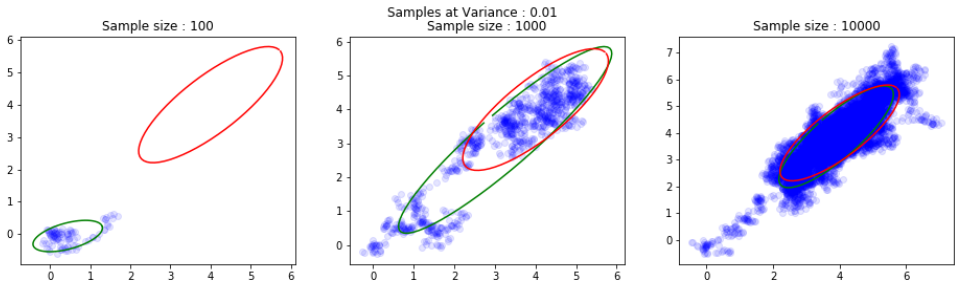

Figure 2: Samples at variance 0.01

At Variance : **1.0**
Rejection Rate Obtained : **0.5987**

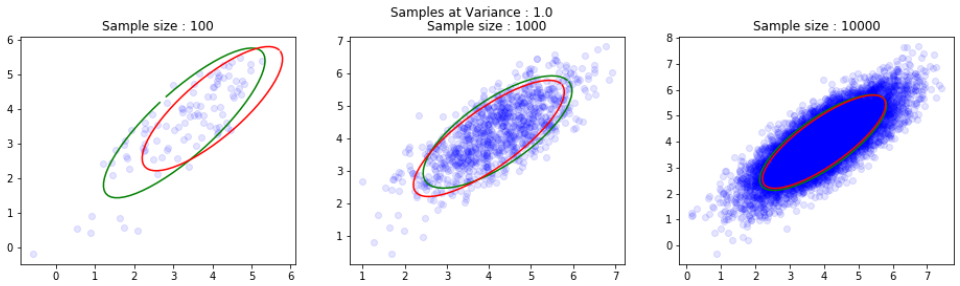

Figure 3: Samples at variance 1.0

At Variance : **100.0**
Rejection Rate Obtained : **0.9883**

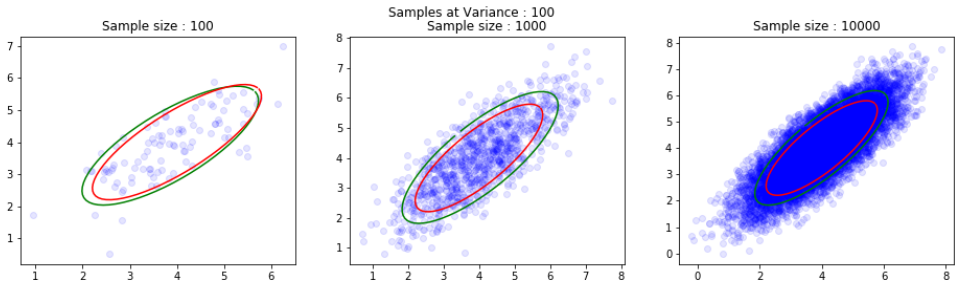

Figure 4: Samples at variance 100.0

Out of three cases the case with **1.0 variance seems the best** since the number of samples required to get a good approximate is less as compared to 0.01 case. Also it's better than the third variance case since variance of 100 will have a lot of rejection and too much varying samples and will be very slow. So the variance of 1.0 will have the best of both worlds better fit with less number of samples and moderate rejection rate.

Observation : The gaussian fit over the high variance sampler is bigger than the original. This may be because of highly varying samples.

This concludes the answer to given question.