Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 1

Student Name: Ritesh Kumar
Roll Number: 160575
Date: March 14, 2019

Speeding Up Gaussian Processes Consider Gaussian Process(GP) regression where $y_n = f(\mathbf{x}_n) + \epsilon_n$ with f modeled by $\mathcal{GP}(0, \kappa)$ where GP mean function is 0 and kernel/covariance function is κ, and noise $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. We will consider noiseless setting, so $y_n = f(\mathbf{x}_n) = f_n$. Given N training inputs $(\mathbf{X}, \mathbf{f}) = \{\mathbf{x}_n, f_n\}_{n=1}^N$, we have seen that the posterior predictive distribution for a new input $\mathbf{x}_*$ is

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(f_*|\mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{f}, \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_*)$$

In the above, $\mathbf{K}$ is the $N \times N$ kernel matrix of training inputs and $\mathbf{k}_*$ is $N \times 1$ vector of kernel based similarities of $\mathbf{x}_*$ with each of the training inputs. The above has $\mathcal{O}(N^3)$ cost due to $N \times N$ matrix inversion.
Let's consider a way to reduce this cost to make GPs more scalable. To do this, suppose there are another set of *pseudo* training inputs $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_M\}$ with $M \ll N$, along with their respective noiseless *pseudo* outputs $\mathbf{t} = \{t_1, ...t_M\}$ modeled by the same GP, ie $t_m = f(\mathbf{z}_m)$. Note that $(\mathbf{Z}, \mathbf{t})$ are not known.
Now we have to assume that the likelihood for each training output f_n to be modeled by a posterior predictive having the same form as the GP regression's posterior predictive but with $(\mathbf{Z}, \mathbf{t})$ acting as "pseudo" training data.

$$p(f_n|\mathbf{x}_n, \mathbf{Z}, \mathbf{t}) = \mathcal{N}(f_n|\widetilde{\mathbf{k}}_*^\top \widetilde{\mathbf{K}}^{-1}\mathbf{t}, \kappa(\mathbf{x}_n, \mathbf{x}_n) - \widetilde{\mathbf{k}}_n^\top \widetilde{\mathbf{K}}^{-1}\widetilde{\mathbf{k}}_n)$$

In the above, $\widetilde{\mathbf{K}}$ is the $M \times M$ kernel matrix of the pseudo inputs $\mathbf{Z}$ and $\widetilde{\mathbf{k}}_n$ is the $M \times 1$ vector of kernel based similarities of $\mathbf{x}_n$ with each of the pseudo inputs $\mathbf{z}_1, ..., \mathbf{z}_M$. Given that

$$p(f_n|\mathbf{x}_n, \mathbf{Z}, \mathbf{t}) = \mathcal{N}\left(\mathbf{x}_n|\mathbf{k}_n^\top \mathbf{K}_M^{-1}\mathbf{t}, \kappa(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{k}_n^\top \mathbf{K}_M^{-1}\mathbf{k}_n\right)$$
$$p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) = \prod_{n=1}^N p(f_n|\mathbf{x}_n, \mathbf{Z}, \mathbf{t}) \qquad = \mathcal{N}(\mathbf{f}|\mathbf{P}\mathbf{K}_M^{-1}\mathbf{t}, \mathbf{\Lambda})$$

Here $\mathbf{P}$ is $N \times M$ matrix with $(\mathbf{P})_{nm} = \kappa(\mathbf{x}_n, \mathbf{z}_m)$ and $\mathbf{K}_M$ is $M \times M$ matrix with $(\mathbf{K}_M)_{nm} = \kappa(\mathbf{z}_n, \mathbf{z}_m)$. Also $\mathbf{\Lambda}$ is a diagonal matrix with $(\mathbf{\Lambda})_{ii} = \kappa(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k}_n^\top \mathbf{K}_M^{-1}\mathbf{k}_n$. Also

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) = \int p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}, \mathbf{t})p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})d\mathbf{t}$$

Using Baye's rule to get the posterior over $\mathbf{t}$, we get

$$p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) \propto p(\mathbf{f}|\mathbf{X}, \mathbf{t}, \mathbf{Z})p(\mathbf{t}|\mathbf{Z})$$

Since the pseudo sample points are modelled by the same Gaussian Process, we have $p(\mathbf{t}|\mathbf{Z}) = (\mathbf{t}|0, \mathbf{K}_M)$. Writing the terms in exponent in the RHS of the above proportionality in information form of Gaussian and solving we get

$$p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}|\mathbf{f}}, \mathbf{\Sigma}_{\mathbf{t}|\mathbf{f}})$$

Where we have $\mathbf{\Sigma_{t|f}} = \left(\mathbf{K}_M^{-1}\mathbf{P}^\top\mathbf{\Lambda}^{-1}\mathbf{P}\mathbf{K}_M^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{t|f}} = \mathbf{\Sigma_{t|f}}\mathbf{K}_M^{-1}\mathbf{P}^\top\mathbf{\Lambda}^{-1}\mathbf{f}$. Since $y_* = f_*$. We can write $f_* = \mathbf{k}_*^\top\mathbf{K}_M^{-1}\mathbf{t} + \epsilon$, where $\mathbf{k}_*$ is $M \times 1$ vector with $(\mathbf{k}_*)_i = \kappa(\mathbf{x}_*, \mathbf{z}_i)$, and $\epsilon = \mathcal{N}(0, \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top\mathbf{K}_M^{-1}\mathbf{k}_*)$. Now using the property of linear Gaussian model we get

$$\begin{aligned} p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) &= \mathcal{N}(f_*|\boldsymbol{\mu}_*, \mathbf{\Sigma}_*) \\ \text{where } \boldsymbol{\mu}_* &= \mathbf{k}_*^\top\mathbf{K}_M^{-1}\mathbf{\Sigma_{t|f}}\mathbf{K}_M^{-1}\mathbf{P}^\top\mathbf{\Lambda}^{-1}\mathbf{f} \\ \text{and } \mathbf{\Sigma}_* &= \mathbf{k}_*^\top\mathbf{K}_M^{-1}\mathbf{\Sigma_{t|f}}\mathbf{K}_M^{-1}\mathbf{k}_* + \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top\mathbf{K}_M^{-1}\mathbf{k}_* \end{aligned}$$

Note that the computation of posterior predictive is mainly dominated by the term $\mathbf{\Sigma_{t|f}}$ whose computation cost is now $\mathcal{O}(M^2N)$ which is much less compared to earlier version of the Gaussian Process(where it was $\mathcal{O}(N^3)$)

Part 2

$$p(\mathbf{f}|\mathbf{X}, \mathbf{Z}) = \int p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t})p(\mathbf{t}|\mathbf{Z})d\mathbf{t}$$

We can write $\mathbf{f} = \mathbf{P}\mathbf{K}_M^{-1}\mathbf{t} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = \mathcal{N}(0, \mathbf{\Lambda})$

Again using property of linear Gaussian model, we have

$$\begin{aligned} p(\mathbf{f}|\mathbf{X}, \mathbf{Z}) &= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{\Sigma}) \\ \text{where } \boldsymbol{\mu} &= \mathbf{P}\mathbf{K}_M^{-1}0 = 0 \\ \text{and } \mathbf{\Sigma} &= \mathbf{P}\mathbf{K}_M^{-1}\mathbf{P}^\top + \mathbf{\Lambda} \end{aligned}$$

Hence to solve for $\mathbf{Z}$ via MLE-II we have the following objective function

$$\begin{aligned} \hat{\mathbf{Z}} &= argmax_{\mathbf{Z}} p(\mathbf{f}|\mathbf{X}, \mathbf{Z}) \\ &= argmax_{\mathbf{Z}} \left(-\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{f}^\top\mathbf{\Sigma}^{-1}\mathbf{f}\right) \\ &= argmin_{\mathbf{Z}} \left(\log|\mathbf{\Sigma}| + \mathbf{f}^\top\mathbf{\Sigma}^{-1}\mathbf{f}\right) \end{aligned}$$

The above objective function can be solved using gradient ascent.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ritesh Kumar
Roll Number: 160575
Date: March 14, 2019

QUESTION 2

(Two Flavors of EM for an LVM) Suppose we are given N observations $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and we wish to model them via a latent variable model with M mixture component. Generative story:

- For $n = 1, ..., N$
  - Draw a mixture component id $c_n \sim multinoullli(\pi_1, ...\pi_M)$. Suppose $c_n = m \in \{1, ..., M\}$
  - Generate a K dimensional latent variable $\mathbf{z}_n$ from $p(\mathbf{z}_n|c_n = m) = \mathcal{N}(0, \mathbf{I}_K)$
  - Generate $\mathbf{x}_n$ as $\mathbf{x}_n = \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \epsilon_n$ where $\epsilon_n \in \mathcal{N}(0, \sigma_m^2\mathbf{I}_D)$

Define $\Theta = \{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^M$

<u>*Part 1*</u> Estimate $\{c_n\}_{n=1}^N$ and parameters Θ

1. *Conditional posterior of the latent variables*
   Note that for $p(\mathbf{x}_n|c_n = m, \Theta)$

$$
\begin{aligned}
\mathbf{x}_n &= \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \epsilon_n \\
&\text{so, } \mathbf{x}_n \text{ will be Gaussian with} \\
\mathbb{E}[\mathbf{x}_n] &= \boldsymbol{\mu}_m \\
var(\mathbf{x}_n) &= \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D \\
\implies \mathbf{x}_n &\sim \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right)
\end{aligned}
$$

above results are due to linear transformation of Random Variable
Therefore, we get

$$
\begin{aligned}
p(c_n = m|\mathbf{x}_n, \Theta) &\propto p(c_n = m|\Theta)p(\mathbf{x}_n|c_n = m, \Theta) \\
&= \pi_m\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right) \\
&\text{normalizing we get} \\
p(c_n = m|\mathbf{x}_n, \Theta) &= \frac{\pi_m\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}
\end{aligned}
$$

Therefore, conditional posterior is

$$
p(c_n = m|\mathbf{x}_n, \Theta) = \frac{\pi_m\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}
$$

2. *CLL*

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{c}|\Theta) &= \prod_{n=1}^{N} p(\mathbf{x}_n, c_n|\Theta) \\
&= \prod_{n=1}^{N} \prod_{m=1}^{M} \left(p(c_n = m|\Theta) p(\mathbf{x}_n|c_n = m, \Theta)\right)^{c_{nm}} \\
&= \prod_{n=1}^{N} \prod_{m=1}^{M} \left(\pi_m \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D)\right)^{c_{nm}}
\end{aligned}
$$

where $c_{nm} = \mathbb{I}[c_n = m]$
$\Longrightarrow$

$$
\log\left(p(\mathbf{X}, \mathbf{c}|\Theta)\right) = \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\left[\log(\pi_m) + \log(\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D)\right]
$$

and

3. *Expected CLL*

$$
\begin{aligned}
\mathbb{E}\left[\log\left(p(\mathbf{X}, \mathbf{c}|\Theta)\right)\right] &= \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}\left[c_{nm}\right]\left[\log(\pi_m) + \log(\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D)\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{nm}\left[\log(\pi_m) + \log(\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D)\right]
\end{aligned}
$$

where $\gamma_{nm} = \mathbb{E}\left[\mathbb{I}[c_n = m]\right] = p(c_n = m|\mathbf{x}_n, \Theta)$

4. The only expected values required is γ_{nm} given by

$$
\gamma_{nm} = \frac{\pi_m \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^{M}\pi_l \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}
$$

5. *M step update equations for* Θ

$$
\begin{aligned}
\hat{\pi}_m &= \frac{N_m}{N} \\
\hat{\boldsymbol{\mu}}_m &= \frac{1}{N_m}\sum_{n=1}^{N}\gamma_{nm}\mathbf{x}_n \\
\hat{\mathbf{W}}_m\hat{\mathbf{W}}_m^\top + \hat{\sigma}_m^2\mathbf{I}_D &= \frac{1}{N_m}\sum_{n=1}^{N}\gamma_{nm}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top
\end{aligned}
$$

Here $N_m = \sum_{n=1}^{N}\gamma_{nm}$, $\hat{\sigma}_m^2 = \frac{1}{D-K}\sum_{k=k+1}^{D}\lambda_k$ and $\hat{\mathbf{W}}_m = \mathbf{L}_K(\mathbf{U}_K - \hat{\sigma}_m^2\mathbf{I}_K)^{\frac{1}{2}}\mathbf{R}$ where $\mathbf{L}_K$ is a $D \times K$ matrix of top K eigen vectors of $\hat{\mathbf{W}}_m\hat{\mathbf{W}}_m^T + \hat{\sigma}_m^2\mathbf{I}_D$, $\mathbf{U}_K$ is $K \times K$ diagonal matrix of top K eigenvalues and $\mathbf{R}$ is the $K \times K$ rotation matrix. The above method is computationaly expensive because of eigenvalue decomposition. Note that we are not estimating the extra unknowns $\mathbf{z}_n$'s.

6. *The overall sketch of the EM algorithm*

- Initialize $\Theta = \{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^M$ to $\Theta^{(0)}$ and set $t = 1$
- E-step
  For all $n = 1, ..., N$ and $m = 1, ...M$

$$\gamma_{nm}^{(t)} = \frac{\pi_m^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_m^{(t-1)}, \mathbf{W}_m^{(t-1)} \mathbf{W}_m^{(t-1)^\top} + \sigma_m^{(t-1)^2} \mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_l^{(t-1)}, \mathbf{W}_l^{(t-1)} \mathbf{W}_l^{(t-1)^\top} + \sigma_l^{(t-1)^2} \mathbf{I}_D\right)}$$

- M-step: RHS values are of $time = t - 1$
  For $m = 1, ..., M$

$$\begin{aligned}
N_m &= \sum_{n=1}^N \gamma_{nm} \\
\hat{\pi}_m &= \frac{N_m}{N} \\
\hat{\boldsymbol{\mu}}_m &= \frac{1}{N_m} \sum_{n=1}^N \gamma_{nm} \mathbf{x}_n \\
\hat{\mathbf{W}}_m \hat{\mathbf{W}}_m^\top + \hat{\sigma}_m^2 \mathbf{I}_D &= \frac{1}{N_m} \sum_{n=1}^N \gamma_{nm} (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top \\
\hat{\sigma}_m^2 &= \frac{1}{D - K} \sum_{k=k+1}^D \lambda_k \\
\hat{\mathbf{W}}_m &= \mathbf{L}_K (\mathbf{U}_K - \hat{\sigma}_m^2 \mathbf{I}_K)^{\frac{1}{2}} \mathbf{R}
\end{aligned}$$

- Set $t = t + 1$ and goto step 2 is not converged

7. *Corresponding Stepwise Online algorithm*

- Initialize $\Theta = \{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^M$ to $\Theta^{(0)}$ and set $t = 1$
- Pick a random example $\mathbf{x}_n$
- Compute γ_{nm} for every m
- Compute the MLE estimates of the global parameters $\hat{\Theta}$ using only $\mathbf{x}_n$

$$\begin{aligned}
\hat{\pi}_m &= \gamma_{nm} \\
\hat{\boldsymbol{\mu}}_m &= \mathbf{x}_n \\
\hat{\mathbf{W}}_m \hat{\mathbf{W}}_m^\top + \hat{\sigma}_m^2 \mathbf{I}_D &= (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top \\
\hat{\sigma}_m^2 &= \frac{1}{D - K} \sum_{k=k+1}^D \lambda_k \\
\hat{\mathbf{W}}_m &= \mathbf{L}_K (\mathbf{U}_K - \hat{\sigma}_m^2 \mathbf{I}_K)^{\frac{1}{2}} \mathbf{R}
\end{aligned}$$

- Compute the learning rate $\epsilon^{(t)}$
- Update every parameter as

$$\Theta^{(t)} = (1 - \epsilon^{(t)}) \Theta^{(t-1)} + \epsilon^{(t)} \hat{\Theta}$$

- Set $t = t + 1$ and go to step 2 if not converged

Part 2 Estimate $\{\mathbf{z}_n, c_n\}_{n=1}^N$ and parameters Θ

1. *Conditional posterior of the latent variables*
   Note that for $p(\mathbf{x}_n|\mathbf{z}_n, c_n = m, \Theta)$

$$\begin{aligned}
\mathbf{x}_n &= \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \epsilon_n \\
&\text{so, } \mathbf{x}_n \text{ will be Gaussian with} \\
\mathbb{E}\left[\mathbf{x}_n\right] &= \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n \\
var(\mathbf{x}_n) &= \sigma_m^2\mathbf{I}_D \\
\implies \mathbf{x}_n &\sim \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)
\end{aligned}$$

   above results are due to linear transformation of Random Variable
   Therefore, we get

$$\begin{aligned}
p(c_n = m, \mathbf{z}_n|\mathbf{x}_n, \Theta) &\propto p(c_n = m|\Theta)p(\mathbf{z}_n|c_n = m, \Theta)p(\mathbf{x}_n|c_n = m, \Theta) \\
&= \pi_m\mathcal{N}\left(\mathbf{z}_n|0, \mathbf{I}_K\right)\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right) \\
&\text{normalizing we get} \\
p(c_n = m, \mathbf{z}_n|\mathbf{x}_n, \Theta) &= \frac{\pi_m\mathcal{N}\left(\mathbf{z}_n|0, \mathbf{I}_K\right)\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}
\end{aligned}$$

   Therefore, conditional posterior is

$$p(c_n = m, \mathbf{z}_n|\mathbf{x}_n, \Theta) = \frac{\pi_m\mathcal{N}\left(\mathbf{z}_n|0, \mathbf{I}_K\right)\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}$$

   Note that the denominator is the of same as in part 1 since it is marginal distribution of $\mathbf{x}_n$ in both cases.
   Individual conditional posteriors

   (a) Conditional posterior of c_n would be same as derived in the above solution

$$p(c_n = m|\mathbf{x}_n, \Theta) = \frac{\pi_m\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^\top + \sigma_m^2\mathbf{I}_D\right)}{\sum_{l=1}^M \pi_l\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{W}_l\mathbf{W}_l^\top + \sigma_l^2\mathbf{I}_D\right)}$$

   (b) For the extimation of $p(\mathbf{z}_n|c_n = m, \mathbf{x}_n, \Theta) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. By using Linear transformation of Gaussian we find that

$$\begin{aligned}
\boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n\mathbf{W}_m^\top(\sigma_m^2\mathbf{I}_D)^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m) \\
\boldsymbol{\Sigma}_n &= \sigma_m^2\left(\sigma_m^2\mathbf{I}_K + \mathbf{W}_m^\top\mathbf{W}_m\right)^{-1}
\end{aligned}$$

2. *CLL*

$$\begin{aligned}
p(\mathbf{X}, \mathbf{Z}, \mathbf{c}|\Theta) &= \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n, c_n|\Theta) \\
&= \prod_{n=1}^N\prod_{m=1}^M \left(p(c_n = m|\Theta)p(\mathbf{z}_n|c_n = m, \Theta)p(\mathbf{x}_n|c_n = m, \mathbf{z}_n, \Theta)\right)^{c_{nm}} \\
&= \prod_{n=1}^N\prod_{m=1}^M \left(\pi_m\mathcal{N}\left(\mathbf{z}_n|0, \mathbf{I}_K\right)\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)\right)^{c_{nm}}
\end{aligned}$$

where $c_{nm} = \mathbb{I}[c_n = m]$
$\Longrightarrow$

$$\log\left(p(\mathbf{X},\mathbf{c}|\Theta)\right) = \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\left[\log(\pi_m) + \log\mathcal{N}\left(\mathbf{z}_n|0,\mathbf{I}_K\right) + \log\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)\right]$$
$$= \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\Bigg(-\frac{D}{2}\log(\sigma_m^2) - \frac{1}{2\sigma_m^2}(\mathbf{x}_n-\boldsymbol{\mu}_m)^\top(\mathbf{x}_n-\boldsymbol{\mu}_m) - \frac{1}{2\sigma_m^2}(\mathbf{z}_n\mathbf{W}_m^\top\mathbf{W}_m\mathbf{z}_n)$$
$$+ \frac{1}{2\sigma_m^2}\Big((\mathbf{W}_m\mathbf{z}_n)^\top(\mathbf{x}_n-\boldsymbol{\mu}_m) + (\mathbf{x}_n-\boldsymbol{\mu}_m)^\top(\mathbf{W}_m\mathbf{z}_n)\Big) - \frac{1}{2}\mathbf{z}_n^\top\mathbf{z}_n + \log\pi_m\Bigg)$$
$$= \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\Bigg(-\frac{D}{2}\log\sigma_m^2 - \frac{1}{2\sigma_m^2}(\mathbf{x}_n-\boldsymbol{\mu}_m)^\top(\mathbf{x}_n-\boldsymbol{\mu}_m) - \frac{1}{2\sigma_m^2}Tr(\mathbf{z}_n\mathbf{z}_n^\top\mathbf{W}_m^\top\mathbf{W}_m)$$
$$+ \frac{1}{\sigma_m^2}(\mathbf{x}_n-\boldsymbol{\mu}_m)^\top\mathbf{W}_m\mathbf{z}_n - \frac{1}{2}Tr(\mathbf{z}_n\mathbf{z}_n^\top) + \log\pi_m\Bigg)$$

and

3. Expected CLL

$$\mathbb{E}\left[\log\left(p(\mathbf{X},\mathbf{z}_n,\mathbf{c}|\Theta)\right)\right] = \mathbb{E}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\left[\log(\pi_m) + \log\mathcal{N}\left(\mathbf{z}_n|0,\mathbf{I}_K\right) + \log\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)\right]\right]$$

We have to evaluate $\mathbb{E}_{p(\mathbf{z}_n,c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}Tr(\mathbf{z}_n\mathbf{z}_n^\top)]$
Note that

$$\mathbb{E}_{p(\mathbf{z}_n,c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}Tr(\mathbf{z}_n\mathbf{z}_n^\top)] = \int\sum_{l=1}^{M} c_{nm}Tr(\mathbf{z}_n\mathbf{z}_n^\top)p(\mathbf{z}_n,c_n=m|\mathbf{x}_n,\Theta)d\mathbf{z}_n$$
$$= \int\sum_{l=1}^{M} c_{nm}Tr(\mathbf{z}_n\mathbf{z}_n^\top)p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)p(c_n=m|\mathbf{x}_n,\Theta)d\mathbf{z}_n$$
$$= \sum_{l=1}^{M} c_{nm}\gamma_{nm}\int Tr(\mathbf{z}_n\mathbf{z}_n^\top)\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)d\mathbf{z}_n$$

Therefore

$$\mathbb{E}_{p(\mathbf{z}_n,c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}Tr(\mathbf{z}_n\mathbf{z}_n^\top)] = \mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[Tr(\mathbf{z}_n\mathbf{z}_n^\top]\mathbb{E}_{p(c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}]$$

Therefore, finding expectation of joint conditional distribution when $\mathbf{z}_n$ and c_n occur together is same as taking respective marginal conditional posterior. Therefore, we only require to find $\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n]$, $\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n\mathbf{z}_n^\top]$ and $\mathbb{E}_{p(c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}]$

$$\mathbb{E}\left[\log\left(p(\mathbf{X},\mathbf{z}_n,\mathbf{c}|\Theta)\right)\right] = \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}[c_{nm}]\left[\log(\pi_m) + \log\mathcal{N}\left(\mathbf{z}_n|0,\mathbf{I}_K\right) + \log\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D\right)\right]$$
$$= \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}[c_{nm}]\Bigg(-\frac{D}{2}\log\sigma_m^2 - \frac{1}{2\sigma_m^2}(\mathbf{x}_n-\boldsymbol{\mu}_m)^\top(\mathbf{x}_n-\boldsymbol{\mu}_m)$$
$$- \frac{1}{2\sigma_m^2}Tr(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\mathbf{W}_m^\top\mathbf{W}_m) + \frac{1}{\sigma_m^2}(\mathbf{x}_n-\boldsymbol{\mu}_m)^\top\mathbf{W}_m\mathbb{E}[\mathbf{z}_n] - \frac{1}{2}Tr(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top])$$
$$+ \log\pi_m\Bigg)$$

4. *Expected value*

$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n] = \gamma_{nm}$$
$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top$$
$$\mathbb{E}_{p(c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}] = \boldsymbol{\mu}_n$$

5. *M-step update equations*

$$\hat{\pi}_m = \frac{\sum_{n=1}^N \gamma_{nm}}{N}$$
$$\hat{\boldsymbol{\mu}}_m = \frac{1}{N_m}\sum_{n=1}^N \gamma_{nm}(\mathbf{x}_n - \mathbf{W}_m\boldsymbol{\mu}_n)$$
$$\hat{\mathbf{W}}_m = \left(\sum_{n=1}^N \gamma_{nm}(\mathbf{x}_n - \boldsymbol{\mu}_m)\boldsymbol{\mu}_n^\top\right)\left(\sum_{n=1}^N \gamma_{nm}(\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\right)^{-1}$$
$$\hat{\sigma}_m^2 = \frac{1}{DN_m}\sum_{n=1}^N \left((\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m) - 2(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top\hat{\mathbf{W}}_m\boldsymbol{\mu}_n + Tr\left((\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\hat{\mathbf{W}}_m^\top\hat{\mathbf{W}}_m\right)\right)$$

6. *Overall EM Algorithm*

- Initailize $\Theta = \Theta^{(0)}$, set $t = 1$
- E-Step

$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n] = \gamma_{nm}\ \forall m, n$$
$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top\ \forall n$$
$$\mathbb{E}_{p(c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}] = \boldsymbol{\mu}_n\ \forall n$$

- M-Step

$$\hat{\pi}_m = \frac{\sum_{n=1}^N \gamma_{nm}}{N}$$
$$\hat{\boldsymbol{\mu}}_m = \frac{1}{N_m}\sum_{n=1}^N \gamma_{nm}(\mathbf{x}_n - \mathbf{W}_m\boldsymbol{\mu}_n)$$
$$\hat{\mathbf{W}}_m = \left(\sum_{n=1}^N \gamma_{nm}(\mathbf{x}_n - \boldsymbol{\mu}_m)\boldsymbol{\mu}_n^\top\right)\left(\sum_{n=1}^N \gamma_{nm}(\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\right)^{-1}$$
$$\hat{\sigma}_m^2 = \frac{1}{DN_m}\sum_{n=1}^N \left((\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m) - 2(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top\hat{\mathbf{W}}_m\boldsymbol{\mu}_n + Tr\left((\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\hat{\mathbf{W}}_m^\top\hat{\mathbf{W}}_m\right)\right)$$

- $t = t + 1$, goto step 2 if not converged

7. *Stepwise(online) EM algorithm*

- Initialize $\Theta = \Theta^{(0)}$, set $t = 1$
- Pick a random sample $\mathbf{x}_n$ from the data

- E-Step

$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n] = \gamma_{nm} \ \forall m$$
$$\mathbb{E}_{p(\mathbf{z}_n|c_n=m,\mathbf{x}_n,\Theta)}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top \ \forall n$$
$$\mathbb{E}_{p(c_n=m|\mathbf{x}_n,\Theta)}[c_{nm}] = \boldsymbol{\mu}_n \ \forall n$$

- M-Step

$$\hat{\pi}_m = \frac{\gamma_{nm}}{N}$$
$$\hat{\boldsymbol{\mu}}_m = \frac{1}{N_m}\gamma_{nm}(\mathbf{x}_n - \mathbf{W}_m\boldsymbol{\mu}_n)$$
$$\hat{\mathbf{W}}_m = \left(\gamma_{nm}(\mathbf{x}_n - \boldsymbol{\mu}_m)\boldsymbol{\mu}_n^\top\right)\left(\gamma_{nm}(\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\right)^{-1}$$
$$\hat{\sigma}_m^2 = \frac{1}{DN_m}\left((\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m) - 2(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^\top\hat{\mathbf{W}}_m\boldsymbol{\mu}_n + Tr\left((\mathbf{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top)\hat{\mathbf{W}}_m^\top\hat{\mathbf{W}}_m\right)\right)$$

  Compute the learning rate and let the above computed update paramters be $\hat{\Theta}$

$$\Theta^{(t)} = (1 - \eta_t)\Theta^{(t-1)} + \eta_t\hat{\Theta}$$

- $t = t + 1$, goto step 2 if not converged

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 3

Student Name: Ritesh Kumar
Roll Number: 160575
Date: March 14, 2019

(Mean-field VI for Sparse Bayesian Linear Regression) Assume N observations $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ generated from a linear regression model $y_n \sim \mathcal{N}\left(y_n|\mathbf{w}^\top \mathbf{x}_n, \beta^{-1}\right)$. Assume a Gaussian prior on $\mathbf{w}$ with different component-wise precisions, i.e. . Also

- $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, diag(\alpha_1^{-1}, ..., \alpha_D^{-1}))$
- $\beta \sim Gamma(\beta|a_0, b_0)$
- $\alpha_d \sim Gamma(\alpha_d|e_0, f_0)\forall d$
- $Gamma(\eta|\tau_1, \tau_2) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)}\eta^{\tau_1-1}\exp(-\tau_2\eta)$

To Derive: Mean-field VI algorithm for approximating the posterior distribution $p(\mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{y}, \mathbf{X})$.
Derivation:
We have to first find $p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X})$

$$\begin{aligned}
p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X}) &= p(\mathbf{y}|\mathbf{w}, \beta, \mathbf{X})p(\mathbf{w}|\alpha_1, ..., \alpha_D)p(\beta)p(\alpha_1, ..., \alpha_D) \\
&= \prod_{n=1}^{N} p(y_n|\mathbf{w}, \mathbf{x}_n, \beta)p(\mathbf{w}|\alpha_1, ..., \alpha_D)p(\beta)\prod_{d=1}^{D} p(\alpha_d)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X}) &= \sum_{n=1}^{N} \log p(y_n|\mathbf{w}, \mathbf{x}_n, \beta) + \log p(\mathbf{w}|\alpha_1, ..., \alpha_D) + \log \beta + \sum_{d=1}^{D} \log p(\alpha_d) \\
&= \sum_{n=1}^{N} \log\left(\sqrt{\frac{\beta}{2\pi}}\exp\left(\frac{-\beta}{2}(y_n - \mathbf{w}^\top \mathbf{x}_n)^2\right)\right) + \log\left(\sqrt{\frac{\alpha_1...\alpha_D}{(2\pi)^D}}\exp\left(\frac{-\mathbf{w}^\top\Sigma\mathbf{w}}{2}\right)\right) \\
&+ \log\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\beta^{a_0-1}\exp(-b_0\beta)\right) + \sum_{d=1}^{D}\log\left(\frac{f_0^{e_0}}{\Gamma(e_0)}\alpha_d^{e_0-1}\exp(-f_0\alpha_d)\right) \\
&\propto \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^\top\mathbf{x}_n)^2 + \frac{1}{2}\sum_{d=1}^{D}\log\alpha_d - \frac{1}{2}\mathbf{w}^\top\Sigma\mathbf{w} + (a_0 - 1)\log\beta - b_0\beta \\
&+ (e_0 - 1)\sum_{d=1}^{D}\log\alpha_d - f_0\sum_{d=1}^{D}\alpha_d
\end{aligned}$$

where $\Sigma = diag(\alpha_1, ..., \alpha_D)$
For $\mathbf{w}$

$$\begin{aligned}
\log q_{\mathbf{w}}^*(\mathbf{w}) &= \mathbb{E}_{q_{\beta,\alpha_1,...\alpha_D}}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X})\right] + const \\
&= \mathbb{E}\left[-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^\top\mathbf{x}_n)^2 - \frac{1}{2}\mathbf{w}^\top\Sigma\mathbf{w}\right] + const \\
&= \frac{-1}{2}\left\{\mathbf{w}^\top\left(\mathbb{E}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top + diag(\mathbb{E}[\alpha_1], ..., \mathbb{E}[\alpha_2])\right)\mathbf{w} - 2\mathbf{w}^\top\mathbb{E}[\beta]\sum_{n=1}^{N}y_n\mathbf{x}_n\right\} + const
\end{aligned}$$

Therefore, $\mathbf{w}$ has Gaussian form

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w}} &= \left(\mathbb{E}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top + diag(\mathbb{E}[\alpha_1], ..., \mathbb{E}[\alpha_2])\right)^{-1}\mathbb{E}[\beta]\sum_{n=1}^{N} y_n\mathbf{x}_n \\
\boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\mathbb{E}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top + diag(\mathbb{E}[\alpha_1], ..., \mathbb{E}[\alpha_2])\right)^{-1}
\end{aligned}$$

For β

$$\begin{aligned}
\log q^*_\beta(\beta) &= \mathbb{E}_{q_{\mathbf{w},\alpha_1,...\alpha_D}}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X})\right] + const \\
&= \mathbb{E}\left[\frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^\top\mathbf{x}_n)^2 + (a_0 - 1)\log\beta - b_0\beta\right] + const \\
&= \left(\frac{N}{2} + a_0 - 1\right)\log\beta - \beta\left(\sum_{n=1}^{N}\frac{1}{2}\mathbb{E}\left[(y_n - \mathbf{w}^\top\mathbf{x}_n)^2\right] + b_0\right) + const
\end{aligned}$$

Therefore, β has Gamma form

$$\beta \sim Gamma(\beta|a, b)$$

where

$$\begin{aligned}
a &= \frac{N}{2} + a_0 \\
b &= \sum_{n=1}^{N}\frac{1}{2}\mathbb{E}\left[(y_n - \mathbf{w}^\top\mathbf{x}_n)^2\right] + b_0
\end{aligned}$$

Note that for taking the above expectation we would require $\mathbb{E}[\mathbf{w}]$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^\top]$
For α_d, $d = 1, ..., D$

$$\begin{aligned}
\log q^*_{\alpha_d}(\alpha_d) &= \mathbb{E}_{q_{\mathbf{w},\beta,\alpha_1,...,\alpha_{d-1},\alpha_{d+1},...,\alpha_D}}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{X})\right] + const \\
&= \mathbb{E}\left[\frac{\log\alpha_d}{2} - \frac{w_d^2\alpha_d}{2} + (e_0 - 1)\log\alpha_d - f_0\alpha_d\right] + const \\
&= \left(\frac{1}{2} + e_0 - 1\right)\log\alpha_d - \alpha_d\left(f_0 + \frac{\mathbb{E}[w_d^2]}{2}\right) + const
\end{aligned}$$

Therefore, α_d has Gamma form

$$\alpha_d \sim Gamma(\alpha_d|e_d, f_d)$$

where

$$\begin{aligned}
e_d &= \frac{1}{2} + e_0 \\
f_d &= f_0 + \frac{\mathbb{E}[w_d^2]}{2}
\end{aligned}$$

The expectations required for above calculations are following

$$
\begin{aligned}
\mathbb{E}[\mathbf{w}] &= \boldsymbol{\mu}_{\mathbf{w}} \\
\mathbb{E}[\mathbf{w}\mathbf{w}^\top] &= \boldsymbol{\Sigma}_{\mathbf{w}} + \boldsymbol{\mu}_{\mathbf{w}}\boldsymbol{\mu}_{\mathbf{w}}^\top \\
\mathbb{E}[w_d^2] &= \boldsymbol{\Sigma}_{\mathbf{w}dd} + \mu_{wd}^2 \\
\mathbb{E}[\beta] &= \frac{a}{b} \\
\mathbb{E}[\alpha_d] &= \frac{e_d}{f_d} \ \forall d
\end{aligned}
$$

Note that the update equation one parameter is dependent on other parameters and therefore, the updates would be cyclic.

Algorithm

- Set $e_d = \frac{1}{2} + e_0$ and $a = \frac{N}{2} + a_0$
- Initialize $f_d \ \forall d$ and b
- Find all expected values required using above formula
- Set $t = 1$. Until not converged

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w}} &= \left(\mathbb{E}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top + diag(\mathbb{E}[\alpha_1],...,\mathbb{E}[\alpha_2])\right)^{-1}\mathbb{E}[\beta]\sum_{n=1}^{N}y_n\mathbf{x}_n \\
\boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\mathbb{E}[\beta]\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top + diag(\mathbb{E}[\alpha_1],...,\mathbb{E}[\alpha_2])\right)^{-1} \\
b &= \sum_{n=1}^{N}\frac{1}{2}\mathbb{E}\left[(y_n - \mathbf{w}^\top\mathbf{x}_n)^2\right] + b_0 \\
f_d &= f_0 + \frac{\mathbb{E}[w_d^2]}{2} \ \forall d \\
\mathbb{E}[\beta] &= \frac{a}{b} \\
\mathbb{E}[\alpha_d] &= \frac{e_d}{f_d} \ \forall d \\
t &= t + 1
\end{aligned}
$$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 4

Student Name: Ritesh Kumar
Roll Number: 160575
Date: March 14, 2019

VI for Bayesian Logistic Regression Bayesian logistic regression is non-conjugate model. Given

- N examples $\{\mathbf{x}_n, y_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \{-1, +1\}$
- $p(y_n|\mathbf{w}, \mathbf{x}_n) = \sigma(y_n \mathbf{w}^\top \mathbf{x}_n)$
- $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$
- λ is fixed
- $q(\mathbf{w}|\phi) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ modeled as $\mathbf{L}\mathbf{L}^\top$ where $\mathbf{L}$ is a $D \times D$ real valued matrix

We can write the ELBO expression as

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}\left[\log p(\mathbf{X}, \mathbf{w}) - \log q(\mathbf{w}|\phi)\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \log \sigma(y_n \mathbf{w}^T \mathbf{x}_n) + \log \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}) - \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]
\end{aligned}$$

Define $\zeta = \sum_{n=1}^{N} \log \sigma(y_n \mathbf{w}^T \mathbf{x}_n) + \log \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}) - \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

1. **Black-box VI based on score function gradients using Monte-Carlo approximation**
   We are given

$$\begin{aligned}
\log q(\mathbf{Z}|\phi) &= \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&= \log \left(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(\frac{(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{w}-\boldsymbol{\mu})}{2} \right) \right)
\end{aligned}$$

   Therefore

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}} \log q(\mathbf{Z}|\phi) &= \nabla_{\boldsymbol{\mu}} \log \left(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(\frac{(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{w}-\boldsymbol{\mu})}{2} \right) \right) \\
&= (\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{w}-\boldsymbol{\mu})
\end{aligned}$$

   and

$$\begin{aligned}
\nabla_{\mathbf{L}} \log q(\mathbf{Z}|\phi) &= \nabla_{\mathbf{L}} \log \left(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(\frac{(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{w}-\boldsymbol{\mu})}{2} \right) \right) \\
&= -\frac{1}{2}\left(\boldsymbol{\Lambda}^\top - \boldsymbol{\Lambda}^\top (\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}^\top \right) \frac{\partial \boldsymbol{\Sigma}}{\partial \mathbf{L}} \\
&= -\left(\mathbf{L}^{-\top} - \mathbf{L}^{-\top}\mathbf{L}^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^\top \mathbf{L}^{-\top} \right)
\end{aligned}$$

Therefore,

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) = \mathbb{E}\left[(\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{w}-\boldsymbol{\mu})\zeta\right]$$
$$\nabla_{\mathbf{L}}\mathcal{L}(q) = \mathbb{E}\left[-\left(\mathbf{L}^{-\top} - \mathbf{L}^{-\top}\mathbf{L}^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^\top\mathbf{L}^{-\top}\right)\zeta\right]$$

Given S samples $\{\mathbf{w}_S\}_{s=1}^S$ from $q(\mathbf{w}|\phi)$, we can get gradients as follows

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left[(\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{w}_S-\boldsymbol{\mu})\zeta_S\right]$$
$$\nabla_{\mathbf{L}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left[-\left(\mathbf{L}^{-\top} - \mathbf{L}^{-\top}\mathbf{L}^{-1}(\mathbf{w}_S-\boldsymbol{\mu})(\mathbf{w}_S-\boldsymbol{\mu})^\top\mathbf{L}^{-\top}\right)\zeta_S\right]$$
$$\text{where } \zeta_S = \sum_{n=1}^{N}\log\sigma(y_n\mathbf{w}_S^T\mathbf{x}_n) + \log\mathcal{N}(\mathbf{w}_S|\mathbf{0},\lambda^{-1}\mathbf{I}) - \log\mathcal{N}(\mathbf{w}_S|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

Algorithm

- Initialize $\phi = \{\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)} = \phi^{(0)}$ and set $t = 1$
- Sample S points from $q(\mathbf{w}|\phi^{(t-1)})$, let them be $\{\mathbf{w}_1\}_{m=1}^M$
- Select B data points randomly, call them $\{\mathbf{x}_n, y_n\}_{n=1}^B$
- Compute gradients of ELBO wrt $\boldsymbol{\mu}$ using the selected data points

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left[(\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{w}_S-\boldsymbol{\mu})\zeta_S\right]$$
$$\text{where } \zeta_S = \sum_{n=1}^{B}\log\sigma(y_n\mathbf{w}_S^T\mathbf{x}_n) + \log\mathcal{N}(\mathbf{w}_S|\mathbf{0},\lambda^{-1}\mathbf{I}) - \log\mathcal{N}(\mathbf{w}_S|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

- Update $\boldsymbol{\mu}$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta\nabla_{\boldsymbol{\mu}}\mathcal{L}(q)$$

- Compute gradients of ELBO wrt $\mathbf{L}$ using the selected points

$$\nabla_{\mathbf{L}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left[-\left(\mathbf{L}^{-\top} - \mathbf{L}^{-\top}\mathbf{L}^{-1}(\mathbf{w}_S-\boldsymbol{\mu})(\mathbf{w}_S-\boldsymbol{\mu})^\top\mathbf{L}^{-\top}\right)\zeta_S\right]$$
$$\text{where } \zeta_S = \sum_{n=1}^{B}\log\sigma(y_n\mathbf{w}_S^T\mathbf{x}_n) + \log\mathcal{N}(\mathbf{w}_S|\mathbf{0},\lambda^{-1}\mathbf{I}) - \log\mathcal{N}(\mathbf{w}_S|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

- Goto step 2 if not converged

2. **Pathwise gradient descent method**: Reparamaterize $\mathbf{w}$ as $\mathbf{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$ where $\mathbf{L} = chol(\boldsymbol{\Sigma})$ or $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}, \mathbf{v} \sim \mathcal{N}(0, \mathbf{I})$ and $q(\mathbf{w}|\phi) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
The ELBO gradient can be written as

$$\nabla_\phi\mathcal{L}(q) = \mathbb{E}_{q_\phi(|\phi)}\left[\log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) - \log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi)\right]$$

Therefore,

$$\nabla_\phi \mathcal{L}(q) = \mathbb{E}_{q_\phi(\mathbf{w}|\phi)}\left[\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}) - \log q(\mathbf{w}|\phi)\right]$$

Now,

$$\log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) = \sum_{n=1}^{N}\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n - \log\left[1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)\right]\right) - \frac{D}{2}\log(2\pi\lambda^{-1}) - \frac{\lambda}{2}((\boldsymbol{\mu} -$$

Taking gradient wrt $\boldsymbol{\mu}$ and $\mathbf{L}$

$$\nabla_{\boldsymbol{\mu}} \log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) = \sum_{n=1}^{N}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})$$

$$\nabla_{\mathbf{L}} \log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) = \left(\sum_{n=1}^{N}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})\right)\mathbf{v}^\top$$

We find gradient of $\log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi)$:

$$\log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi) = -\frac{1}{2}\log(|\mathbf{L}\mathbf{L}^\top|) - \frac{1}{2}\mathbf{v}^\top\mathbf{v}$$

$$\nabla_{\boldsymbol{\mu}} \log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi) = 0$$

$$\nabla_{\mathbf{L}} \log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi) = -\mathbf{L}^{-\top}$$

Given S iid random samples $\{\mathbf{v}_s\}_{s=1}^S$ from $p(\mathbf{v}|0, \mathbf{I})$ we can compute a Monte-Carlo approximation as

$$\nabla_\phi \mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\nabla_\phi\left[\log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) - \log q(\boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\phi)\right]$$

So, the gradients are as follows

$$\nabla_{\boldsymbol{\mu}} \log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) \approx \frac{1}{S}\sum_{s=1}^{S}\sum_{n=1}^{N}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})$$

$$\nabla_{\mathbf{L}} \log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X}) \approx \frac{1}{S}\sum_{s=1}^{S}\left(\sum_{n=1}^{N}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})\right)\mathbf{v}^\top + \mathbf{L}^{-\top}$$

Algorithm

- Initialize $\phi = \phi^{(0)}$ and $t = 1$. Let learning rate for $\boldsymbol{\mu}$ as $\eta_{\boldsymbol{\mu}}$ and $\mathbf{L}$ as $\eta_{\mathbf{L}}$
- Draw S samples $\{\mathbf{v}_s^{(t)}\}_{s=1}^S$ from the distribution $\mathcal{N}(0, \mathbf{I})$
- Pick B random examples $\{\mathbf{x}_n, y_n\}_{n=1}^B$ and update ϕ as follows

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\sum_{n=1}^{B}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})$$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta_{\boldsymbol{\mu}}\nabla_{\boldsymbol{\mu}}\mathcal{L}(q)$$

$$\nabla_{\mathbf{L}} \mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left(\sum_{n=1}^{B}\left(\frac{y_n \mathbf{x}_n}{1 + \exp\left(y_n(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})^\top \mathbf{x}_n\right)}\right) - \lambda(\boldsymbol{\mu} + \mathbf{L}\mathbf{v})\right)\mathbf{v}^\top + \mathbf{L}^{-\top}$$

$$\mathbf{L}^{(t)} = \mathbf{L}^{(t-1)} + \eta_{\mathbf{L}}\nabla_{\mathbf{L}}\mathcal{L}(q)$$

- $t = t + 1$ and goto step 2 if not converged

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ritesh Kumar
Roll Number: 160575
Date: March 14, 2019

QUESTION 5

Question 5 images are in the programming folder