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

QUESTION 1

Student Name: Subham Kumar
Roll Number: 160707
Date: August 19, 2019

Since $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, ..., \mathbf{z}^{(S)}$ are drawn i.i.d from $p(\mathbf{z})$,

$$E\left[f(\mathbf{z}^{(i)})\right] = E\left[f(\mathbf{z}^{(j)})\right] = E\left[f(\mathbf{z})\right] \ \ \forall \, i \neq j$$

$$var\left[f(\mathbf{z}^{(i)})\right] = var\left[f(\mathbf{z}^{(j)})\right] = var\left[f(\mathbf{z})\right] \ \ \forall \, i \neq j$$

Now,

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

$$= \frac{1}{S}\sum_{s=1}^{S} E\left[f(\mathbf{z}^s)\right] \ \ \text{(using linearity of expectation)}$$

$$= \frac{S \times E\left[f(\mathbf{z})\right]}{S} = E\left[f(\mathbf{z})\right]$$

Hence monte-carlo approximation is unbiased.
Also,

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

Using the fact that $var\left[a_1X_1 + a_2X_2 + ... + a_sX_s\right] = a_1^2 var[X_1] + a_2^2 var[X_2] + ... + a_s^2 var[X_S]$ where $X_1, X_2, .., X_s$ are i.i.d.,we have,

$$var\left[\frac{1}{S}\sum_{s=1}^{S} f(\mathbf{z}^{(s)})\right] = \frac{1}{S^2}\sum_{s=1}^{S} var\left[f(\mathbf{z}^{(s)})\right]$$

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

Student Name: Subham Kumar
Roll Number: 160707
Date: August 19, 2019

QUESTION 2

For the augmented model $p(y_n, z_n|\mathbf{w}, \mathbf{x}_n, \sigma^2, \nu) = \mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n, \frac{\sigma^2}{z_n})Gamma(z_n|\frac{\nu}{2}, \frac{\nu}{2})$

$$\begin{aligned}
p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{z}, \rho^2, \sigma^2) &\propto p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathbf{z}, \sigma^2)p(\mathbf{w}|\rho^2) \\
&\propto \prod_{n=1}^{N} p(y_n|\mathbf{w}, \mathbf{x}_n, z_n, \sigma^2)p(\mathbf{w}|\rho^2) \\
&\propto \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n, \frac{\sigma^2}{z_n})\mathcal{N}(\mathbf{w}|0, \rho^2\mathbf{I}_D) \\
&\propto \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, Diag\left\{\frac{\sigma^2}{z_1}, \frac{\sigma^2}{z_2}, ..., \frac{\sigma^2}{z_N}\right\}\right)\mathcal{N}(\mathbf{w}|0, \rho^2\mathbf{I}_D)
\end{aligned}$$

Writing $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ where $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \rho^2\mathbf{I}_D)$ and $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = Diag\left\{\frac{\sigma^2}{z_1}, \frac{\sigma^2}{z_2}, ..., \frac{\sigma^2}{z_N}\right\}$.
Now using formula for gaussian conditional posterior,we have $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{z}, \sigma^2, \rho^2) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}}\right)$ where,

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \frac{\mathbf{I}_D}{\rho^2}\right)^{-1}$$
$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

Note that $\boldsymbol{\Sigma}^{-1} = Diag\left\{\frac{z_1}{\sigma^2}, \frac{z_2}{\sigma^2}, ..., \frac{z_N}{\sigma^2}\right\}$
Similarily writing conditional posterior for z_n,

$$\begin{aligned}
p(z_n|\mathbf{y}, \mathbf{w}, \mathbf{X}, \mathbf{z}_{-n}, \sigma^2, \nu) &\propto p(y_n|z_n, \mathbf{w}, \mathbf{x}_n, \sigma^2)p(z_n|\nu) \\
&\propto \mathcal{N}\left(y_n|\mathbf{w}^T\mathbf{x}_n, \frac{\sigma^2}{z_n}\right)Gamma(z_n|\frac{\nu}{2}, \frac{\nu}{2}) \\
&\propto z_n^{\frac{\nu+1}{2}-1}exp\left(-z_n\left\{\frac{(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2} + \frac{\nu}{2}\right\}\right)
\end{aligned}$$

which is similar to Gamma distribution with shape parameter $= \frac{\nu+1}{2}$ and rate parameter $= \left\{\frac{(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2} + \frac{\nu}{2}\right\}$.
Hence $p(z_n|\mathbf{y}, \mathbf{w}, \mathbf{X}, \mathbf{z}_{-n}, \sigma^2, \nu) = Gamma\left(z_n|\frac{\nu+1}{2}, \frac{(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2} + \frac{\nu}{2}\right)$

The Gibbs Sampling Algorithm:
1.Initialize $\mathbf{w}^{(0)}$ randomly.Set t=1
2.While $t \leq T$,

Sample $z_n^{(t)}$ from $Gamma\left(z_n|\frac{\nu+1}{2}, \frac{(y_n - \mathbf{w}^{(t-1)T}\mathbf{x}_n)^2}{2\sigma^2} + \frac{\nu}{2}\right)\ \forall\, n$
Sample $\mathbf{w}^{(t)}$ from $\mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}^{(t)}_{\mathbf{w}|\mathbf{y}}, \boldsymbol{\Sigma}^{(t)}_{\mathbf{w}|\mathbf{y}}\right)$
Set t=t+1

$$p(z_{d,n} = k|\mathbf{z}_{-d,n}, \mathbf{w}) \propto p(w_{d,n}|z_{d,n} = k, \mathbf{z}_{-d,n}, \mathbf{w}_{-d,n})p(z_{d,n} = k|\mathbf{z}_{-d,n})$$

The first term in R.H.S. can be written as:

$$p(w_{d,n}|z_{d,n} = k, \mathbf{z}_{-d,n}, \mathbf{w}_{-d,n}) = \int p(w_{d,n}|\boldsymbol{\phi}_k)p(\boldsymbol{\phi}_k|\mathbf{z}_{-d,n}, \mathbf{w}_{-d,n})d\boldsymbol{\phi}_k$$

$$= \int \phi_{k,w_{d,n}}\, p(\boldsymbol{\phi}_k|\mathbf{z}_{-d,n}, \mathbf{w}_{-d,n})d\boldsymbol{\phi}_k$$

$$= E_{p(\boldsymbol{\phi}_k|\mathbf{z}_{-d,n},\mathbf{w}_{-d,n})}\left[\phi_{k,w_{d,n}}\right]$$

Now,

$$p(\boldsymbol{\phi}_k|\mathbf{z}, \mathbf{w}, \eta) \propto p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}_k)p(\boldsymbol{\phi}_k|\eta)$$

$$\propto \prod_{v=1}^{V} \phi_{k,v}^{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{I}[w_{d,n}=v]\mathbb{I}[z_{d,n}=k]+\eta-1}$$

which is $Dir\left(\eta_1, \eta_2, ..., \eta_V\right)$ where $\eta_v = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{I}[w_{d,n} = v]\mathbb{I}[z_{d,n} = k] + \eta$.
Hence,

$$p(\boldsymbol{\phi}_k|\mathbf{z}_{-d,n}, \mathbf{w}_{-d,n}) = Dir\left(\eta_1^{'}, \eta_2^{'}, ..., \eta_V^{'}\right)$$

where $\eta_v^{'} = \sum_{t=1}^{D}\sum_{m=1}^{N_d}\mathbb{I}[w_{t,m} = v]\mathbb{I}[z_{t,m} = k] + \eta$ not including $(t, m) = (d, n)$.
Hence,$E_{p(\boldsymbol{\phi}_k|\mathbf{z}_{-d,n},\mathbf{w}_{-d,n})}\left[\phi_{k,w_{d,n}}\right] = \frac{\eta_{w_{d,n}}^{'}}{V\eta+N_{k,-dn}}$ where $N_{k,-dn} = \sum_{t=1}^{D}\sum_{m=1}^{N_d}\mathbb{I}[z_{t,m} = k]$ not including $(t, m) = (d, n)$.
Now solving for the second term i.e.

$$p(z_{d,n} = k|\mathbf{z}_{-d,n}) = \int p(z_{d,n} = k|\mathbf{z}_{-d,n}, \boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d|\mathbf{z}_{-d,n})d\boldsymbol{\theta}_d$$

$$= \int \theta_{d,k}\, p(\boldsymbol{\theta}_d|\mathbf{z}_{-d,n})d\boldsymbol{\theta}_d$$

$$= E_{p(\boldsymbol{\theta}_d|\mathbf{z}_{-d,n})}\left[\theta_{d,k}\right]$$

Now,

$$p(\boldsymbol{\theta}_d|\mathbf{z}_{d,1:N_d}) \propto p(\mathbf{z}_{d,1:N_d}|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d|\alpha)$$

$$\propto \prod_{k=1}^{K} \theta_{d,k}^{\sum_{n=1}^{N_d}\mathbb{I}[z_{d,n}=k]+\alpha-1}$$

which is $Dir\left(\alpha_1, \alpha_2, ..., \alpha_K\right)$ where $\alpha_k = \sum_{n=1}^{N_d}\mathbb{I}[z_{d,n} = k] + \alpha$. Hence

$$p(\boldsymbol{\theta}_d|\mathbf{z}_{-d,n}) = Dir\left(\alpha_1^{'}, \alpha_2^{'}, ..., \alpha_K^{'}\right)$$

where $\alpha_k^{'} = \sum_{m=1,m\neq n}^{N_d} \mathbb{I}[z_{d,m} = k] + \alpha$.

Hence,$E_{p(\boldsymbol{\theta}_d|\mathbf{z}_{-d,n})}[\theta_{d,k}] = \frac{\alpha_k^{'}}{K\alpha+N_d-1}$

Hence

$$p(z_{d,n} = k|\mathbf{z}_{-d,n}, \mathbf{w}) \propto \frac{\eta_{w_{d,n}}^{'}}{V\eta + N_{k,-dn}} \frac{\alpha_k^{'}}{K\alpha + N_d - 1}$$

$$p(z_{d,n} = k|\mathbf{z}_{-d,n}, \mathbf{w}) = \pi_k = \frac{\frac{\eta_{w_{d,n}}^{'}}{V\eta+N_{k,-dn}} \frac{\alpha_k^{'}}{K\alpha+N_d-1}}{\sum_{l=1}^{K} \frac{\eta_{w_{d,n}}^{'}}{V\eta+N_{l,-dn}} \frac{\alpha_l^{'}}{K\alpha+N_d-1}}$$

So,$p(z_{d,n}|\mathbf{z}_{-d,n}, \mathbf{w})$ is a $multinoulli(\pi_1, \pi_2, ..., \pi_K)$.

The expression implies that the the probability of the word $w_{d,n}$ belonging to topic k depends on the proportion of the number of times the words across the document belonged to topic k (excluding current occurence) and proportion of the number of times the word $w_{d,n}$ across the corpus belonged to topic k (excluding current occurence).$z_{d,n}$ which is drawn from $\boldsymbol{\theta}_d$ depends on the document d, so we look across the document d,whereas for word $w_{d,n}$ we look across entire corpus because it depends on topic vectors which are kind of support for the entire corpus.

The Gibbs Sampling Algorithm:

1.Initialize $\mathbf{z}^{(0)}$,set t=1.

2.While $t \leq T$,

Sample $z_{d,n}^{(t)}$ from $multinoulli(\pi_1, \pi_2, ..., \pi_K)\ \ \forall\, n, d$

Note: Most recent values of $z_{l,m}((l, m) \neq (d, n))$ are used to calculate $\pi_1, \pi_2, ..., \pi_K$ Set t=t+1

We will use monte-carlo approximation using the samples $\mathbf{z}$ for $E[\boldsymbol{\theta_d}] = (E[\theta_{d,1}], ..., E[\theta_{d,K}])$ as well as $E[\boldsymbol{\phi}_k] = (E[\phi_{k,1}], ..., E[\phi_{k,V}])$, where

$$E[\theta_{d,k}] = \frac{1}{S} \sum_{s=1}^{S} \frac{\alpha + N_{d,k}^{(s)}}{K\alpha + N_d}$$

and $N_{d,k}^{(s)} = \sum_{n=1}^{N_d} \mathbb{I}[z_{d,l}^{(s)} = k]$ which is number of words in document d assigned to topic k given $\mathbf{z}_{d,1:N_d}^{(s)}$. Also

$$E[\phi_{k,v}] = \frac{1}{S} \sum_{s=1}^{S} \frac{\eta + N_{kv}^{(s)}}{V\eta + N_k^{(s)}}$$

where $N_k^{(s)} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n}^{(s)} = k]$(i.e. number of words belonging to topic k over the corpus given $\mathbf{z}^{(s)}$) and $N_{kv}^{(s)} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}[w_{d,n} = v]\mathbb{I}[z_{d,n}^{(s)} = k]$(i.e. number of times word v belonged to topic k in complete corpus given $\mathbf{z}^{(s)}$).

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

Student Name: Subham Kumar
Roll Number: 160707
Date: August 19, 2019

QUESTION 4

Given $p(X_{nm}|\mathbf{u}_n, \mathbf{v}_m) = Poisson(X_{nm}|\sum_{k=1}^{K} u_{nk}v_{mk})$. We can augment the representation of X_{nm} as $X_{nm} = \sum_{k=1}^{K} X_{nmk}$ so that,

$$p(X_{nmk}|u_{nk}, v_{mk}) = Poisson(X_{nmk}|u_{nk}v_{mk})$$

.
Note that now we also need to infer these latent variables now.

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

which resembles a Gamma distribution with shape parameter $= \sum_{m=1}^{M} X_{nmk} + a$ and rate parameter $= b + \sum_{m=1}^{M} v_{mk}$.
Here $\Theta = \{a, b, c, d\}$ which are assumed to be known.
Similarily $p(v_{mk}|X_{\{1:N\}mk}, \mathbf{v}_{-mk}, \Theta) = Gamma\left(v_{mk}|\sum_{n=1}^{N} X_{nmk} + c, d + \sum_{n=1}^{N} u_{nk}\right)$
Lemma: If $\mathbf{X} = (X_1, X_2, ..., X_K)$ are poisson distributed with parameters $\lambda_1, \lambda_2, ..., \lambda_3$ respectively,then the joint distribution of $\mathbf{X}$ conditioned on $Y = \sum_{i=1}^{K} X_i$ will be a multinouli distribution.
Proof:

$$p(\mathbf{X}) = \prod_{i=1}^{K} Poisson(X_i|\lambda_i)$$

. Also note that $p(Y) = Poisson(Y|\lambda)$ where $\lambda = \sum_{i=1}^{K} \lambda_k$ Now

$$p(\mathbf{X}|Y) = \frac{p(\{X_1 = x_1, X_2 = x_2, ..., X_K = x_k\} \bigcap \{Y = y\})}{p(Y = y)}$$

$$p(\{X_1 = x_1, X_2 = x_2, ..., X_K = x_k\} \bigcap \{Y = y\}) = \begin{cases} 0 & if\ y \neq \sum_{i=1}^{K} x_i \\ p(\mathbf{X}) & otherwise \end{cases}$$

Hence,

$$p(\mathbf{X}|Y) = \frac{\prod_{i=1}^{K} \lambda_i^{x_i} \times y!}{\lambda^{\sum_{i=1}^{K} x_i} x_1!x_2!...x_k!} = \frac{y!}{x_1!x_2!...x_K!} \prod_{i=1}^{K} \left(\frac{\lambda_i}{\lambda}\right)^{x_i}$$

which is $Multinoulli(\mathbf{X}|\frac{\lambda_1}{\lambda}, \frac{\lambda_2}{\lambda}, ..., \frac{\lambda_K}{\lambda})$
Now to infer latent variables we will use above lemma,

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

The Gibbs Sampling Algorithm:

1. $\forall n, m, k$ Sample $u_{nk}^{(0)}, v_{mk}^{(0)}$.Set t=1;

2. while $t \le T$,

$\forall n, m$ Sample $\{X_{nmk}\}_{k=1}^{K}$ from $Multinoulli\left((X_{nm1}, X_{nm2}, ..., X_{nmK}) | \left\{\frac{u_{n1}v_{m1}}{\mathbf{u}_n^T\mathbf{v}_m}, ..., \frac{u_{nK}v_{mK}}{\mathbf{u}_n^T\mathbf{v}_m}\right\}^{(t-1)}\right)$

$\forall n, k$ Sample $u_{nk}^{(t)}$ from $Gamma(u_{nk} | \sum_{m=1}^{M} X_{nmk}^{(t)} + a), b + \sum_{m=1}^{M} v_{mk}^{(t-1)})$

$\forall m, k$ Sample $v_{mk}^{(t)}$ from $Gamma\left(v_{mk} | \sum_{n=1}^{N} X_{nmk}^{(t)} + c, d + \sum_{n=1}^{N} u_{nk}^{(t)}\right)$

Set t=t+1