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

QUESTION 1

Student Name: Shashank Shailabh
Roll Number: 170655
Date: May 14, 2021

$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \sim p(\mathbf{x}|\theta)$ and prior distribution is $p(\theta)$.
Posterior distribution for a model m is calculated using log marginal-likelihood.

$$\log p(\mathbf{x}|m) = \int q(\theta) \log \left\{ \frac{p(\mathbf{x}, \theta)}{q(\theta)} \right\} d\theta + KL(q(\theta)||p(\theta|\mathbf{x}))$$

Since log marginal-likelihood is constant w.r.t θ so

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

This shows that the above expression is same as Bayes rule for finding posterior distribution of θ.
Intuitively, the above objective function is the lower bound of the original function so maximizing will give the posterior distribution of θ. KL term in forces the $q(\theta)$ to be similar to prior which is simple.

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: May 14, 2021

QUESTION 2

Using mean-field variational assumption, finding joint distribution-

$$p(\mathbf{w}, \alpha_1, .., \alpha_D, \beta, \mathbf{y}|\mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha_1, .., \alpha_D)p(\alpha_1, .., \alpha_D)p(\beta)$$

$$\log p(\mathbf{w}, \alpha_1, .., \alpha_D, \beta, \mathbf{y}|\mathbf{X}) = \sum_{n=1}^{N} \log p(y_n|\mathbf{x}_n, \mathbf{w}, \beta) + \log p(\mathbf{w}|\alpha_1, .., \alpha_D) + \sum_{d=1}^{D} \log p(\alpha_d) + \log p(\beta)$$

$$\begin{aligned}\log p(\mathbf{w}, \alpha_1, .., \alpha_D, \beta, \mathbf{y}|\mathbf{X}) \quad = \quad & \textstyle\sum_{n=1}^{N} \log \mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n, \beta^{-1}) \; + \; \log \mathcal{N}(\mathbf{w}|0, diag(\alpha_1^{-1}, .., \alpha_D^{-1})) \\ & + \textstyle\sum_{d=1}^{D} \log Gamma(\alpha_d|e_0, f_0) + \log Gamma(\beta|a_0, b_0)\end{aligned}$$

To calculate the the q distribution w.r.t a particular parameter, ignore all other parameters -

Derivation of $q(\mathbf{w})$ - Taking only terms corresponding to w-

$$q(\mathbf{w}) \propto e^{\mathbb{E}_{q(\beta),q(\alpha_1),..,\alpha_D}[\sum_{n=1}^{N} \log \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \beta^{-1}) + \log \mathcal{N}(\mathbf{w}|0, diag(\alpha_1^{-1}, .., \alpha_D^{-1}))]}$$

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

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

$$q(\mathbf{w}) \propto e^{\frac{-1}{2}[-2\sum_{n=1}^{N}(\mathbb{E}_{q(\beta)}[\beta]y_n\mathbf{x}_n^T)\mathbf{w} + \mathbf{w}^T(\mathbb{E}_{q(\beta)}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + diag(\mathbb{E}_{q(\alpha_1)}[\alpha_1], .., \mathbb{E}_{q(\alpha_D)}[\alpha_D]))\mathbf{w}]}$$

Now, it can seen that $q(\mathbf{w})$ is gaussian distribution.
Therefore,

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

where,

$$\boldsymbol{\Sigma}_N = [\mathbb{E}_{q(\beta)}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + diag(\mathbb{E}_{q(\alpha_1)}[\alpha_1], .., \mathbb{E}_{q(\alpha_D)}[\alpha_D])]^{-1}$$

$$\boldsymbol{\mu}_N = \mathbb{E}_{q(\beta)}[\beta]\boldsymbol{\Sigma}_N(\sum_{n=1}^{N} y_n\mathbf{x}_n)$$

Derivation of $q(\alpha_d)$ - only terms corresponding to α_d $\qquad \forall d = 1, .., D$

$$q(\alpha_d) \propto e^{[\mathbb{E}_{q(\mathbf{w})}[\log \mathcal{N}(\mathbf{w}|diag(\alpha_1^{-1}, .., \alpha_D^{-1})) + \log Gamma(\alpha_d|e_0, f_0)]]}$$

$$q(\alpha_d) \propto e^{[\mathbb{E}_{q(\mathbf{w})}[\frac{1}{2}\sum_{d=1}^{D}(\log \alpha_d - \alpha_d w_d^2) + (e_0 - 1)\log \alpha_d - f_0\alpha_d]]}$$

$$q(\alpha_d) \propto e^{[\mathbb{E}_{q(\mathbf{w})}[(e_0 - \frac{1}{2})\log \alpha_d - (f_0 + \frac{w_d^2}{2}\alpha_d)]]}$$

$$q(\alpha_d) \propto e^{[(e_0 + \frac{1}{2})\log \alpha_d - (f_0 + \frac{1}{2}\mathbb{E}_{q(\mathbf{w})}[w_d^2])\alpha_d]}$$

Therefore,

$$q(\alpha_d) = Gamma(\alpha_d|e_0 + \frac{1}{2}, f_0 + \frac{1}{2}\mathbb{E}_{q(\mathbf{w})}[w_d^2]) \qquad \forall d = 1,..,D$$

where $\mathbb{E}_{q(\mathbf{w})}[w_d^2] = (\boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T)_{d,d}$
Derivation for $q(\beta)$ -

$$q(\beta) \propto e^{[\mathbb{E}_{q(\mathbf{w})}[\sum_{n=1}^{N}\log\mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n,\beta^{-1})+\log Gamma(\beta|a_0,b_0)]]}$$

$$q(\beta) \propto e^{\mathbb{E}_{q(\mathbf{w})}[(\frac{N}{2}+a_0-1)\log\beta-(b_0+\frac{1}{2}\sum_{n=1}N(y_n-\mathbf{w}^T\mathbf{x}_n)^2)\beta]}$$

$$q(\beta) \propto e^{\mathbb{E}_{q(\mathbf{w})}[(\frac{N}{2}+a_0-1)\log\beta-(b_0+\frac{1}{2}\sum_{n=1}N(y_n^2+\mathbf{x}_n^T\mathbb{E}_{q(\mathbf{w})}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n-2y_n\mathbb{E}_{q(\mathbf{w})}[\mathbf{w}]^T\mathbf{x}_n))\beta]}$$

Therefore,

$$q(\beta) = Gamma(\beta|a_0 + \frac{N}{2}, b_0 + \frac{1}{2}\sum_{n=1} N(y_n^2 + \mathbf{x}_n^T\mathbb{E}_{q(\mathbf{w})}[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n - 2y_n\mathbb{E}_{q(\mathbf{w})}[\mathbf{w}]^T\mathbf{x}_n))$$

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: May 14, 2021

QUESTION 3

$$p(x_n|\lambda_n) = Poisson(x_n|\lambda_n) = \frac{\lambda_n^{x_n} e^{-\lambda_n}}{x_n!} \qquad \forall \ \ n = 1, ..., N$$

$$p(\lambda_n|\alpha,\beta) = Gamma(\lambda_n|\alpha,\beta) = \frac{\beta^\alpha \lambda_n^{\alpha-1} e^{-\beta\lambda_n}}{\Gamma(\alpha)} \qquad \forall \ \ n = 1, ..., N$$

$$p(\alpha|a,b) = Gamma(\alpha|a,b) = \frac{b^a \alpha^{a-1} e^{-b\alpha}}{\Gamma(a)}$$

$$p(\beta|c,d) = Gamma(\beta|c,d) = \frac{d^c \beta^{c-1} e^{-d\beta}}{\Gamma(c)}$$

To find the conditional posterior for $\lambda_1, ..., \lambda_N, \alpha, \beta$, take only those distributions that corresponding to the parameter.

Conditional posterior of λ_n -

$$p(\lambda_n|X, \lambda_{-n}, \alpha, \beta) \propto Poisson(x_n|\lambda_n) Gamma(\lambda_n|\alpha, \beta)$$

Poisson and Gamma distribution are conjugate in this case and hence conditional posterior will have closed form solution with Gamma distribution.

$$\begin{aligned} p(\lambda_n|X, \lambda_{-n}, \alpha, \beta) &\propto \frac{\lambda_n^{x_n} e^{-\lambda_n}}{x_n!} \frac{\beta^\alpha \lambda_n^{\alpha-1} e^{-\beta\lambda_n}}{\Gamma(\alpha)} \\ &\propto \lambda_n^{x_n+\alpha-1} e^{-\lambda_n(\beta+1)} \end{aligned}$$

Therefore closed form conditional posterior distribution.

$$p(\lambda_n|X, \lambda_{-n}, \alpha, \beta) = Gamma(\lambda_n|\alpha + x_n, \beta + 1) \qquad \forall \ \ n = 1, ..., N$$

Conditional posterior of α -

Similarly for α, conditional posterior calculation involve gamma and gamma distribution which are not conjugate.

$$p(\alpha|X, \lambda_1, .., \lambda_n, \beta) \propto \prod_{n=1}^{N} Gamma(\lambda_n|, \alpha, \beta) Gamma(\alpha|a, b)$$

$$p(\alpha|X, \lambda_1, .., \lambda_n, \beta) \propto \frac{b^a \alpha^{a-1} e^{-b\alpha}}{\Gamma(a)} \prod_{n=1}^{N} \frac{\beta^\alpha \lambda_n^{\alpha-1} e^{-\beta\lambda_n}}{\Gamma(\alpha)}$$

$$p(\alpha|X, \lambda_1, .., \lambda_n, \beta) \propto \frac{\alpha^{a-1} \beta^{N\alpha} e^{-b\alpha} \left(\prod_{n=1}^{N} \lambda_n\right)^{\alpha-1}}{(\Gamma(\alpha))^N}$$

Therefore, no closed form conditional posterior distribution.

$$p(\alpha|X, \lambda_1, .., \lambda_n, \beta) \propto \frac{\alpha^{a-1}\beta^{N\alpha}e^{-b\alpha}\left(\prod_{n=1}^{N}\lambda_n\right)^{\alpha-1}}{(\Gamma(\alpha))^N}$$

Conditional posterior of β -
Similarly for β, conditional posterior calculation involve gamma and gamma distribution which are conjugate to each other.

$$p(\beta|X, \lambda_1, .., \lambda_n, \alpha) \propto \prod_{n=1}^{N} Gamma(\lambda_n|, \alpha, \beta)Gamma(\beta|c, d)$$
$$p(\beta|X, \lambda_1, .., \lambda_n, \alpha) \propto \frac{d^c\beta^{c-1}e^{-d\beta}}{\Gamma(c)} \prod_{n=1}^{N} \frac{\beta^{\alpha}\lambda_n^{\alpha-1}e^{-\beta\lambda_n}}{\Gamma(\alpha)}$$
$$p(\beta|X, \lambda_1, .., \lambda_n, \alpha) \propto \beta^{c+N\alpha-1}e^{-\beta(d+\sum_{n=1}^{N}\lambda_n)}$$

Therefore closed form conditional posterior distribution.

$$p(\beta|X, \lambda_1, .., \lambda_n, \alpha) = Gamma(\beta|c + N\alpha, d + \sum_{n=1}^{N}\lambda_n)$$

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

QUESTION 4

Student Name: Shashank Shailabh
Roll Number: 170655
Date: May 14, 2021

$p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j) = \mathcal{N}(r_{ij}|\mathbf{u}_i^T\mathbf{v}_j, \beta^{-1})$ where $\mathbf{u}_i$ and $\mathbf{v}_j$ are $i-th$ row and $j-th$ column of $\mathbf{R}$.

Using the samples generated by Gibbs sampler for mean and variance calculation.

$$p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j)p(\mathbf{u}_i, \mathbf{v}_j|\mathbf{R})d\mathbf{u}_i d\mathbf{v}_j = \mathbb{E}_{p(\mathbf{u}_i,\mathbf{v}_j|\mathbf{R})}[p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j)] = \frac{1}{S}\sum_{s=1}^{S} p(r_{ij}|\mathbf{u}_i^{(s)}, \mathbf{v}_j^{(s)})$$

Now expression for sample based approximation of mean,

$$\begin{aligned}
\mathbb{E}[r_{ij}] = \int r_{ij}p(r_{ij}|\mathbf{R})dr_{ij} &= \int r_{ij}\left(\frac{1}{S}\sum_{s=1}^{S}\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})\right)dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S}\int r_{ij}\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})dr_{ij} \\
&= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})}[r_{ij}] \\
&= \frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}
\end{aligned}$$

Now expression for sample based approximation of variance,

$$\begin{aligned}
var[r_{ij}] = \mathbb{E}[r_{ij}^2] - (\mathbb{E}[r_{ij}])^2 &= \int r_{ij}^2\left(\frac{1}{S}\sum_{s=1}^{S}\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})\right)dr_{ij} - \left(\frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}\right)^2 \\
&= \frac{1}{S}\sum_{s=1}^{S}\left[\int r_{ij}^2\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})dr_{ij}\right] - \left(\frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}\right)^2 \\
&= \frac{1}{S}\sum_{s=1}^{S}\left[\mathbb{E}_{\mathcal{N}(r_{ij}|(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}, \beta^{-1})}[r_{ij}^2]\right] - \left(\frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}\right)^2 \\
&= \frac{1}{S}\sum_{s=1}^{S}\left(\beta^{-1} + [\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}]^2\right) - \left(\frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}\right)^2
\end{aligned}$$

Therefore,

$$\mathbb{E}[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}$$

$$var[r_{ij}] = \frac{1}{S}\sum_{s=1}^{S}\left(\beta^{-1} + [\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}]^2\right) - \left(\frac{1}{S}\sum_{s=1}^{S}(\mathbf{u}_i^{(s)})^T\mathbf{v}_j^{(s)}\right)^2$$

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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: May 14, 2021

QUESTION 5

$p(x) \propto e^{sin(x)}$ and $\tilde{p} = e^{sin(x)}$ for $-\pi \leq x \leq \pi$.
Proposal distribution $q(x) = \mathcal{N}(x|0, \sigma^2)$

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

$$M \geq \frac{e^{sin(x)}}{\frac{e^{\frac{-x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}} = \sqrt{2\pi\sigma^2}e^{sin(x)+\frac{x^2}{2\sigma^2}}$$

Now, differentiate to find the maximum of $\frac{\tilde{p}(x)}{q(x)}$. Both σ and x are independent.
Differentiating w.r.t. σ will give

$$x^2 = \sigma^2$$

Differentiating w.r.t. x will give

$$x + \sigma^2 cos(x) = 0$$

On solving, the required solution for σ is approximately 2 and $M \geq 21$.
Therefore a suitable value of $M = 25$ and $\sigma = 2$.

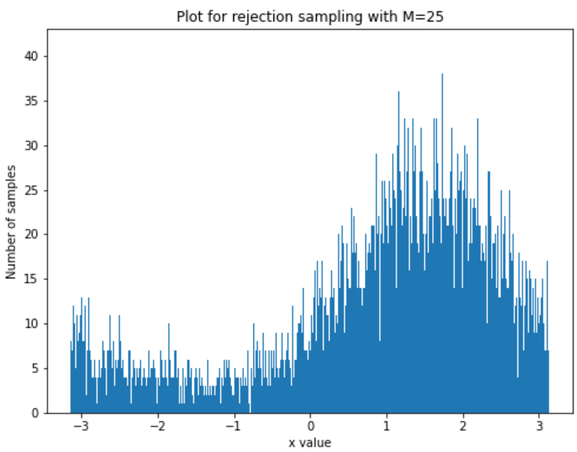

Figure 1: Histogram plot for rejection sampling.