Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 1

Student Name: Abhas Kumar
Roll Number: 20111001
Date: May 17, 2021

Given observations $x_1, x_2 ..., x_N$ drawn i.i.d. from an likelihood model $p(\mathbf{x}|\theta)$, and a prior distribution $p(\theta)$ on the model parameters θ,we need to prove that solving given problem is equivalent to the Bayes rule for finding the posterior distribution of θ. The Given equation is

$$\begin{aligned}
&\arg\min_{q(\theta)} -\sum_{n=1}^{N}\left[\int q(\theta)\log p(\mathbf{x}_n|\theta)d\theta\right] + KL(q(\theta)||p(\theta)) \\
&= \arg\min_{q(\theta)} -\left[\int \sum_{n=1}^{N} q(\theta)\log p(\mathbf{x}_n|\theta)d\theta\right] + KL(q(\theta)||p(\theta)) \\
&= \arg\min_{q(\theta)} -\left[\int q(\theta) \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)d\theta\right] + KL(q(\theta)||p(\theta)) \\
&= \arg\min_{q(\theta)} -\left[\int q(\theta) \log \prod_{n=1}^{N} p(x_n|\theta)d\theta\right] + KL(q(\theta)||p(\theta))
\end{aligned}$$

Now since, $KL(q(\theta)||p(\theta)) = -\int q(\theta)\log\left(\frac{p(\theta)}{q(\theta)}\right)d(\theta)$, we get

$$\arg\min_{q(\theta)} -\int q(\theta)\log\prod_{n=1}^{N} p(x_n|\theta)d\theta - \int q(\theta)\log\left(\frac{p(\theta)}{q(\theta)}\right)d(\theta)$$

$$\begin{aligned}
&= \arg\min_{q(\theta)} -\left[\int q(\theta)\log\left(\frac{\prod_{n=1}^{N} p(x_n|\theta)p(\theta)}{q(\theta)}\right)d\theta\right] \\
&= \arg\min_{q(\theta)} -\left[\int q(\theta)\log\left(\frac{p(\mathbf{X}|\theta)p(\theta)}{q(\theta)}\right)d\theta\right], \text{ where } p(\mathbf{X}|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \\
&= \arg\min_{q(\theta)} \left[KL\Big(q(\theta)||p(\mathbf{X}/\theta)p(\theta)\Big)\right]
\end{aligned}$$

We know that KL divergence is minimum when both distributions are same i.e $q(\theta) \propto p(\mathbf{X}/\theta)p(\theta)$ which can be normalised to obtain a **PDF** integrating to **1**.

$$\begin{aligned}
q(\theta) &\propto p(\mathbf{X}/\theta)p(\theta) \\
\implies q(\theta) &= \frac{p(\mathbf{X}/\theta)p(\theta)}{p(\mathbf{X})} = p(\theta|\mathbf{X})
\end{aligned}$$

which is the **Bayes Rule**, hence solving the given objective function is equivalent to the Bayes rule for finding the posterior distribution of θ.
Intuitively, the objective function tries to find a distribution i.e $q(\theta)$ that explains the data well and is also as close to prior as possible by minimising the $KL(q(\theta)||p(\theta))$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 2

Student Name: Abhas Kumar
Roll Number: 20111001
Date: May 17, 2021

Given N observations $(x_1, y_1), ..., (x_N, y_N)$ generated from a linear regression model $y_n \propto \mathcal{N}(y_n \mid w^T x_n, \beta^{-1})$. Assume a Gaussian prior on w with different component-wise precisions, where we have, $p(w) = \mathcal{N}(w \mid 0, diag(\alpha_1^{-1}, ...\alpha_D^{-1}))$ Also assume gamma priors on the noise precision β and prior's precisions $\{\alpha_d\}_{d=1}^{D}$, i.e $\beta \sim Gamma(\beta|a_0, b_0))$ Given parametrization of the gamma is

$$Gamma(\eta|\tau_1, \tau_2) = \frac{\tau_2^{\tau_1}}{\tau_1}\eta^{(\tau_1-1)}exp(-\tau_2\eta)$$

$$\log p(\mathbf{w}, \mathbf{y}, \beta, \alpha_1^{-1}, ...\alpha_D^{-1} \mid \mathbf{X}) = \log p(\mathbf{y} \mid \mathbf{w}, \beta, \mathbf{X})p(\mathbf{w} \mid \alpha_1, ..., \alpha_D)p(\beta)p(\alpha_1, ..., \alpha_D)$$

$$= \log\left(\prod_{n=1}^{N} p(y_n \mid \mathbf{w}, \mathbf{x}_n, \beta)p(\mathbf{w} \mid \alpha_1, ..., \alpha_D)p(\beta)\prod_{d=1}^{D} p(\alpha_d)\right)$$

$$= \sum_{n=1}^{N} \log p(y_n \mid \mathbf{w}, x_n, \beta) + \log p(\mathbf{w} \mid \alpha_1, ..., \alpha_D) + \log\beta + \sum_{d=1}^{D}\log p(\alpha_d)$$

$$= \sum_{n=1}^{N} log\left(\sqrt{\frac{\beta}{2\pi}}exp(\frac{-\beta}{2}(y_n - \mathbf{w}^T\mathbf{x}_n)^2)\right) + \log\left(\sqrt{\frac{\alpha_1, ..., \alpha_D}{(2\pi)^D}}exp\left(\frac{-\mathbf{w}^T diag(\alpha_1, ..., \alpha_D)\mathbf{w}}{2}\right)\right)$$

$$+ \log\left(\frac{b_0^{a_0}}{\tau(a_0)}\beta^{a_0-1}exp(-b_0\beta)\right) + \sum_{d=1}^{D}\log\left(\frac{f_0^{e_0}}{\tau(e_0)}\alpha_d^{e_0-1}exp(-f_0\alpha_d)\right)$$

$$\propto \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \frac{1}{2}\sum_{d=1}^{D}\log\alpha_d - \frac{1}{2}w^T\sum\nolimits_w + (a_0 - 1)\log\beta - b_0\beta$$

$$+(e_0 - 1)\sum_{d=1}^{D}\log\alpha_d - f_0\sum_{d=1}^{D}\alpha_d$$

- To estimate $\mathbf{w}$, we can write $\log q_{\mathbf{w}}^*(\mathbf{w})$ as following after ignoring the constants.

$$= E_{q_\beta, \alpha_1, ...\alpha_D}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D \mid \mathbf{X})\right]$$

$$= E\left[-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 - \frac{1}{2}\mathbf{w}^T diag(\alpha_1, ..., \alpha_D)\mathbf{w}\right]$$

$$= \frac{-1}{2}\left\{\mathbf{w}^T\left(E[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + diag(E[\alpha_1], ..., E[\alpha_2])\right)\mathbf{w} - 2\mathbf{w}^T E[\beta]\sum_{n=1}^{N} y_n\mathbf{x}_n\right\}$$

Above equation has a Gaussian form, where we can find

$$Mean = \boxed{\mu_w = \left(E[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + diag(E[\alpha_1], ..., E[\alpha_2]\right)^{-1} E[\beta]\sum_{n=1}^{N} y_n\mathbf{x}_n} \quad (1)$$

$$Covariance = \boxed{\Lambda_w = \left(E[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + diag(E[\alpha_1], ..., E[\alpha_2]\right)^{-1}} \quad (2)$$

Therefore, w has Gaussian form

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mu_w, \Lambda_w)$$

• Similarly, to estimate $\{\alpha_d\}_{d=1}^{D}$, we can write $\log q^*_{\alpha_d}(\alpha_d)$ as following after ignoring the constants.

$$= E\left[\frac{\log \alpha_d}{2} - \frac{{w_d}^2\alpha_d}{2} + (l_0 - 1)\log \alpha_d - m_0\alpha_d\right]$$
$$= \left(\frac{1}{2} + l^0 - 1\right)\log \alpha_d - \alpha_d\left(m_0 + \frac{E[w_d^2]}{2}\right)$$

Therefore, as β, α_d also has a **Gamma** form i.e

$$\alpha_d \sim \textbf{Gamma}(\alpha_d|l_d, m_d)$$

where,

$$\boxed{l_d = l_0 + \frac{1}{2}}$$

$$\boxed{m_d = m_0 + \frac{E[w_d^2]}{2}}$$

• Again, to estimate β, we can write $\log q^*_\beta(\beta)$ as following after ignoring the constants.

$$= E\left[\frac{N}{2}\log \beta - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + (\log \beta) * (a_0 - 1) - b_0\beta\right]$$
$$= (\log \beta) * (\frac{N}{2} + a_0 - 1) - \beta\left(\sum_{n=1}^{N}\frac{N}{2}E[(y_n - \mathbf{w}^T\mathbf{x}_n)^2]\right) + b_0$$

Above equation has a **Gamma** form, where we can find

$$\beta \sim \textbf{Gamma}(\beta|a_0, b_0))$$

where,

$$\boxed{a = a_0 + \frac{N}{2}}$$

$$\boxed{b = b_0 + \sum_{n=1}^{N}\frac{1}{2}E[(y_n - \mathbf{w}^T\mathbf{x}_n)^2]}$$

Mean-field VI algorithm:

1. Initialize $t = 1$, l_d $\forall d$ and b, also $m_d = m_0 + \frac{1}{2}$ and $a = a_0 + \frac{N}{2}$
2. Calculate following expectations

$$E[\mathbf{w}] = \mu_w$$
$$E[\mathbf{w}\mathbf{w}^T] = \Lambda_w + \mu_w \mu_w^T$$
$$E[w_d^2] = \Lambda_{w_{dd}} + \mu_{w_d}^2$$
$$E[\beta] = \frac{a}{b}$$
$$E[\alpha_d] = \frac{l_d}{m_d} \forall d$$

3. Repeat until not converged

   4. Calculate μ_w using equation 1
   5. Calculate Λ_w using equation 2
   6. Calculate $b = \sum_{n=1}^{N} \frac{1}{2} E[(y_n - w^T.x_n)^2] + b_0$
   7. Calculate $E[\beta] = \dfrac{a}{b}$ and $E[\alpha_d] = \dfrac{l_d}{m_d} \forall d$
   8. Calculate $m_d = m_0 + \dfrac{E[w_d^2]}{2}$ $\forall d$
   9. $t = t + 1$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 3

Student Name: Abhas Kumar
Roll Number: 20111001
Date: May 17, 2021

We are given a bunch of count-valued observations $x_1, x_2..., x_N$, generated from the following hierarchical model:

$p(x_n|\lambda_n) = \textbf{Poisson}(x_n|\lambda_n)$,
$p(\lambda_n|\alpha,\beta) = \textbf{Gamma}(\lambda_n|\alpha,\beta)$, where $n = 1, 2..., N$
$p(\alpha|a,b) = \textbf{Gamma}(\alpha|a,b)$, and
$p(\beta|c,d) = \textbf{Gamma}(\beta|c,d)$. where a, b, c, d are fixed.

Joint Probability Distribution to find the conditional posteriors,

$$\begin{aligned} p(\mathbf{X},\lambda,\alpha,\beta,a,b,c,d) &= \prod_{n=1}^{N}\Big(p(x_n|\lambda_n)\ p(\lambda_n|\alpha,\beta)\Big)p(\alpha|a,b)p(\beta|c,d) \\ &= \prod_{n=1}^{N}\Big(\textbf{Poisson}(x_n|\lambda_n)\ \textbf{Gamma}(\lambda_n|\alpha,\beta)\Big)\textbf{Gamma}(\alpha|a,b)\textbf{Gamma}(\beta|c,d) \end{aligned}$$

To do Gibbs sampling for this model, we need to derive the conditional posterior (CP) of each variable $\lambda_1, \lambda_2..., \lambda_N, \alpha$, and β, using the markov blanket of that parameter.

- Conditional Posterior for λ_n

$$p(x_n|\lambda_n) * p(\lambda_n|\alpha,\beta) = \textbf{Possion}(x_n|\alpha,\beta) * \textbf{Gamma}(\lambda_n|\alpha,\beta)$$

$$= \left(\frac{\lambda_n^{x_n}\mathbf{e}^{-\lambda_n}}{x_n!}\right) * \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_n^{\alpha-1}\mathbf{e}^{-\beta\lambda_n}\right)$$

$$\propto \left(\lambda_n^{x_n}\mathbf{e}^{-\lambda_n}\right)\left(\lambda_n^{\alpha-1}\mathbf{e}^{-\beta\lambda_n}\right)$$

$$\propto \lambda_n^{(x_n+\alpha-1)}\mathbf{e}^{(-\lambda_n(\beta+1))}$$

$$\implies p(\lambda_n|\alpha,\beta) = \textbf{Gamma}(x_n+\alpha, \beta+1)$$

****Conditional Posterior for $\lambda_1, \lambda_2, ..., \lambda_N$ has closed form**.

- Conditional Posterior for α

$$\begin{aligned}
p(\alpha|\mathbf{X},\lambda,\beta,a,b) &= \frac{p(\lambda_n|\alpha,\beta)p(\alpha|a,b)}{p(\lambda|\mathbf{X},\beta)} \\
&\propto \prod_{n=1}^{N}\Big(\mathbf{Gamma}(\lambda_n|\alpha,\beta)\Big)\mathbf{Gamma}(\alpha|a,b) \\
&\propto \prod_{n=1}^{N}\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda_n^{\alpha-1}\mathbf{e}^{-\beta\lambda_n}\right)\left(\alpha^{a-1}\ e^{-\alpha b}\right) \\
&\propto \frac{\alpha^{a-1}\ \beta^{N\alpha}}{(\Gamma(\alpha))^{N}}\ e^{\left(-\alpha b\ -\ \sum_{n=1}^{N}\lambda_n\beta\right)}\ \prod_{n=1}^{N}\lambda_n^{\alpha-1}
\end{aligned}$$

****Conditional Posterior for α has no closed form**.

- Conditional Posterior for β

$$\begin{aligned}
p(\beta|\mathbf{X},\lambda,\alpha) &= \frac{p(\lambda_n|\alpha,\beta)p(\beta|c,d)}{p(\lambda|\mathbf{X},\alpha)} \\
&\propto \prod_{n=1}^{N}p(\lambda_n|\alpha,\beta)p(\beta|c,d) \\
&\propto \prod_{n=1}^{N}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda_n^{\alpha-1}\mathbf{e}^{-\beta\lambda_n}\times\frac{d^{c}}{\Gamma(c)}\beta^{c-1}\mathbf{e}^{-d\beta}
\end{aligned}$$

****Conditional Posterior for β has no closed form**.

Only λ_n has a closed form CP out of λ, α and β was only with $p(\lambda_n|\alpha,\beta) = \mathbf{Gamma}(\mathbf{x}_n + \alpha, \beta + 1)$.Hence we can only do Gibbs Sampling for λ_n.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 4

Student Name: Abhas Kumar
Roll Number: 20111001
Date: May 17, 2021

The posterior predictive distribution of each r_{ij} is given as

$$p(r_{ij} \mid R) \;=\; \int p(r_{ij} \mid u_i, v_j)\; p(u_i, v_j \mid R)\; du_i\; dv_j$$

We are given a set of S samples $\{U^{(s)}, V^{(s)}\}_{s=1}^{S}$ generated by a Gibbs sampler for this matrix factorization model, where $U^{(s)} = \{u_i\}_{i=1}^{N}$ and $V^{(s)} = \{v_i\}_{j=1}^{M}$.

The **PPD** $p(r_{ij} \mid R)$ can be approximated using sampling based approximation as

$$\begin{aligned} p(r_{ij} \mid R) \;&\approx\; \frac{1}{S} \sum_{s=1}^{S} p\left(r_{ij} \mid u_i^{(s)}, v_j^{(s)}\right) \\ &\approx\; \frac{1}{S} \sum_{s=1}^{S} \mathcal{N}\left(r_{ij} \mid u_i^{(s)^T} v_j^{(s)}, \beta^{-1}\right) \end{aligned}$$

Then Expectation of r_{ij} can be found using,

$$\begin{aligned} E[r_{ij}] \;&=\; \int r_{ij}\left(\frac{1}{S} \sum_{s=1}^{S} \mathcal{N}\left(r_{ij} \mid u_i^{(s)^T} v_j^{(s)}, \beta^{-1}\right)\right) dr_{ij} \\ &=\; \frac{1}{S} \sum_{s=1}^{S} \int r_{ij}\mathcal{N}\left(r_{ij} \mid u_i^{(s)^T} v_j^{(s)}, \beta^{-1}\right) dr_{ij} \\ &=\; \frac{1}{S} \sum_{s=1}^{S} \left(u_i^{(s)^T} v_j^{(s)}\right) \end{aligned}$$

Similarly, $E[r_{ij}^2]$ can be calculated using

$$\begin{aligned} E[r_{ij}^2] \;&=\; \int r_{ij}^2\left(\frac{1}{S} \sum_{s=1}^{S} \mathcal{N}\left(r_{ij} \mid u_i^{(s)^T} v_j^{(s)}, \beta^{-1}\right)\right) dr_{ij} \\ &=\; \int r_{ij}^2\left(\frac{1}{S} \mathcal{N}\left(r_{ij} \mid \sum_{s=1}^{S} u_i^{(s)^T} v_j^{(s)}, S * \beta^{-1}\right)\right) dr_{ij} \\ &=\; \frac{1}{S}\sum_{s=1}^{S} \left(u_i^{(s)^T} v_j^{(s)}\right)^2 + \beta^{-1} \end{aligned}$$

Now we can calculate the variance using $E[r_{ij}^2], (E[r_{ij}])^2$ as the Variance,

$$\begin{aligned} Var[r_{ij}] \;&=\; E[r_{ij}^2] - (E[r_{ij}])^2 \\ &=\; \frac{1}{S}\sum_{s=1}^{S} \left(u_i^{(s)^T} v_j^{(s)}\right)^2 + \beta^{-1} - \frac{1}{S^2}\left(\sum_{s=1}^{S} \left(u_i^{(s)^T} v_j^{(s)}\right)\right)^2 \end{aligned}$$

Rejection Sampling

We want to use Rejection Sampling to sample from $p(x)$ and using a proposal distribution $q(x) = N(x|0, \sigma^2)$.We need to figure out the optimal value of the constant M such that $M * q(z) \geq \tilde{p}(x)$, as required in Rejection Sampling. Using this value of M and some suitably chosen σ^2 , we need to draw 10,000 samples from $p(x)$ distribution.

Finding optimal value of the constant M such that $M * q(z) \geq \tilde{p}(x)$

$$M \geq max_x \frac{\tilde{p}(x)}{q(x)}$$

$$\therefore M \geq max_x \frac{exp(sin(x))}{\mathcal{N}(x \mid 0,\ \sigma^2)}$$

Putting the value of Gaussian Distribution in the above equation,

$$\therefore M \geq max_x \frac{exp(sin(x))}{\frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-\frac{1}{2}\left(\frac{x-0}{\sigma}\right)^2}}$$

$$\therefore M \geq max_x \sqrt{2\pi}\ \sigma\ exp\left(sin(x) + \frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)$$

Now, in order to maximize the above quantity, we need to maximize the quantity in exponential term i.e., $sin(x) + \frac{1}{2}\left(\frac{x}{\sigma}\right)^2$.

$$\therefore M \geq \sqrt{2\pi}\ \sigma\ exp\left(1 + \frac{\pi^2}{2\sigma^2}\right)$$

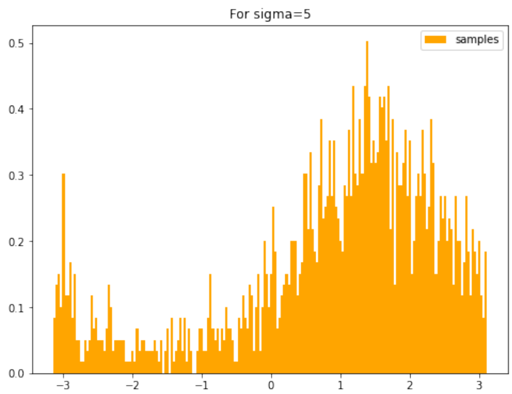

Figure 1: Plot showing the resulting histogram of the samples for suitably chosen $\sigma = 5$

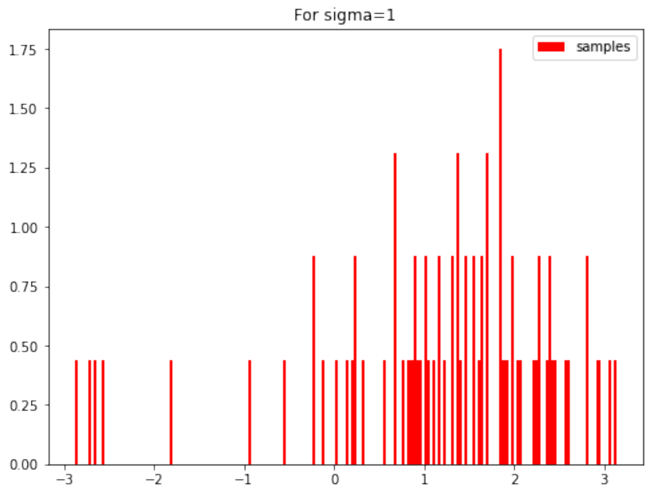

Figure 2: Plot showing the resulting histogram of the samples for low value of $\sigma = 1$

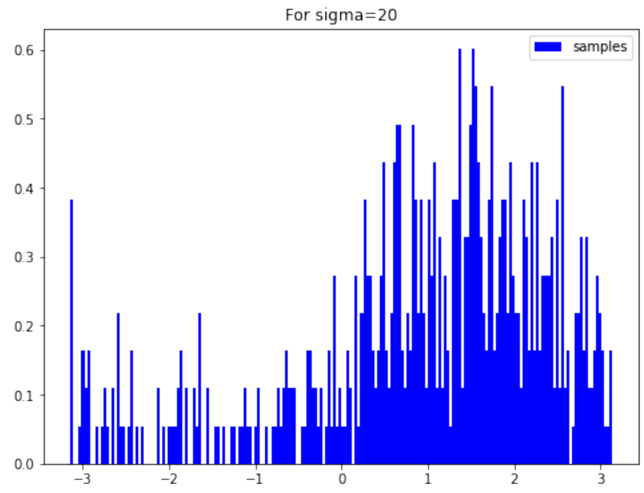

Figure 3: Plot showing the resulting histogram of the samples for high value of $\sigma = 20$