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

QUESTION 1

Student Name: Suryateja B.V.
Roll Number: 160729
Date: March 14, 2019

To solve part 1, we first need to compute $p(\mathbf{t}|\mathbf{X},\mathbf{f},\mathbf{Z})$.

$$p(\mathbf{t}|\mathbf{X},\mathbf{f},\mathbf{Z}) \propto p(\mathbf{f}|\mathbf{X},\mathbf{t},\mathbf{Z})\, p(t|\mathbf{Z}) = \mathcal{N}\left(\mathbf{f}|\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\mathbf{t},\mathbf{P}\right)\mathcal{N}\left(\mathbf{t}|0,\bar{\mathbf{K}}\right)$$

where $\bar{\mathbf{k}} = \left[\bar{\mathbf{k}_1}\bar{\mathbf{k}_2}\cdots\bar{\mathbf{k}_N}\right]$ is an $M \times N$ matrix and $\mathbf{P}$ is a diagonal matrix with each diagonal entry given by $p_{ii} = \kappa(\mathbf{x}_i,\mathbf{x}_i) - \bar{\mathbf{k}_i}^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}_i}$. Consider the exponent in the RHS side. We have,

$$\left(\left(\mathbf{f}-\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\mathbf{t}\right)^T\mathbf{P}^{-1}\left(\mathbf{f}-\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\mathbf{t}\right)+\mathbf{t}^T\bar{\mathbf{K}}^{-1}\mathbf{t}\right) = \begin{bmatrix}\mathbf{t}\\ \mathbf{f}\end{bmatrix}^T\begin{bmatrix}\bar{\mathbf{K}}^{-1}+\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1} & -\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\\ -\mathbf{P}^{-1}\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1} & \mathbf{P}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{f}\end{bmatrix}$$

We know $p(\mathbf{t}|\mathbf{X},\mathbf{f},\mathbf{Z}) = \mathcal{N}\left(\mathbf{t}|\mu_{\mathbf{t}|\mathbf{f}},\Sigma_{\mathbf{t}|\mathbf{f}}\right)$. Using Gaussian conditional formulae, we get

$$\begin{aligned}\Sigma_{\mathbf{t}|\mathbf{f}} = \Lambda_{\mathbf{tt}}^{-1} &= \left(\bar{\mathbf{K}}^{-1}+\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\right)^{-1}\\ &= \bar{\mathbf{K}}-\bar{\mathbf{k}}\left(\mathbf{P}+\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\right)^{-1}\bar{\mathbf{k}}^T = \bar{\mathbf{K}}-\bar{\mathbf{k}}\mathbf{Q}^{-1}\bar{\mathbf{k}}^T \quad \text{(using Woodbury Inverse)}\end{aligned}$$

Note that the variance in $\mathbf{t}$ has reduced after conditioning on $\mathbf{f}$. The matrix $\mathbf{Q}$ is of the form,

$$\mathbf{Q} = \begin{bmatrix}\kappa(\mathbf{x}_1,\mathbf{x}_1) & \bar{\mathbf{k}}_1^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_2 & \cdots & \bar{\mathbf{k}}_1^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_N\\ \bar{\mathbf{k}}_2^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_1 & \kappa(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \bar{\mathbf{k}}_2^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_N\\ \vdots & \vdots & \ddots & \vdots\\ \bar{\mathbf{k}}_N^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_1 & \bar{\mathbf{k}}_N^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_2 & \cdots & \kappa(\mathbf{x}_N,\mathbf{x}_N)\end{bmatrix}$$

We'd prefer the initial form for $\Sigma_{\mathbf{t}|\mathbf{f}}$ which involves an inverse of $M \times M$ matrix instead of the form obtained after using the Woodbury identity, since it involves inverse of $\mathbf{Q}$ which is $N \times N$. The time to compute the initial form is of $\mathcal{O}\left(M^2N\right)$ owing to the computation of $\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}$. Note that computing $\mathbf{P}^{-1}$ is $\mathcal{O}(N)$ since $\mathbf{P}$ is a diagonal matrix. Computing $\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}$ and its transpose take $\mathcal{O}\left(M^2N\right)$, and then multiplying both along with diagonal matrix $\mathbf{P}^{-1}$ also takes $\mathcal{O}\left(M^2N\right)$.

We can write $\mathbf{f} = \bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\mathbf{t}+\epsilon$, where $\epsilon \sim \mathcal{N}(\epsilon|\mathbf{0},\mathbf{P})$. Then, $p(\mathbf{f}|\mathbf{X},\mathbf{Z}) = \mathcal{N}(\mathbf{f}|\mu_{\mathbf{f}},\Sigma_{\mathbf{f}})$ where $\mu_{\mathbf{f}} = \mathbb{E}[\mathbf{f}] = \bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}(\mathbf{0}) + \mathbf{0} = \mathbf{0}$ and $\Sigma_{\mathbf{f}} = \bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}} + \mathbf{P} = \mathbf{Q}$. Using $\mu_{\mathbf{f}}$, we can now compute $\mu_{\mathbf{t}|\mathbf{f}}$ as follows :

$$\mu_{\mathbf{t}|\mathbf{f}} = \mu_{\mathbf{t}} - \Lambda_{\mathbf{tt}}^{-1}\Lambda_{\mathbf{tf}}(\mathbf{f}-\mu_{\mathbf{f}}) = \Sigma_{\mathbf{t}|\mathbf{f}}\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\mathbf{f}$$

We know $p(y_\star|\mathbf{x}_\star,\mathbf{f},\mathbf{X},\mathbf{Z}) = \int p(y_\star|\mathbf{x}_\star,\mathbf{f},\mathbf{X},\mathbf{Z},\mathbf{t})\, p(\mathbf{t}|\mathbf{f},\mathbf{X},\mathbf{Z})\, d\mathbf{t}$.
Now, we can write $y_\star = \mathbf{f}_\star = \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\mathbf{t}+\epsilon$, with $\mathbf{t} \sim \mathcal{N}\left(\mu_{\mathbf{t}|\mathbf{f}},\Sigma_{\mathbf{t}|\mathbf{f}}\right)$ and $\epsilon \sim \mathcal{N}\left(\epsilon|\mathbf{0},\kappa(\mathbf{x}_\star,\mathbf{x}_\star) - \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_\star\right)$.
Then, $p(y_\star|\mathbf{x}_\star,\mathbf{f},\mathbf{X},\mathbf{Z}) = \mathcal{N}(y_\star|\mu_\star,\Sigma_\star)$ where

$$\begin{aligned}\mu_\star &= \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\mu_{\mathbf{t}|\mathbf{f}} = \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\Sigma_{\mathbf{t}|\mathbf{f}}\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\mathbf{f} = \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\left(\bar{\mathbf{K}}^{-1}+\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\right)^{-1}\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}\mathbf{P}^{-1}\mathbf{f}\\ \Sigma_\star &= \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\Sigma_{\mathbf{t}|\mathbf{f}}\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_\star + \kappa(\mathbf{x}_\star,\mathbf{x}_\star) - \bar{\mathbf{k}}_\star^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_\star\end{aligned}$$

Before, the posterior predictive involved the inversion of $\mathbf{K}$, which takes $\mathcal{O}\left(N^3\right)$ time complexity. With pesudo training data, we see that computation of $\Sigma_{\mathbf{t}|\mathbf{f}}$ takes $\mathcal{O}\left(M^2N\right)$, a significant reduction in time taken.

For part (2), we derive the MLE-II solution, that is, maximise the marginal likelihood $p\left(\mathbf{f}|\mathbf{X},\mathbf{Z}\right)$. Earlier, we have derived the same - $p\left(\mathbf{f}|\mathbf{X},\mathbf{Z}\right) = \mathcal{N}\left(\mathbf{f}|\mu_{\mathbf{f}},\Sigma_{\mathbf{f}}\right)$ where $\mu_{\mathbf{f}} = \mathbf{0}$ and $\Sigma_{\mathbf{f}} = \bar{\mathbf{k}}^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{k}} + \mathbf{P} = \mathbf{Q}$. The MLE-II objective is given by,

$$\arg\max_{\mathbf{Z}} \log\left(p\left(\mathbf{f}|\mathbf{X},\mathbf{Z}\right)\right) = \arg\min_{\mathbf{Z}} \left(\log|\Sigma_{\mathbf{f}}| + \mathbf{f}^T\Sigma_{\mathbf{f}}^{-1}\mathbf{f}\right)$$

There is dependence on $\mathbf{Z}$ through $\mathbf{Q}$, as in, $\bar{\mathbf{k}}_i = \left[\kappa\left(\mathbf{x}_i,\mathbf{z}_1\right)\kappa\left(\mathbf{x}_i,\mathbf{z}_2\right)\cdots\kappa\left(\mathbf{x}_i,\mathbf{z}_M\right)\right]$

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

Student Name: Suryateja B.V.
Roll Number: 160729
Date: March 14, 2019

QUESTION

2

0.1 EM - I : not involving $\mathbf{z}_n$

We first marginalize over $\mathbf{z}_n$. Let us compute $p(\mathbf{x}_n|c_n = m, \Theta)$. Given $c_n = m$, we can write $\mathbf{x}_n$ as $\mathbf{x}_n = \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \boldsymbol{\epsilon}_n$ where $\mathbf{z}_n \sim \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{I}_K)$ and $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\boldsymbol{\epsilon}_n|\mathbf{0}, \sigma_m^2\mathbf{I}_D)$

$$\mathbb{E}[\mathbf{x}_n] = \boldsymbol{\mu}_m \quad \text{and} \quad V(\mathbf{x}_n) = \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D = \mathbf{V}_m$$

So, $p(\mathbf{x}_n|c_n = m, \Theta) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)$
Now let us compute the posterior probabilities of latent variables.

$$p(c_n = m|\mathbf{x}_n, \Theta) \propto p(\mathbf{x}_n|c_n = m, \Theta)\, p(c_n = m) = \pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{V}_m)$$
$$\implies p(c_n = m|\mathbf{x}_n, \Theta) = \frac{\pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{V}_m)}{\sum_{l=1}^{M}\pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{V}_l)} = r_{nm}$$

The CLL is given as follows :

$$p(\mathbf{x}, \mathbf{c}|\Theta) = \prod_{n=1}^{N}\prod_{m=1}^{M}\left(p(\mathbf{x}_n|c_n = m, \Theta)\, p(c_n = m|\Theta)\right)^{\mathbb{I}[c_n=m]}$$
$$\implies \log p(\mathbf{x}, \mathbf{c}|\Theta) = \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{I}[c_n = m]\left(\log\pi_m - \frac{1}{2}\log|\mathbf{V}_m| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^T\mathbf{V}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m)\right)$$

The only expectation we need is $\mathbb{E}[\mathbb{I}[c_n = m]] = r_{nm}$. So the expected CLL is given by,

$$\sum_{n=1}^{N}\sum_{m=1}^{M}r_{nm}\left(\log\pi_m - \frac{1}{2}\log|\mathbf{V}_m| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^T\mathbf{V}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m)\right) \tag{1}$$

The M-step update equations are as follows :

$$\hat{\pi}_m = \frac{\sum_{n=1}^{N}r_{nm}}{N} = \frac{N_m}{N}$$
$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{n=1}^{N}r_{nm}\mathbf{x}_n}{N_m}$$
$$\hat{\mathbf{V}}_m = \frac{\sum_{n=1}^{N}r_{nm}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m)^T}{N_m}$$

$$\mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D = \hat{\mathbf{V}}_m \implies \hat{\mathbf{W}}_m = \mathbf{U}_K\left(\mathbf{L}_K - \hat{\sigma}_m^2\mathbf{I}_K\right)^{1/2}\mathbf{R} \quad \text{and} \quad \hat{\sigma}_m^2 = \frac{1}{D-K}\sum_{k=K+1}^{D}\lambda_k$$

where $\mathbf{U}_K$ is a $D \times K$ matrix of top K eigen vectors of $\hat{\mathbf{V}}_m$, $\mathbf{L}_K$: $K \times K$ diagonal matrix of top K eigen values $\lambda_1, \lambda_2, \cdots, \lambda_K$, $\mathbf{R}$ is a $K \times K$ rotation matrix. While this method avoids z_n estimates, it is expensive due to eigen decomposition.

1 Initialize $\Theta = \left\{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\right\}_{m=1}^{M} = \Theta^{(0)}$. Set t = 1;
2 **E-Step**

$$r_{nm}^{(t)} = \frac{\pi_m^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_m^{(t-1)}, \mathbf{V}_m^{(t-1)}\right)}{\sum_{l=1}^{M} \pi_m^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_l^{(t-1)}, \mathbf{V}_l^{(t-1)}\right)} \quad \forall n, m$$

;
3 **M-Step** - Do for all m

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

;
4 Go to E-Step if not converged.

The stepwise online algorithm sketch is as follows :

1 Initialize $\Theta = \left\{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\right\}_{m=1}^{M} = \Theta^{(0)}$. Set t = 1;
2 Pick a random example $\mathbf{x}_n$;
3 Compute $r_{nm}^{(t)}$ for all m;
4 Compute learning rate ϵ_t;
5 Compute $\hat{\Theta}$ using only example $\mathbf{x}_n$;
6 $\Theta^{(t)} = (1 - \epsilon_t)\,\Theta^{(t-1)} + \epsilon_t \hat{\Theta}$;
7 Go to Step-2 if Θ not converged;

0.2 EM - II : Include Estimating $\mathbf{z}_n$

Conditional posterior of c_n is same as before (since it is independent of $\mathbf{z}_n$), and is given by:

$$p\left(c_n = m|\mathbf{x}_n, \mathbf{z}_n, \Theta\right) = \frac{\pi_m \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{V}_m\right)}{\sum_{l=1}^{M} \pi_m \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{V}_l\right)} = r_{nm}$$

We sort of have a mixture of M PPCA models. Each model coming from a particular class m. For a single PPCA model we had a latent variable $\mathbf{z}_n$ for each $\mathbf{x}_n$. Now, in case of mixture of PPCA, we have M latent variables $\{\mathbf{z}_{nm}\}_{m=1}^{M}$ for each $\mathbf{x}_n$. Conditional posterior of $\mathbf{z}_{nm}$ is as follows:

$$p\left(\mathbf{z}_{nm}|\mathbf{x}_n, c_n = m, \Theta\right) \propto p\left(\mathbf{x}_n|\mathbf{z}_{nm}, c_n = m, \Theta\right) p\left(\mathbf{z}_{nm}|\Theta\right) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_{nm}, \sigma_m^2\mathbf{I}_D\right) \mathcal{N}\left(\mathbf{z}_{nm}|0, \mathbf{I}_K\right)$$

Using Gaussian conditional properties, we get $p\left(\mathbf{z}_{nm}|\mathbf{x}_n, c_n = m, \Theta\right) = \mathcal{N}\left(\mathbf{z}_{nm}|\boldsymbol{\mu}_{nm}, \boldsymbol{\Sigma}_{nm}\right)$ where $\boldsymbol{\Sigma}_{nm} = \sigma_m^2\left(\mathbf{W}_m^T\mathbf{W}_m + \sigma_m^2\mathbf{I}_K\right)^{-1} = \sigma_m^2\mathbf{M}_m^{-1}$ and $\boldsymbol{\mu}_{nm} = \mathbf{M}_m^{-1}\mathbf{W}_m^T\left(\mathbf{x}_n - \boldsymbol{\mu}_m\right)$.
The joint posterior distribution is as follows :

$$p\left(\mathbf{z}_{nm}, c_n = m|\mathbf{x}_n, \Theta\right) = p\left(\mathbf{z}_{nm}|\mathbf{x}_n, c_n = m, \Theta\right) p\left(c_n = m|\mathbf{x}_n, \Theta\right) = r_{nm}\mathcal{N}\left(\mathbf{z}_{nm}|\mu_{nm}, \boldsymbol{\Sigma}_{nm}\right)$$

Note that $\mathbf{z}_{nm}$ inherently holds the information that it belongs to m^{th} class, so we can say $p\left(\mathbf{z}_{nm}|\mathbf{x}_n, c_n = m, \Theta\right) = p\left(\mathbf{z}_{nm}|\mathbf{x}_n, \Theta\right)$. This means that the joint posterior contains two independent terms. Hence when we take expectations wrt joint posterior (as in M-step), we can decouple and take expectation wrt individual terms $p\left(\mathbf{z}_{nm}|\mathbf{x}_n, \Theta\right)$ and $p\left(c_n = m|\mathbf{x}_n, \Theta\right)$.

The CLL is given as follows :

$$p\left(\mathbf{x}, \mathbf{c}, \mathbf{z}|\Theta\right) = \prod_{n=1}^{N}\prod_{m=1}^{M}\left(p\left(\mathbf{x}_n|\mathbf{z}_n, c_n = m, \Theta\right) p\left(\mathbf{z}_{nm}|c_n = m, \Theta\right) p\left(c_n = m|\Theta\right)\right)^{\mathbb{I}[c_n=m]}$$

$$\log p\left(\mathbf{x}, \mathbf{c}|\Theta\right) =$$

$$\sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{I}\left[c_n = m\right]\left(\log\pi_m - \frac{D}{2}\log\sigma_m^2 - \frac{1}{2\sigma_m^2}\left(\mathbf{x}_n - \boldsymbol{\mu}_m - \mathbf{W}_m\mathbf{z}_{nm}\right)^T\left(\mathbf{x}_n - \boldsymbol{\mu}_m - \mathbf{W}_m\mathbf{z}_{nm}\right) - \frac{1}{2}\mathbf{z}_{nm}^T\mathbf{z}_{nm}\right)$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{I}\left[c_n = m\right]\left(\log\pi_m - \frac{D}{2}\log\sigma_m^2 - \frac{1}{2\sigma_m^2}\left\|\mathbf{x}_n - \boldsymbol{\mu}_m\right\|^2 - \frac{1}{2\sigma_m^2}\mathrm{tr}\left(\mathbf{z}_{nm}\mathbf{z}_{nm}^T\mathbf{W}_m^T\mathbf{W}_m\right)\right.$$
$$\left. + \frac{2}{2\sigma_m^2}\left(\mathbf{x}_n - \boldsymbol{\mu}_m\right)^T\mathbf{W}_m\mathbf{z}_{nm} - \frac{1}{2}\mathrm{tr}\left(\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right)\right)$$

$$\mathrm{ECLL} = \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}\left[\mathbb{I}\left[c_n = m\right]\right]\left(\log\pi_m - \frac{D}{2}\log\sigma_m^2 - \frac{1}{2\sigma_m^2}\left\|\mathbf{x}_n - \boldsymbol{\mu}_m\right\|^2 - \frac{1}{2\sigma_m^2}\mathrm{tr}\left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\mathbf{W}_m^T\mathbf{W}_m\right)\right.$$
$$\left. + \frac{2}{2\sigma_m^2}\left(\mathbf{x}_n - \boldsymbol{\mu}_m\right)^T\mathbf{W}_m\mathbb{E}\left[\mathbf{z}_{nm}\right] - \frac{1}{2}\mathrm{tr}\left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\right)\right)$$

where $\mathbb{E}\left[\mathbf{z}_{nm}\right] = \boldsymbol{\mu}_{nm}$ and $\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right] = \boldsymbol{\mu}_{nm}\boldsymbol{\mu}_{nm}^T + \Sigma_{nm}$ and $\mathbb{E}\left[\mathbb{I}\left[c_n = m\right]\right] = r_{nm}$

The M-Step updates are given as follows :

$$\hat{\pi}_m = \frac{\sum_{n=1}^{N} r_{nm}}{N} = \frac{N_m}{N}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{n=1}^{N} r_{nm} \left(\mathbf{x}_n - \mathbf{W}_m \mathbb{E}\left[\mathbf{z}_{nm}\right]\right)}{N_m}$$

$$\hat{\mathbf{W}}_m = \left(\sum_{n=1}^{N} r_{nm} \left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m\right) \mathbb{E}\left[\mathbf{z}_{nm}\right]^T\right) \left(\sum_{n=1}^{N} r_{nm} \mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\right)^{-1}$$

$$\hat{\sigma}_m^2 = \frac{1}{DN_m} \left(\sum_{n=1}^{N} \left(\left\|\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m\right\|^2 - 2\mathbb{E}\left[\mathbf{z}_{nm}\right]^T \hat{\mathbf{W}}_m^T \left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m\right) + \operatorname{tr}\left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right] \hat{\mathbf{W}}_m^T \hat{\mathbf{W}}_m\right)\right)\right)$$

1 Initialize $\Theta = \left\{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\right\}_{m=1}^{M} = \Theta^{(0)}$. Set t = 1;

2 **E-Step** $\forall n, m$

$$r_{nm}^{(t)} = \frac{\pi_m^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_m^{(t-1)}, \mathbf{V}_m^{(t-1)}\right)}{\sum_{l=1}^{M} \pi_m^{(t-1)} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_l^{(t-1)}, \mathbf{V}_l^{(t-1)}\right)}$$

$$\mathbf{M}_m^{(t)} = \left(\mathbf{W}_m^T\right)^{(t-1)} \mathbf{W}_m^{(t-1)} + \left(\sigma_m^2\right)^{(t-1)} \mathbf{I}_K$$

$$\mathbb{E}\left[\mathbf{z}_{nm}\right]^{(t)} = \left(\mathbf{M}_m^{-1}\right)^{(t)} \left(\mathbf{W}_m^T\right)^{(t-1)} \left(\mathbf{x}_n - \boldsymbol{\mu}_m\right)$$

$$\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]^{(t)} = \mathbb{E}\left[\mathbf{z}_{nm}\right]^{(t-1)} \left(\mathbb{E}\left[\mathbf{z}_{nm}\right]^T\right)^{(t)} + \left(\sigma_m^2\right)^{(t-1)} \left(\mathbf{M}_m^{-1}\right)^{(t)}$$

;

3 **M-Step** - Do for all m

$$\hat{\pi}_m^{(t)} = \frac{\sum_{n=1}^{N} r_{nm}^{(t)}}{N} = \frac{N_m^{(t)}}{N}$$

$$\hat{\boldsymbol{\mu}}_m^{(t)} = \frac{\sum_{n=1}^{N} r_{nm}^{(t)} \left(\mathbf{x}_n - \mathbf{W}_m^{(t-1)} \mathbb{E}\left[\mathbf{z}_{nm}\right]^{(t)}\right)}{N_m^{(t)}}$$

$$\hat{\mathbf{W}}_m^{(t)} = \left(\sum_{n=1}^{N} r_{nm}^{(t)} \left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m^{(t)}\right) \left(\mathbb{E}\left[\mathbf{z}_{nm}\right]^T\right)^{(t)}\right) \left(\sum_{n=1}^{N} r_{nm}^{(t)} \left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\right)^{(t)}\right)^{-1}$$

$$\left(\hat{\sigma}_m^2\right)^{(t)} = \frac{1}{DN_m^{(t)}} \sum_{n=1}^{N} r_{nm}^{(t)} \left(\left\|\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m^{(t)}\right\|^2 - 2 \left(\mathbb{E}\left[\mathbf{z}_{nm}\right]^T\right)^{(t)} \left(\hat{\mathbf{W}}_m^T\right)^{(t)} \left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m^{(t)}\right) + \operatorname{tr}\left(\left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\right)^{(t)} \left(\hat{\mathbf{W}}_m^T \hat{\mathbf{W}}_m\right)^{(t)}\right) \right)$$

$$t = t + 1$$

;

4 Go to E-Step if not converged.

The stepwise online algorithm sketch is as follows :

1 Initialize $\Theta = \left\{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\right\}_{m=1}^{M} = \Theta^{(0)}$. Set t = 1;
2 Pick a random example $\mathbf{x}_n$;
3 Compute $r_{nm}^{(t)}, \mathbf{M}_m^{(t)}, \left(\mathbb{E}\left[\mathbf{z}_{nm}\right]\right)^{(t)}, \left(\mathbb{E}\left[\mathbf{z}_{nm}\mathbf{z}_{nm}^T\right]\right)^{(t)}$ for all m;
4 Compute learning rate ϵ_t;
5 Compute $\hat{\Theta}$ using only example $\mathbf{x}_n$;
6 $\Theta^{(t)} = (1 - \epsilon_t)\,\Theta^{(t-1)} + \epsilon_t \hat{\Theta}$;
7 Go to Step-2 if Θ not converged;

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

QUESTION 3

Student Name: Suryateja B.V.
Roll Number: 160729
Date: March 14, 2019

Let the mean field approximation of posterior be given by $q(\mathbf{w}, \beta, \alpha) = q(\mathbf{w}|\phi_1)\, q(\beta|\phi_2) \prod_{d=1}^{D} q(\alpha_d|\phi_d)$. The complete data log likelihood is given by,

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha|\mathbf{X}, \theta) &= \log p(\mathbf{y}|\mathbf{w}, \beta, \alpha, \mathbf{X}, \theta) + \log p(\mathbf{w}|\alpha, \theta) + \log p(\beta|\theta) + \sum_{d=1}^{D} \log p(\alpha_d|\theta) \\ &= \log \beta \left(\frac{N}{2} + a_0 - 1\right) - \frac{\beta}{2}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + 2b_0\right) \\ &\quad - \frac{1}{2}\sum_{d=1}^{D} w_d^2 \alpha_d + \left(e_0 - \frac{1}{2}\right)\sum_{d=1}^{D} \log \alpha_d - f_0 \sum_{d=1}^{D} \alpha_d + \text{const.}\end{aligned}$$

where $\theta = \{a_0, b_0, e_0, f_0\}$, $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_D]$, $\mathbf{y} = [y_1, y_2, \cdots, y_N]$ and $\mathbf{X}$ is the matrix of $\{\mathbf{x}_n\}_{n=1}^{N}$. Now, using the mean-field VI algorithm, the optimal distributions are given as follows.

$$\begin{aligned}\log \hat{q}(\mathbf{w}) &= \mathbb{E}_{q(\alpha)q(\beta)}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha|\mathbf{X}, \theta)\right] \\ &= -\frac{\mathbb{E}[\beta]}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\left(\mathbf{X}^T\mathbf{X} + \frac{1}{\mathbb{E}[\beta]}\mathbb{E}_\alpha\left[diag(\alpha_1, \cdots, \alpha_D)\right]\right)\mathbf{w} + 2b_0\right) \\ &\quad + \text{constant terms}\end{aligned}$$

Comparing the above with the log of normal distribution $\mathcal{N}(\mathbf{w}|\mu_N, \Sigma_N)$, we get,

$$\begin{aligned}\Sigma_N &= \left(\mathbb{E}[\beta]\,\mathbf{X}^T\mathbf{X} + \mathbb{E}_\alpha\left[diag(\alpha_1, \cdots, \alpha_D)\right]\right)^{-1} \\ \mu_N &= \Sigma_N \mathbb{E}[\beta]\,\mathbf{X}^T\mathbf{y}\end{aligned}$$

Now, let us solve for β.

$$\begin{aligned}\log \hat{q}(\beta) &= \mathbb{E}_{q(\alpha)q(\mathbf{w})}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha|\mathbf{X}, \theta)\right] \\ &= \log \beta \left(\frac{N}{2} + a_0 - 1\right) - \frac{\beta}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}] + \text{tr}\left(\mathbf{X}^T\mathbf{X}\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right]\right) + 2b_0\right) + \text{const.}\end{aligned}$$

Comparing the above with the log of Gamma$(\beta|a_N, b_N)$, we get,

$$\begin{aligned}a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}] + \text{tr}\left(\mathbf{X}^T\mathbf{X}\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right]\right)\right)\end{aligned}$$

Next, we solve for α.

$$\begin{aligned}\log \hat{q}(\alpha) &= \mathbb{E}_{q(\beta)q(\mathbf{w})}\left[\log p(\mathbf{y}, \mathbf{w}, \beta, \alpha|\mathbf{X}, \theta)\right] \\ &= -\frac{1}{2}\sum_{d=1}^{D} \mathbb{E}[w_d]^2 \alpha_d + \left(e_0 - \frac{1}{2}\right)\sum_{d=1}^{D} \log \alpha_d - f_0 \sum_{d=1}^{D} \alpha_d + \text{const.}\end{aligned}$$

Comparing the above with $\prod_{d=1}^{D} \text{Gamma}\left(\alpha_d | e_{Nd}, f_{Nd}\right)$, we get

$$
\begin{aligned}
e_{Nd} &= e_0 + \frac{1}{2} \quad \forall d \\
f_{Nd} &= f_0 + \frac{1}{2}\mathbb{E}\left[w_d^2\right]
\end{aligned}
$$

Now, let us write down the expectations.

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{w}\right] &= \mu_N \\
\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right] &= \Sigma_N + \mu_N \mu_N^T \\
\mathbb{E}\left[\beta\right] &= \frac{a_N}{b_N} \\
\mathbb{E}\left[\alpha_d\right] &= \frac{e_{Nd}}{f_{Nd}} \\
\mathbb{E}\left[w_d^2\right] &= \left(\Sigma_N\right)_{dd} + \mu_{Nd}^2
\end{aligned}
$$

Note that the updates of each of the optimum distributions depend on other distributions. So, we are required to perform cyclic updates. The algorithm is as follows: $\mathbf{P}$ and $\mathbf{K}$ are introduced

Given : $\mathbf{X}$, $\mathbf{y}, a_0, b_0, e_0, f_0$.;
Set $e_{Nd} = e_0 + \frac{1}{2} \quad \forall d$. Set $a_N = a_0 + \frac{N}{2}$;
Set $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. Set $\mathbf{P} = \mathbf{X}^T\mathbf{y}$. Set $\mathbf{Q} = \mathbf{y}^T\mathbf{y}$;
Initialize $b_N^{(0)} = b_0 + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} + \text{tr}\left(\mathbf{K}\right)\right)$;
Initialize $f_{Nd}^{(0)} = f_0 + \frac{1}{2} \quad \forall d$;
Initialize $\mathbb{E}\left[\beta\right]^{(0)} = \frac{a_N}{b_N^{(0)}}$;
Initialize $\mathbb{E}\left[\alpha_d\right]^{(0)} = \frac{e_{Nd}}{f_{Nd}^{(0)}} \quad \forall d$;
Set t = 0. While ELBO not converged:

$$
\begin{aligned}
\Sigma_N^{(t+1)} &= \left(\mathbb{E}\left[\beta\right]^{(t)}\mathbf{K} + \mathbb{E}_\alpha\left[diag\left(\alpha_1, \cdots, \alpha_D\right)\right]^{(t)}\right)^{-1} \\
\mu_N^{(t+1)} &= \Sigma_N^{(t+1)}\mathbb{E}\left[\beta\right]^{(t)}\mathbf{P} \\
b_N^{(t+1)} &= b_0 + \frac{1}{2}\left(\mathbf{Q} - 2\mathbf{P}^T\mathbb{E}\left[\mathbf{w}\right]^{(t+1)} + \text{tr}\left(\mathbf{K}\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right]^{(t+1)}\right)\right) \\
\mathbb{E}\left[\beta\right]^{(t+1)} &= \frac{a_N}{b_N^{(t+1)}} \\
f_{Nd}^{(t+1)} &= f_0 + \frac{1}{2}\mathbb{E}\left[w_d^2\right]^{(t+1)} \quad \forall d \\
\mathbb{E}\left[\alpha_d\right]^{(t+1)} &= \frac{e_{Nd}}{f_{Nd}^{(t+1)}} \quad \forall d \\
t &= t + 1
\end{aligned}
$$

to avoid recomputation. Instead of making random initializations, we have assumed arbitarily that $\mathbf{w}$ is initially a standard normal variable.

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

Student Name: Suryateja B.V.
Roll Number: 160729
Date: March 14, 2019

QUESTION

4

1 BBVI

$$\nabla_\phi L(\phi) = \mathbb{E}_q\left[\nabla_\phi \log q(\mathbf{w}|\phi)\left(\sum_{n=1}^{N}\log\sigma\left(y_n\mathbf{w}^T\mathbf{x}_n\right) + \log p(\mathbf{w}) - \log q(\mathbf{w}|\phi)\right)\right]$$

$$\nabla_\mu \log q(\mathbf{w}|\phi) = \frac{d}{d\boldsymbol{\mu}}\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})\right) = \Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu}) = \left(\mathbf{L}\mathbf{L}^T\right)^{-1}(\mathbf{w}-\boldsymbol{\mu})$$

$$\begin{aligned}\nabla_{\mathbf{L}} \log q(\mathbf{w}|\phi) &= \frac{d}{d\mathbf{L}}\left(-\frac{1}{2}\log|\mathbf{L}\mathbf{L}^T|\right) + \frac{d}{d\boldsymbol{\Sigma}}\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)\frac{d\boldsymbol{\Sigma}}{d\mathbf{L}} \\ &= -\mathbf{L}^{-T} + \left(\left(\mathbf{L}\mathbf{L}^T\right)^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T\left(\mathbf{L}\mathbf{L}^T\right)^{-1}\right)\mathbf{L}\end{aligned}$$

$$\log p(\mathbf{w}) = \frac{D}{2}\log\frac{\lambda}{2\pi} - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$\log q(\mathbf{w}|\phi) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{L}\mathbf{L}^T| - \frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T\left(\mathbf{L}\mathbf{L}^T\right)^{-1}(\mathbf{w}-\boldsymbol{\mu})$$

Instead of taking an expectation wrt $q(w|\phi)$, we draw S samples from the distribution and do a Monte carlo approximation. The algorithm for BBVI with minibatch size of 1 is given in the next page. We note that the gradients derived does not depend on the type of model. Hence this algorithm is like a black-box (agnostic).

1 Initialize $\phi = \phi^{(0)} = \left\{\boldsymbol{\mu}^{(0)}, \mathbf{L}^{(0)}\right\}$. Set t = 1 ;
2 Pick a random example n;
3 Draw S samples $\left\{\mathbf{w}_s^{(t)}\right\}_{s=1}^{S}$ from $q\left(\mathbf{w}|\phi^{(t-1)}\right)$;
4 Compute learning rates $\eta_\mu^{(t)}$, $\eta_L^{(t)}$;
5

$$\nabla_\mu L\left(\phi^{(t)}\right) = \frac{1}{S}\sum_{s=1}^{S}\left(\nabla_\mu \log q\left(\mathbf{w}_s^{(t)}|\phi^{(t-1)}\right)\left(\log\sigma\left(y_n\left(\mathbf{w}_s^T\right)^{(t)}\mathbf{x}_n\right) + \log p\left(\mathbf{w}_s^{(t)}\right) - \log q\left(\mathbf{w}_s^{(t)}|\phi^{(t-1)}\right)\right)\right)$$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta_\mu^{(t)}\nabla_\mu L\left(\phi^{(t)}\right)$$

$$\nabla_L L\left(\phi^{(t)}\right) = \frac{1}{S}\sum_{s=1}^{S}\left(\nabla_L \log q\left(\mathbf{w}_s^{(t)}|\boldsymbol{\mu}^{(t)}, \mathbf{L}^{(t-1)}\right)\left(\log\sigma\left(y_n\left(\mathbf{w}_s^T\right)^{(t)}\mathbf{x}_n\right) + \log p\left(\mathbf{w}_s^{(t)}\right) - \log q\left(\mathbf{w}_s^{(t)}|\boldsymbol{\mu}^{(t)}, \mathbf{L}^{(t-1)}\right)\right)\right)$$

$$\mathbf{L}^{(t)} = \mathbf{L}^{(t-1)} + \eta_L^{(t)}\nabla_L L\left(\phi^{(t)}\right)$$

$$t = t + 1$$

6 Go to Step-2 if ELBO not converged.

2 Reparametrization trick

Here we write $\mathbf{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{v}|0,1)$. So, the elbo gradient is as follows -

$$\nabla_\phi L(\phi) = \mathbb{E}_q\left[\nabla_\phi\left(\sum_{n=1}^{N}\log\sigma\left(y_n\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)^T\mathbf{x}_n\right) + \log p\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right) - \log q\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}|\phi\right)\right)\right]$$

$$\nabla_\phi L(\phi) = \mathbb{E}_{p(\mathbf{v})}\left[\nabla_\phi f\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)\right]$$

$$\log p\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right) = \frac{D}{2}\log\frac{\lambda}{2\pi} - \frac{\lambda}{2}\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)^T\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)$$

$$\log q\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}|\phi\right) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{L}\mathbf{L}^T| - \frac{1}{2}\mathbf{v}^T\mathbf{v}$$

$$\nabla_\mu f\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right) = \sum_{n=1}^{N}\left(1-\sigma\left(y_n\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)^T\mathbf{x}_n\right)\right)y_n\mathbf{x}_n - \lambda\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right) - 0$$

$$\nabla_{\mathbf{L}} f\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right) = \sum_{n=1}^{N}\left(1-\sigma\left(y_n\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)^T\mathbf{x}_n\right)\right)y_n\mathbf{x}_n\mathbf{v}^T - \lambda\left(\boldsymbol{\mu}+\mathbf{L}\mathbf{v}\right)\mathbf{v}^T - \mathbf{L}^{-T}$$

We use monte carlo gradien again, by drawing S samples from the distribution $p(\mathbf{v})$. Algorithm is given in the next page

1 Initialize $\phi = \phi^{(0)} = \left\{\boldsymbol{\mu}^{(0)}, \mathbf{L}^{(0)}\right\}$. Set t = 1 ;

2 Pick a random example n;

3 Draw S samples $\left\{\mathbf{v}_s^{(t)}\right\}_{s=1}^{S}$ from $\mathcal{N}\left(\mathbf{v}|0, \mathbf{I}\right)$;

4 Compute learning rates $\eta_\mu^{(t)}$, $\eta_L^{(t)}$;

5

$$\nabla_\mu L\left(\phi^{(t)}\right) = \frac{1}{S}\sum_{s=1}^{S} \nabla_\mu f\left(\boldsymbol{\mu}^{(t-1)} + \mathbf{L}^{(t-1)}\mathbf{v}_s^{(t)}\right)$$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \eta_\mu^{(t)} \nabla_\mu L\left(\phi^{(t)}\right)$$

$$\nabla_L L\left(\phi^{(t)}\right) = \frac{1}{S}\sum_{s=1}^{S} \nabla_L f\left(\boldsymbol{\mu}^{(t)} + \mathbf{L}^{(t-1)}\mathbf{v}_s^{(t)}\right)$$

$$\mathbf{L}^{(t)} = \mathbf{L}^{(t-1)} + \eta_L^{(t)} \nabla_L L\left(\phi^{(t)}\right)$$

$$t = t + 1$$

6 Go to Step-2 if ELBO not converged.

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

Student Name: Suryateja B.V.
Roll Number: 160729
Date: March 14, 2019

QUESTION 5

A few notes -

- logsumexp trick has been used to calculate since some probabilities are very small.
- The various mean images obtained resemble original numbers.
- Online EM is as expected very fast compared to full batch EM.
- Sometimes mean images, repeat, as in, resemble almost same numbers. This is due to either poor initialization, or the algorithm has captured two different types of the same digit. For example, 0 tilted towards right and 0 tilted towards left might be two different mean images.
- All images are available in submission file. You can also view images by varying K in the Jupyter notebook file.