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

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 1

The provided objective function can be simplified as follows,

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

$$= -\int q(\theta)\log p(\mathbf{X}|\theta)d\theta - \int q(\theta)\log\frac{p(\theta)}{q(\theta)}d\theta$$

$$= -\int q(\theta)\log\frac{p(\theta)p(\mathbf{X}|\theta)}{q(\theta)}d\theta = KL(q(\theta)||p(\theta)p(\mathbf{X}|\theta))$$

$KL(q(\theta)||p(\theta)p(\mathbf{X}|\theta))$ attains its minimum value at $q(\theta) = p(\theta)p(\mathbf{X}|\theta)$. According to Bayes rule, $p(\theta|\mathbf{X}) \propto p(\theta)p(\mathbf{X}|\theta)$. Therefore, by minimising the provided objective function, $q(\theta)$ approximates the Bayes posterior upto a proportionality constant.

Intuitively, the first term in the objective function is the negative of the expectation of complete data log likelihood with respect to $q(\theta)$. The second term is the KL divergence of prior on θ and $q(\theta)$. Therefore, by minimising the objective, we are trying to find a $q(\theta)$ that maximises the expected complete likelihood and minimise the KL divergence between posterior and prior. It seems like a decent trade-off.

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

QUESTION 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

Mean-field variational inference algorithm for approximating the posterior distribution:

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

Update equation for mean-field VI:

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\log p(\mathbf{Z}_j|\mathbf{X}, \mathbf{Z}_{-j})] + const$$

- For $\mathbf{w}$:

$$\begin{aligned}\log q_{\mathbf{w}}^*(\mathbf{w}) &= \mathbb{E}_{\beta,\alpha_1,...,\alpha_D}[\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\alpha}, \beta)] + const\\ &= \mathbb{E}_{\beta,\alpha_1,...,\alpha_D}[\log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \beta)p(\mathbf{w}|\boldsymbol{\alpha})] + const\\ &= \mathbb{E}_{\beta,\alpha_1,...,\alpha_D}[-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 - \frac{1}{2}\mathbf{w}^T diag(\alpha_1, ..., \alpha_D)\mathbf{w}] + const\end{aligned}$$

  This is similar to calculating the posterior for a Gaussian-Gaussian likelihood and prior pair. Therefore we can write,

$$q_{\mathbf{w}}^*(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}}, \Sigma_{\mathbf{w}}) \text{ where,}$$

$$\boldsymbol{\mu}_{\mathbf{w}} = \Sigma_{\mathbf{w}}\mathbb{E}[\beta]\mathbf{X}^T\mathbf{y}$$
$$\Sigma_{\mathbf{w}} = (\mathbb{E}[\beta]\mathbf{X}^T\mathbf{X} + diag(\mathbb{E}[\alpha_1], ..., \mathbb{E}[\alpha_D]))^{-1}$$

- For β:

$$\begin{aligned}\log q_{\beta}^*(\beta) &= \mathbb{E}_{\mathbf{w},\alpha_1,...,\alpha_D}[\log p(\beta|\mathbf{y}, \mathbf{X}, \boldsymbol{\alpha}, \mathbf{w})] + const\\ &= \mathbb{E}_{\mathbf{w},\alpha_1,...,\alpha_D}[\log p(\mathbf{y}|\beta, \mathbf{X}, \mathbf{w})p(\beta)] + const\end{aligned}$$

  By observation, form will be similar to the posterior of a Gamma-Gaussian prior-likelihood pair. Therefore,

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

- For α_d:

$$\begin{aligned}\log q_{\alpha_d}^*(\alpha_d) &= \mathbb{E}_{\mathbf{w},\beta,\alpha_1,..\alpha_{d-1},\alpha_{d+1},..,\alpha_D}[\log p(\alpha_d|\mathbf{w}_d)] + const\\ &= \mathbb{E}_{\mathbf{w},\beta,\boldsymbol{\alpha}_{-d}}[\log p(\mathbf{w}_d|\alpha_d)p(\alpha_d)] + const\end{aligned}$$

  By observation, form will be similar to the posterior of a Gamma-Gaussian prior-likelihood pair. Therefore,

$$q_{\alpha_d}^*(\alpha_d) = \text{Gamma}\left(\alpha_d|e_0 + \frac{1}{2}, f_0 + \frac{\mathbb{E}[w_d^2]}{2}\right)$$

The required expectations for the algorithm can be written as,
$\mathbb{E}[\mathbf{w}] = \boldsymbol{\mu}_{\mathbf{w}}$
$\mathbb{E}[\mathbf{w}^T\mathbf{w}] = \Sigma_w + \boldsymbol{\mu}^T\boldsymbol{\mu}$
$\mathbb{E}[w_d^2] = (\Sigma_w)_{dd} + (\boldsymbol{\mu}_{w_d})^2$
$\mathbb{E}[\beta] = \frac{a_0+\frac{N}{2}}{b_0+\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(y_n-\mathbf{w}^T\mathbf{x}_n)^2]}$
$\mathbb{E}[\alpha_d] = \frac{e_0+\frac{1}{2}}{f_0+\frac{\mathbb{E}[w_d^2]}{2}}$

Final Mean-field VI algorithm to approximate the posterior:

1. Calculate $\mathbb{E}(\beta)$ and $\mathbb{E}[\alpha_d]$ using initialized values.
2. $i := 1$. Until not converged,
   - Update $\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}$
   - Update $\mathbb{E}[\beta], \mathbb{E}[\alpha_d], \mathbb{E}[w_d^2] \forall d$
   - If maximum update $<$ stopping criteria, algorithm has converged
     Else, $i = i + 1$

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

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION

3

- Conditional posterior for $\lambda_n \forall n \in [1, N]$:

$$p(\lambda_n|x_n, \alpha, \beta) \propto p(x_n|\lambda_n, \alpha, \beta)p(\lambda_n|\alpha, \beta)$$

$$\implies p(\lambda_n|x_n, \alpha, \beta) \propto \text{Poisson}(x_n|\lambda_n)\text{Gamma}(\lambda_n|\alpha, \beta)$$

$$\implies p(\lambda_n|x_n, \alpha, \beta) \propto \lambda_n^{x_n+\alpha-1} e^{-(\beta+1)\lambda_n}$$

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

- Conditional posterior for α:

$$p(\alpha|\boldsymbol{\lambda}, \alpha, \beta) \propto p(\alpha|a, b) \prod_{n=1}^{N} p(\lambda_n|\alpha, \beta)$$

$$\implies p(\alpha|\boldsymbol{\lambda}, \alpha, \beta) \propto \text{Gamma}(\alpha|a, b) \prod_{n=1}^{N} \text{Gamma}(\lambda_n|\alpha, \beta)$$

$$\implies p(\alpha|\boldsymbol{\lambda}, \alpha, \beta) \propto \frac{\alpha^{a-1} e^{-b\alpha} (\lambda_1.....\lambda_N)^{\alpha-1}}{\Gamma(\alpha)^N}$$

The above expression does not match any known probability distribution and clearly there is no local conjugacy. Therefore, we cannot obtain a closed form expression for CP of α.

- Conditional posterior for β:

$$p(\beta|\boldsymbol{\lambda}, \alpha, \beta) \propto p(\beta|c, d) \prod_{n=1}^{N} p(\lambda_n|\alpha, \beta)$$

$$\implies p(\beta|\boldsymbol{\lambda}, \alpha, \beta) \propto \text{Gamma}(\beta|c, d) \prod_{n=1}^{N} \text{Gamma}(\lambda_n|\alpha, \beta)$$

$$\implies p(\beta|\boldsymbol{\lambda}, \alpha, \beta) \propto \beta^{N\alpha+c-1} e^{-\beta(d+\sum_{n=1}^{N} \lambda_n)}$$

$$\implies p(\beta|\boldsymbol{\lambda}, \alpha, \beta) = \text{Gamma}(\beta|N\alpha + c, d + \sum_{n=1}^{N} \lambda_n)$$

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 4

We can write any entry of matrix $\mathbf{R}$ as,

$$r_{ij} = \mathbf{u}_i^T \mathbf{v}_j + \epsilon_{ij}$$

$$\implies \mathbb{E}[r_{ij}] = \mathbb{E}[\mathbf{u}_i^T \mathbf{v}_j] + \mathbb{E}[\epsilon_{ij}]$$

We know that $\epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1})$. Now we can use our S samples to approximate the mean as,

$$\mathbb{E}[r_{ij}] \approx \sum_{s=1}^{S} \frac{1}{S}[(\mathbf{u}_i^T)^{(s)} \mathbf{v}_j^{(s)}] + 0$$

Similarly for variance,

$$\text{Var}(r_{ij}) = \mathbb{E}[r_{ij}^2] - \mathbb{E}[r_{ij}]^2$$

$$\implies \text{Var}(r_{ij}) = \mathbb{E}[(\mathbf{u}_i^T \mathbf{v}_j)^2] + \beta^{-1} - \mathbb{E}[r_{ij}]^2$$

$$\implies \text{Var}(r_{ij}) \approx \sum_{s=1}^{S} \frac{1}{S}[(\mathbf{u}_i^T)^{(s)} \mathbf{v}_j^{(s)}]^2 + \beta^{-1} - (\sum_{s=1}^{S} \frac{1}{S}[(\mathbf{u}_i^T)^{(s)} \mathbf{v}_j^{(s)}])^2$$

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

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 5

To find an optimal value of M, we need to ensure that $Mq(x) \geq \tilde{p}(x)$ for all values of x. Therefore,

$$M \geq \frac{\tilde{p}(x)}{q(x)} \implies M \geq \frac{\exp(\sin(x))}{(2\pi\sigma^2)^{-0.5}\exp(-\frac{x^2}{2\sigma^2})} \forall -\pi \leq x \leq \pi$$

Since max value of $\exp(\sin(x))$ in this range is 1 and the max value of $\exp(\frac{x^2}{2\sigma^2})$ is $\exp(\frac{\pi^2}{2\sigma^2})$, we can calculate a lower bound on M as follows,

$$M \geq (2\pi\sigma^2)^{0.5}\exp(1 + \frac{\pi^2}{2\sigma^2})$$

Choosing an appropriate value for variance, we can build the rejection sampler. For $\sigma^2 = 3$, $M \geq 61.14$ so let $M = 62$.