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

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: May 14, 2021

QUESTION 1

This is Bayes Rule!
Given:

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

$$= \arg\max_{q(\theta)} \int q(\theta)\log p(\mathbf{X}|\theta)d\theta + \int q(\theta)\log\frac{p(\theta)}{q(\theta)}d\theta \tag{2}$$

$$= \arg\max_{q(\theta)} \underbrace{\int q(\theta)\log\frac{p(\mathbf{X},\theta)}{q(\theta)}}_{\mathcal{L}(q)}d\theta \tag{3}$$

The identity we know

$$\log\ p(\mathbf{X}) = \mathcal{L}(q) + KL(q(\theta)\|p(\theta|X)) \tag{4}$$

As the log $p(\mathbf{X})$ is constant wrt to θ

$$\arg\max_{q(\theta)} \mathcal{L}(q) = \arg\min_{q} KL(q(\theta)\|p(\theta|X)) \tag{5}$$

$KL(q(\theta)\|p(\theta|X))$ will attain its minimum value when $q(\theta) = p(\theta|X)$
Therefore by minimizing the given equation we can get the posterior of parameters.
Intitution of given equation

1. The first term is one such a way that best explains the data
2. Second term is regularizer makes sure that $q(\theta)$ close to prior $p(\theta)$

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: May 14, 2021

QUESTION 2

Mean-Field VI for Sparse Bayesian Linear Regression
Given

$$y_n \sim \mathcal{N}(y_n|\omega^T \mathbf{x_n}, \beta^{-1}) \tag{6}$$
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\omega}, \beta^{-1}\mathbf{I}) \tag{7}$$
$$p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}|0, \boldsymbol{\alpha}^{-1}) \text{ where } \boldsymbol{\alpha}^{-1} = diag(\alpha_1^{-1}, ...\alpha_D^{-1}) \tag{8}$$
$$\omega_d \sim \mathcal{N}(\omega_d|0, \alpha_d) \tag{9}$$
$$\beta \sim Gamma(\beta|a_0, b_0) \tag{10}$$
$$\alpha_d \sim Gamma(\alpha_d|e_0, f_0) \tag{11}$$

The joint probability can be written as

$$\log p(\mathbf{y}, \boldsymbol{\omega}, \beta, \alpha_1, ...\alpha_D|\mathbf{X}) = \log p(\mathbf{y}|\beta, \boldsymbol{\omega}\mathbf{X}) + \log p(\boldsymbol{\omega}|\alpha_1, ..\alpha_D) + \log p(\beta) + \sum_{d=1}^{D} \log p(\alpha_d) \tag{12}$$

Now update for $\boldsymbol{\omega}$

$$\log q^*_{\boldsymbol{\omega}}(\boldsymbol{\omega}) = \mathbb{E}_{\beta,\alpha_1...\alpha_D}[\log p(\mathbf{y}, \boldsymbol{\omega}, \beta, \alpha_1, ...\alpha_D|\mathbf{X})] \tag{13}$$
$$= \mathbb{E}_{\beta,\alpha_1...\alpha_D}[\log p(\mathbf{y}|\mathbf{X}\boldsymbol{\omega}, \beta) + \log p(\boldsymbol{\omega}|\alpha_1, ..\alpha_D)] + \text{const} \tag{14}$$
$$= \mathbb{E}_{\beta,\alpha_1...\alpha_D}[\log \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\omega}, \beta^{-1}\mathbf{I}) + \log \mathcal{N}(\boldsymbol{\omega}|0, \boldsymbol{\alpha})] \tag{15}$$
$$\text{using gaussian properties} \tag{16}$$
$$q^*_{\boldsymbol{\omega}}(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}) \tag{17}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = (\mathbb{E}[\boldsymbol{\alpha}] + \mathbb{E}[\beta]\mathbf{X^T X})^{-1} \tag{18}$$
$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \mathbb{E}[\beta]\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\mathbf{X^T y} \tag{19}$$
$$\mathbb{E}[\boldsymbol{\alpha}] = diag(\mathbb{E}[\alpha_1], ..\mathbb{E}[\alpha_D]) \tag{20}$$

update for β

$$\log q^*_{\beta}(\beta) = \mathbb{E}_{\boldsymbol{\omega},\alpha_1,..\alpha_D}[\log p(\mathbf{y}|\mathbf{X}\boldsymbol{\omega}, \beta) + \log p(\beta|a_0, b_0)] + \text{const} \tag{21}$$
$$= \mathbb{E}_{\boldsymbol{\omega},\alpha_1,..\alpha_D}[\log \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\omega}, \beta^{-1}\mathbf{I}) + \log Gamma(\beta|a_0, b_0)] \tag{22}$$
$$\text{by conjugacy} \tag{23}$$
$$q^*_{\beta}(\beta) = Gamma(\beta|a_\beta, b_\beta) \tag{24}$$
$$a_\beta = \frac{N}{2} + a_0 \tag{25}$$
$$b_\beta = \mathbb{E}[\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\omega}\|^2] + b_0 \tag{26}$$
$$= \frac{\mathbf{y^T y} - \mathbf{2y^T X}\mathbb{E}[\boldsymbol{\omega}] + Tr(\mathbf{X^T X}\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathbf{T}])}{2} + b_0 \tag{27}$$

Now update for α_d

$$
\begin{aligned}
\log q^*_{\alpha_d}(\alpha_d) &= \mathbb{E}_{\boldsymbol{\omega},\beta,\alpha_{i\neq d}}[\log p(\omega_d|0,\alpha_d) + \log p(\alpha_d|e_0, f_0)] + \text{const} && (28)\\
&= \mathbb{E}_{\boldsymbol{\omega},\beta,\alpha_{i\neq d}}[\log \mathcal{N}(\omega_d|0,\alpha_d) + \log Gamma(\alpha_d|e_0, f_0)] && (29)\\
&\text{by conjugacy} && (30)\\
q^*_{\alpha_d}(\alpha_d) &= Gamma(\alpha_d|e_{\alpha_d}, f_{\alpha_d}) && (31)\\
e_{\alpha_d} &= \frac{1}{2} + e_0 && (32)\\
f_{\alpha_d} &= \frac{\mathbb{E}[\omega_d^2]}{2} + f_0 && (33)
\end{aligned}
$$

We following expectations are required for above calculations

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\omega}] &= \boldsymbol{\mu_\omega} && (34)\\
\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T] &= \boldsymbol{\Sigma_\omega} + \boldsymbol{\mu_\omega}\boldsymbol{\mu_\omega}^T && (35)\\
\mathbb{E}[\omega_d^2] &= [\boldsymbol{\Sigma_\omega}]_{dd} + [\boldsymbol{\mu_\omega}\boldsymbol{\mu_\omega}^T]_{dd}\ \forall d && (36)\\
\mathbb{E}[\alpha_d] &= \frac{e_{\alpha_d}}{f_{\alpha_d}}\ \forall d && (37)\\
\mathbb{E}[\beta] &= \frac{a_\beta}{b_\beta} && (38)
\end{aligned}
$$

Mean Field VI Algorithm

1. Set $a_\beta = a_0 + \frac{N}{2}$ and $e_{\alpha_d} = e_0 + \frac{1}{2}$
2. Set t=0. Initialize, ${\boldsymbol{\Sigma_w}}^{(0)} = \mathbf{I}_D$ and ${\boldsymbol{\mu_w}}^{(0)} = 0$ so,

$$
\begin{aligned}
f^{(0)}_{\alpha_d} &= f_0 + \frac{1}{2}\\
{b_\beta}^{(0)} &= b_0 + \frac{\mathbf{y}^T\mathbf{y} + Tr(\mathbf{X}^T\mathbf{X})}{2}\\
\mathbb{E}[\beta]^{(0)} &= \frac{a_\beta}{{b_\beta}^{(0)}}\\
\mathbb{E}[\alpha_d]^{(0)} &= \frac{e_{\alpha_d}}{f^{(0)}_{\alpha_d}}, \forall d
\end{aligned}
$$

3. Set t=t+1. Until convergence, repeat.

$$
\begin{aligned}
\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\omega}} &= (\mathbb{E}[\boldsymbol{\alpha}]^{(t-1)} + \mathbb{E}[\beta]^{(t-1)}\mathbf{X^T X})^{-1}\\
\boldsymbol{\mu}^{(t)}_{\boldsymbol{\omega}} &= \mathbb{E}[\beta]^{(t-1)}{\boldsymbol{\Sigma_\omega}}^{(t)}\mathbf{X^T y}\\
{b_\beta}^{(t)} &= \frac{\mathbf{y^T y - 2y^T X}\mathbb{E}[\boldsymbol{\omega}]^{(\mathbf{t})} + Tr(\mathbf{X^T X}\mathbb{E}[\boldsymbol{\omega\omega}^\mathbf{T}]^{(\mathbf{t})})}{2} + b_0\\
f^{(t)}_{\alpha_d} &= \frac{\mathbb{E}[\omega_d^2]^{(t)}}{2} + f_0, \forall d\\
\mathbb{E}[\beta]^{(t)} &= \frac{a_\beta}{{b_\beta}^{(t)}}\\
\mathbb{E}[\alpha_d]^{(t)} &= \frac{e_{\alpha_d}}{f_{\alpha_d^{(t)}}}, \forall d
\end{aligned}
$$

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

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: May 14, 2021

QUESTION 3

Gibbs Sampling
Given

$$x_1, ..x_N \sim Poisson(x_n|\lambda_n) \tag{39}$$
$$\lambda_n \sim Gamma(\lambda_n|\alpha, \beta) \tag{40}$$
$$\alpha \sim Gamma(\alpha|a, b) \tag{41}$$
$$\beta \sim Gamma(\beta|c, d) \tag{42}$$

To perform the gibbs sampling. of posteriror we need CP's. Finding the conditional posteriors (CP)
wrt λ_n $\forall n$

$$p(\lambda_n|\lambda_{-n}, \alpha, \beta, X) \propto p(x_n|\lambda_n)p(\lambda_n|, \alpha, \beta)) \tag{43}$$
$$= Poisson(x_n|\lambda_n)Gamma(\lambda_n|\alpha, \beta) \tag{44}$$
$$\text{by conjugacy} \tag{45}$$
$$p(\lambda_n|\lambda_{-n}, \alpha, \beta, x_n) = Gamma(\lambda_n|\alpha_{\lambda_n}, \beta_{\lambda_n}) \tag{46}$$
$$\alpha_{\lambda_n} = x_n + \alpha \tag{47}$$
$$\beta_{\lambda_n} = \beta + 1 \tag{48}$$

wrt α

$$p(\alpha|\lambda_1, ..., \lambda_N, , \beta, X) \propto \prod_{n=1}^{N} p(\lambda_n|\alpha, \beta)p(\alpha|a, b) \tag{49}$$
$$\propto \prod_{n=1}^{N} Gamma(\lambda_n|\alpha, \beta)Gamma(\alpha|a, b) \tag{50}$$
$$\propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^N [\prod_{n=1}^{N} \lambda_n]^{\alpha-1} exp(-\beta \sum_{n=1}^{N} \lambda_n - b\alpha)\alpha^{a-1} \tag{51}$$

The above equation is not in the closed form wrt α

$$p(\beta|\lambda_1, ..., \lambda_N, , \beta, X) \propto \prod_{n=1}^{N} p(\lambda_n|\alpha, \beta)p(\beta|c, d) \tag{52}$$
$$\propto \prod_{n=1}^{N} Gamma(\lambda_n|\alpha, \beta)Gamma(\beta|c, d) \tag{53}$$
$$\text{by conjugacy} \tag{54}$$
$$p(\beta|\lambda_1, ..., \lambda_N, , \beta, X) = Gamma(c + N\alpha, d + \sum_{n=1}^{N} \lambda_n) \tag{55}$$

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: May 14, 2021

QUESTION 4

Using Samples for Prediction

Consider matrix factorization model $N \times M$

$$\text{likelihood}: \quad p(r_{ij}|\mathbf{u_i}, \mathbf{v_j}) = \mathcal{N}(r_{ij}|\mathbf{u_i^T v_j}, \beta^{-1}) \tag{56}$$

$$\text{ppd}: \quad p(r_{ij}|R) = \int p(r_{ij}|\mathbf{u_i}, \mathbf{v_j}) p(\mathbf{u_i}, \mathbf{v_j}|R) d\mathbf{u}_i d\mathbf{v}_j \tag{57}$$

Given the samples $\{\mathbf{U^{(s)}}, \mathbf{V^{(s)}}\}_{s=1}^S$ from gibbs sampler. PPD will can written as:

$$p(r_{ij}|R) \approx \frac{1}{S} \sum_{s=1}^{S} p(r_{ij}|\mathbf{u_i^{(s)}}, \mathbf{v_j^{(s)}}) \tag{58}$$

Now calculating expectation of r_{ij}

$$\mathbb{E}_{r_{ij}|R}[r_{ij}] = \int r_{ij} p(r_{ij}|R)\ dr_{ij} \tag{59}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \int r_{ij} \mathcal{N}(r_{ij}|\mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}}, \beta^{-1})\ dr_{ij} \tag{60}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}[\mathcal{N}(r_{ij}|\mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}}, \beta^{-1})] \tag{61}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}} \tag{62}$$

Now calculating expectation of r_{ij}

$$var_{r_{ij}|R}(r_{ij}) = \mathbb{E}[r_{ij}^2] - [\mathbb{E}[r_{ij}]]^2 \tag{63}$$

$$\mathbb{E}_{r_{ij}|R}[r_{ij}] = \int r_{ij}^2 p(r_{ij}|R)\ dr_{ij} - (\frac{1}{S} \sum_{s=1}^{S} \mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}})^2 \tag{64}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \int r_{ij}^2 \mathcal{N}(r_{ij}|\mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}}, \beta^{-1})\ dr_{ij} - (\frac{1}{S} \sum_{s=1}^{S} \mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}})^2 \tag{65}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \left(\beta^{-1} + (\boldsymbol{u_i^{T(s)}} \boldsymbol{v_j})^2\right) - (\frac{1}{S} \sum_{s=1}^{S} \mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}})^2 \tag{66}$$

$$= \beta^{-1} + \frac{1}{S} \sum_{s=1}^{S} (\mathbf{u_i^{T(s)}} \mathbf{v_j})^2 - (\frac{1}{S} \sum_{s=1}^{S} \mathbf{u_i^{T(s)}} \mathbf{v_j^{(s)}})^2 \tag{67}$$

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

Student Name: Musale Krushna Pavan
Roll Number: 20111268
Date: May 14, 2021

QUESTION 5

Rejection Sampling
Given

$$p(x) \propto exp(\sin(x)) \qquad -\pi \leq x \leq \pi \tag{68}$$

$$\text{poposal distribution for rejection sampling} \tag{69}$$

$$q(x) = \mathcal{N}(x|0, \sigma^2) \tag{70}$$

For doing rejection sampling we need the following condition

$$Mq(x) \geq \tilde{p}(x) \qquad \forall -\pi \leq x \leq \pi \tag{71}$$

$$M\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{x^2}{2\sigma^2}) \geq exp(\sin(x)) \tag{72}$$

$$M \geq \frac{\sigma\sqrt{2\pi}\ exp(\sin(x))}{exp(-\frac{x^2}{2\sigma^2})} \tag{73}$$

$$\text{Approximating max value of M for } \forall -\pi \leq x \leq \pi \tag{74}$$

$$\max exp(\sin(x)) = e \tag{75}$$

$$\min exp(-\frac{x^2}{2\sigma^2}) = \frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{\pi^2}{2\sigma^2}) \tag{76}$$

$$M \geq \sigma\sqrt{2\pi}exp(\frac{\pi^2}{2\sigma^2} + 1) \tag{77}$$

considering the $\sigma^2 = 1$ we get $M = \sqrt{2\pi}exp(\frac{\pi^2}{2} + 1) = 947.4183259814814$

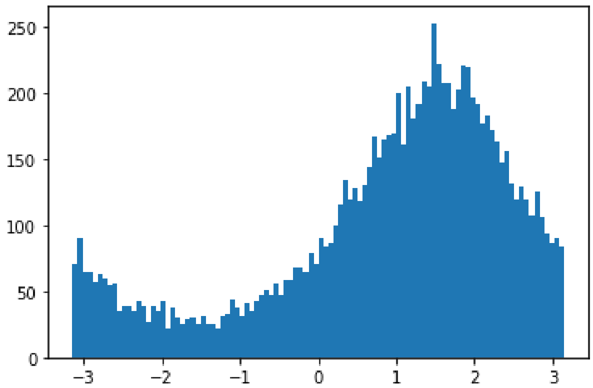

Figure 1: Hist of samples from p(x)