Student Name: Subham Kumar
Roll Number: 160707
Date: March 14, 2019

QUESTION 1

1.Given that

$$p(f_n|\mathbf{x_n}, \mathbf{Z}, \mathbf{t}) = \mathcal{N}\left(\mathbf{x_n}|\mathbf{k}_n^T\mathbf{K}_M^{-1}\mathbf{t}, \kappa(x_n, x_n) - \mathbf{k}_n^T\mathbf{K}_M^{-1}\mathbf{k}_n\right)$$

$$p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) = \prod_{n=1}^{N} p(f_n|\mathbf{x_n}, \mathbf{Z}, \mathbf{t})$$

$$= \mathcal{N}\left(\mathbf{f}|\mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{t}, \mathbf{\Lambda}\right)$$

Here $\mathbf{K}_{NM}$ is $N \times M$ matrix with $[\mathbf{K}_{NM}]_{nm} = \kappa(\mathbf{x}_n, \mathbf{z}_m)$ 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 $[\mathbf{\Lambda}]_{ii} = \kappa(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{k}_n^T\mathbf{K}_M^{-1}\mathbf{k}_n$. 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 posterior over $\mathbf{t}$,we have:

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

Since the pseudo sample points are modelled by same G.P.,we have $p(\mathbf{t}|\mathbf{Z}) = \mathbf{N}(\mathbf{t}|0, \mathbf{K}_M)$.Writing the terms in exponent in the R.H.S. of the above proportionality in information form of Gaussian,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 $\mathbf{\Sigma}_{\mathbf{t}|\mathbf{f}} = \left(\mathbf{K}_M^{-1}\mathbf{K}_{NM}^T\Lambda^{-1}\mathbf{K}_{NM}\mathbf{K}_M^{-1} + \mathbf{K}_M^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{t}|\mathbf{f}} = \mathbf{\Sigma}_{\mathbf{t}|\mathbf{f}}\mathbf{K}_M^{-1}\mathbf{K}_{NM}^T\mathbf{\Lambda}^{-1}\mathbf{f}$.
Since $y_* = f_*$.We can write $f_* = \mathbf{k}_*^T\mathbf{K}_M^{-1}\mathbf{t} + \epsilon$, where $\mathbf{k}_*$ is $M \times 1$ vector with $[\mathbf{k}_*]_i = \kappa(\mathbf{x}_*, \mathbf{z}_i)$, and $\epsilon = \mathcal{N}(0, \kappa\left(\mathbf{x}_*, \mathbf{x}_*\right) - \mathbf{k}_*^T\mathbf{K}_M^{-1}\mathbf{k}_*^T)$ Now using the property of linear gaussian model we get,

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) = \mathcal{N}\left(f_*|\boldsymbol{\mu}_*, \mathbf{\Sigma}_*\right)$$

$$\text{where } \boldsymbol{\mu}_* = \mathbf{k}_*^T\mathbf{K}_M^{-1}\mathbf{\Sigma}_{\mathbf{t}|\mathbf{f}}\mathbf{K}_M^{-1}\mathbf{K}_{NM}^T\mathbf{\Lambda}^{-1}\mathbf{f}$$

$$\text{and } \mathbf{\Sigma}_* = \mathbf{k}_*^T\mathbf{K}_M^{-1}\mathbf{\Sigma}_{\mathbf{t}|\mathbf{f}}\mathbf{K}_M^{-1}\mathbf{k}_* + \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T\mathbf{K}_M^{-1}\mathbf{k}_*$$

Note that the computation of posterior predictive is mainly dominated by the term $\Sigma_{\mathbf{t}|\mathbf{f}}$ whose computation cost is now $\mathcal{O}(M^2N)$ which is much less than the earlier version of GP(provided $M \ll N$) where the computation cost for inversion of $\mathbf{K}_N$ in posterior predictive was $\mathcal{O}(N^3)$.
2.

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

Note that we can also write $\mathbf{f} = \mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{t} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = \mathcal{N}(0, \mathbf{\Lambda})$.
Again using property of linear gaussian model we have,

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

$$\text{where } \boldsymbol{\mu} = \mathbf{K}_{NM}\mathbf{K}_M^{-1} \times 0 = 0$$

$$\text{and } \boldsymbol{\Sigma} = \mathbf{K}_{NM}\mathbf{K}_{M}^{-1}\mathbf{K}_{NM}^{T} + \boldsymbol{\Lambda}$$

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

$$\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} p(\mathbf{f}|\mathbf{X}, \mathbf{Z})$$

$$= \arg\max_{\mathbf{Z}}(-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{f}^{T}\boldsymbol{\Sigma}^{\mathbf{-1}}\mathbf{f})$$

Note that this objective 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: Subham Kumar
Roll Number: 160707
Date: March 14, 2019

QUESTION 2

Flavor 1.

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

$$p(\mathbf{z}_n|c_n = m, \Theta) = \mathcal{N}(\mathbf{z}_n|0, \mathbf{I}_K)$$

Here $\Theta = \left\{\pi_m, \boldsymbol{\mu}_m, \mathbf{W}_m, \sigma_m^2\right\}_{m=1}^M$
Given $c_n = m$, $\mathbf{x}_n$ can be written as $\mathbf{x}_n = \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \boldsymbol{\epsilon}_n$ where $\boldsymbol{\epsilon}_n = \mathcal{N}(0, \sigma_m^2\mathbf{I}_D)$.
Using property of linear gaussian model,we have

$$p(\mathbf{x}_n|c_n = m, \Theta) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{K}_m)$$

where $\mathbf{K}_m = \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D$.
The CLL can be written as:

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

$$= \sum_{n=1}^N\sum_{m=1}^M c_{nm}\left(\log p(\mathbf{x}_n|c_n = m, \Theta) + \log p(c_n = m|\Theta)\right)$$

$$= \sum_{n=1}^N\sum_{m=1}^M c_{nm}\left[\log\pi_m - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^T\mathbf{K}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m) - \frac{1}{2}\log|\mathbf{K}_m| + \text{const}\right]$$

Here $c_{nm} = \mathbb{I}[c_n = m]$.
Expected CLL:

$$E[\log p(\mathbf{X}, \mathbf{c}|\Theta)] = \sum_{n=1}^N\sum_{m=1}^M E[c_{nm}]\left[\log\pi_m - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^T\mathbf{K}_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m) - \frac{1}{2}\log|\mathbf{K}_m| + \text{const}\right]$$

The only required expectation i.e. $E[c_{nm}]$ can be calculated as:

$$E[c_{nm}] = 0 \times p(c_n \neq m|\mathbf{x}_n, \Theta) + 1 \times p(c_n = m|\mathbf{x}_n, \Theta)$$

$$\propto p(\mathbf{x}_n|c_n = m, \Theta)p(c_n = m|\Theta)$$

$$\text{Hence } E[c_{nm}] = \gamma_{nm} = \frac{\pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{K}_m)}{\sum_{l=1}^M \pi_l\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{K}_l)}$$

Note that the conditional posterior $p(c_n = m|\mathbf{x}_n, \Theta)$ is same as γ_{nm}
The M-step parameter updates are as follows:

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{n=1}^N \gamma_{nm}\mathbf{x}_n}{\sum_{n=1}^N \gamma_{nm}}$$

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

$$\hat{\sigma}_m^2 = \frac{1}{D-K}\sum_{k=K+1}^{D}\lambda_k$$

$$\hat{\mathbf{W}}_m = \mathbf{U}_K(\mathbf{L}_K - \hat{\sigma}_m^2\mathbf{I}_K)^{\frac{1}{2}}\mathbf{R}$$

Here $\mathbf{U}_k$ is $D\times K$ matrix of top K eigvectors of $\mathbf{S}_m = \frac{\sum_{n=1}^{N}\gamma_{nm}(\mathbf{x}_n-\hat{\boldsymbol{\mu}}_m)(\mathbf{x_n}-\hat{\boldsymbol{\mu}}_m)^T}{\sum_{n=1}^{N}\gamma_{nm}}$,$L_K$: $K\times K$ diagonal matrix of top K eigvalues $\lambda_1,...,\lambda_K$, $\mathbf{R}$ is a $K\times K$ arbitrary rotation matrix.

The EM algorithm:

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

2.E-step:$\forall\, n, m$:

$$E[c_{nm}^{(t)}] = \gamma_{nm}^{(t)} = \frac{\pi_m^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m^{(t-1)},\mathbf{K}_m^{(t-1)})}{\sum_{l=1}^{M}\pi_l^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l^{(t-1)},\mathbf{K}_l^{(t-1)})}$$

3.M-step:$\forall\, m$:

$$\boldsymbol{\mu}_m^{(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$$

$$\pi_m^{(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}{N}$$

$$\mathbf{S}_m^{(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}(\mathbf{x}_n-\boldsymbol{\mu}_m^{(t)})(\mathbf{x_n}-\boldsymbol{\mu}_m^{(t-1)})^T}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$$

$$\sigma_m^{2(t)} = \frac{1}{D-K}\sum_{k=K+1}^{D}\lambda_k^{(t)}$$

$$\mathbf{W}_m^{(t)} = \mathbf{U}_K^{(t)}(\mathbf{L}_K^{(t)} - \sigma_m^{2\,(t)}\mathbf{I}_K)^{\frac{1}{2}}\mathbf{R}^{(t)}$$

4.If not converged,then set t=t+1,Go to step 2;

The Stepwise EM algorithm:

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

2.While not converged:

Set learning rate γ_t,pick a random sample $\mathbf{x}_n$ and compute $\gamma_{nm}^{(t)}\ \forall\, m$.
Compute $\hat{\Theta}$ using only this sample
Update $\Theta^{(t)} = (1-\gamma_t)\Theta^{(t-1)} + \gamma_t\hat{\Theta}$
Set t=t+1;

Flavor 2.

$$p(\mathbf{z}_n|\mathbf{x}_n, c_n = m, \Theta) \propto p(\mathbf{x}_n|\mathbf{z}_n, c_n = m, \Theta)p(\mathbf{z}_n|c_n = m)$$

Writing $\mathbf{x}_n = \boldsymbol{\mu}_m + \mathbf{W}_m\mathbf{z}_n + \boldsymbol{\epsilon}_n$ where $\boldsymbol{\epsilon}_n = \mathcal{N}(0, \sigma_m^2\mathbf{I}_D)$ and using properties of Linear gaussian model to get gaussian posterior,we have:

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

$$\text{where }\ \boldsymbol{\Sigma}_{nm} = \left(\mathbf{I}_K + \frac{\mathbf{W}_m^T\mathbf{W}_m}{\sigma_m^2}\right)^{-1}$$

$$\text{and }\ \boldsymbol{\mu}_{nm} = \boldsymbol{\Sigma}_{nm}\left(\frac{\mathbf{W}_m^T(\mathbf{x}_n - \boldsymbol{\mu}_m)}{\sigma_m^2}\right)$$

Conditional posterior for $\mathbf{z}_n$ can be found by summing over possible values of c_n i.e.:

$$p(\mathbf{z}_n|\mathbf{x}_n,\Theta) = \sum_{m=1}^{M} p(\mathbf{z}_n|\mathbf{x}_n, c_n = m,\Theta)p(c_n = m|\Theta)$$

$$= \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{nm}, \boldsymbol{\Sigma}_{nm})$$

The conditional posterior for c_n i.e. $p(c_n = m|\mathbf{x}_n,\Theta)$ will remain same as in flavor 1. The CLL can be written as :

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

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

The Expected CLL is :

$$= -\sum_{n=1}^{N}\sum_{m=1}^{M} E[c_{nm}] \left[\frac{D}{2}\log\sigma_m^2 + \frac{1}{2\sigma_m^2}\left\|\mathbf{x}_n - \boldsymbol{\mu}_m\right\|^2 - \frac{E[\mathbf{z}_n]^T\mathbf{W}_m^T(\mathbf{x}_n - \boldsymbol{\mu}_m)}{\sigma_m^2} \right.$$
$$\left. + \frac{1}{2\sigma_m^2}\mathrm{Tr}\left(E[\mathbf{z_n}\mathbf{z_n}^T]\mathbf{W}_m^T\mathbf{W}_m\right) + \frac{1}{2}\mathrm{Tr}\left(E[\mathbf{z}_n\mathbf{z}_n^T]\right) - \log\pi_m + \text{const} \right]$$

Required expectations are $E[c_{nm}]$,$E[\mathbf{z}_n|c_n = m]$,$E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m]$.

$$E[c_{nm}] = \gamma_{nm} = \frac{\pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, \mathbf{K}_m)}{\sum_{l=1}^{M}\pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l, \mathbf{K}_l)}$$

$$E[\mathbf{z}_n|c_n = m] = \boldsymbol{\mu}_{nm}$$

$$E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m] = \boldsymbol{\Sigma}_{nm} + \boldsymbol{\mu}_{nm}\boldsymbol{\mu}_{nm}^T$$

The M-step parameter updates are as follows:

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{n=1}^{N}\gamma_{nm}(\mathbf{x}_n - \mathbf{W}_m E[\mathbf{z}_n|c_n = m])}{\sum_{n=1}^{N}\gamma_{nm}}$$

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

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

$$\hat{\sigma}_m^2 = \frac{\sum_{n=1}^{N}\gamma_{nm}\left\|\mathbf{x}_n - \boldsymbol{\mu}_m\right\|^2 + \mathrm{Tr}\left(E[\mathbf{z_n}\mathbf{z_n}^T]\mathbf{W}_m^T\mathbf{W}_m\right) - 2E[\mathbf{z}_n^T]\mathbf{W}_m^T(\mathbf{x}_n - \boldsymbol{\mu}_m)}{D\sum_{n=1}^{N}\gamma_{nm}}$$

The EM algorithm:

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

2.E-step:$\forall\ n, m$:

$$E[c_{nm}^{(t)}] = \gamma_{nm}^{(t)} = \frac{\pi_m^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m^{(t-1)}, \mathbf{K}_m^{(t-1)})}{\sum_{l=1}^{M}\pi_l^{(t-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_l^{(t-1)}, \mathbf{K}_l^{(t-1)})}$$

$$\boldsymbol{\Sigma}_{nm}^{(t)} = \left(\mathbf{I}_K + \frac{\mathbf{W}_m^{T(t-1)}\mathbf{W}_m^{(t-1)}}{\sigma_m^{2(t-1)}}\right)^{-1}$$

$$E[\mathbf{z}_n|c_n = m]^{(t)} = \boldsymbol{\mu}_{nm}^{(t)} = \boldsymbol{\Sigma}_{nm}^{(t)}\left(\frac{\mathbf{W}_m^{T(t-1)}(\mathbf{x}_n - \boldsymbol{\mu}_m^{(t-1)})}{\sigma_m^{2(t-1)}}\right)$$

$$E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m]^{(t)} = \boldsymbol{\Sigma}_{nm}^{(t)} + \boldsymbol{\mu}_{nm}^{(t)}\boldsymbol{\mu}_{nm}^{T(t)}$$

3.M-step:$\forall\ m$:

$$\boldsymbol{\mu}_m^{(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}(\mathbf{x}_n - \mathbf{W}_m^{(t-1)}E[\mathbf{z}_n|c_n = m]^{(t)})}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$$

$$\pi_m^{(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}{N}$$

$$\mathbf{W}_m^{(t)} = \left(\sum_{n=1}^{N}\gamma_{nm}^{(t)}(\mathbf{x}_n - \boldsymbol{\mu}_m^{(t)})E[\mathbf{z}_n|c_n = m]^{T(t)}\right)\left(\sum_{n=1}^{N}\gamma_{nm}^{(t)}E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m]^{(t)}\right)^{-1}$$

$$\sigma_m^{2(t)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}\left(\left\|\mathbf{x}_n - \hat{\boldsymbol{\mu}}_m^{(t)}\right\|^2 + \mathrm{Tr}\left(E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m]^{(t)}\mathbf{W}_m^{T(t)}\mathbf{W}_m^{(t)}\right) - 2E[\mathbf{z}_n|c_n = m]^{T(t)}\mathbf{W}_m^{T(t)}(\mathbf{x}_n - \boldsymbol{\mu}_m^{(t)})\right)}{D\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$$

4.If not converged,then set t=t+1,Go to step 2;

The Stepwise EM algorithm:

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

2.While not converged:

Set learning rate γ_t,pick a random sample $\mathbf{x}_n$

Compute $\gamma_{nm}^{(t)}, \boldsymbol{\Sigma}_{nm}^{(t)}, E[\mathbf{z}_n|c_n = m]^{(t)}, E[\mathbf{z}_n\mathbf{z}_n^T|c_n = m]^{(t)}\ \ \forall\ m.$

Compute $\hat{\Theta}$ using only this sample

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

Set t=t+1;

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

QUESTION 3

Student Name: Subham Kumar
Roll Number: 160707
Date: March 14, 2019

The joint distribution can be written as:

$$p(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\alpha}|\mathbf{X},\theta)=\left[\prod_{n=1}^{N}\mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n,\beta^{-1})\mathcal{N}(\mathbf{w}|0,\mathrm{diag}(\alpha_1^{-1},..,\alpha_D^{-1}))Gamma(\beta|a_0,b_0)\prod_{d=1}^{D}Gamma(\alpha_d|e_0,f_0)\right]$$

Here $\boldsymbol{\alpha}=[\alpha_1,..,\alpha_D]^T$ and $\theta=\{a_0,b_0,e_0,f_0\}$ The log-joint can be now written as:

$$\log p(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\alpha}|\mathbf{X},\theta)=\Bigg(\left(a_0-1+\frac{N}{2}\right)\log\beta-\beta\left[b_0+\frac{\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2}{2}\right]$$
$$+\left(e_0+\frac{1}{2}-1\right)\sum_{d=1}^{D}\log\alpha_d-\sum_{d=1}^{D}\alpha_d\left(\frac{w_d^2}{2}\right)-f_0\sum_{d=1}^{D}\alpha_d+\text{constant}\Bigg)$$

Now consider the mean field V.I. approximation of posterior as :

$$p(\mathbf{w},\beta,\boldsymbol{\alpha}|\mathbf{y},\mathbf{X})=q(\mathbf{w}|\lambda_1)q(\beta|\lambda_2)\prod_{d=1}^{D}q(\alpha_d|\phi_d)$$

VI approximation for β can be found as follows:

$$\log q^*(\beta)=E_{q(\boldsymbol{\alpha})q(\mathbf{w})}[\log p(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\alpha}|\mathbf{X},\theta)]$$

$$=\left(a_0-1+\frac{N}{2}\right)\log\beta-\beta\left[b_0+\frac{E_{q(\mathbf{w})}\left[\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\right]}{2}\right]+\text{const}$$

$$=\left(a_0-1+\frac{N}{2}\right)\log\beta-\beta\left[b_0+\frac{E_{q(\mathbf{w})}\left[(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right]}{2}\right]+\text{const}$$

$$=\left(a_0-1+\frac{N}{2}\right)\log\beta-\beta\left[b_0+\frac{\mathbf{y}^T\mathbf{y}-2\mathbf{y}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\right]+\mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\mathbf{w}^T\right]\right)}{2}\right]+\text{const}$$

The above form is similar to log of a Gamma Distribution whose shape and rate parameters are as follows:

$$q^*(\beta)=Gamma(\beta|\tau_1,\tau_2)$$
$$\text{where }\ \tau_1=a_0+\frac{N}{2}$$
$$\text{and }\ \tau_2=b_0+\frac{\mathbf{y}^T\mathbf{y}-2\mathbf{y}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\right]+\mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\mathbf{w}^T\right]\right)}{2}$$

Similarily,solving for $\boldsymbol{\alpha}$ component-wise

$$\log q^*(\alpha_d)=\left(e_0+\frac{1}{2}-1\right)\log\alpha_d-\alpha_d\left(f_0+\frac{E_{q(\mathbf{w})}[w_d^2]}{2}\right)+\text{constant}\ \ \forall\, d$$

This is again similar to log of a Gamma Distribution whose shape and rate parameters are as follows:

$$q^*(\alpha_d) = Gamma(\alpha_d|e_d, f_d)$$

$$\text{where } \ e_d = e_0 + \frac{1}{2}$$

$$\text{and } \ f_d = \left(f_0 + \frac{E_{q(\mathbf{w})}[w_d^2]}{2}\right)$$

Now solving for $\mathbf{w}$:

$$\log q^*(\mathbf{w}) = E_{q(\beta)}\left[-\beta\left[b_0 + \frac{(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})}{2}\right] - \frac{1}{2}\mathbf{w}^T\text{diag}(\alpha_1, ..., \alpha_D)\mathbf{w}\right] + \text{const}$$

$$= \frac{-1}{2}\left(\mathbf{w}^T\left[E_{q(\beta)}[\beta]\mathbf{X}^T\mathbf{X} + \text{diag}\big(E_{q(\alpha_1)}[\alpha_1], ..., E_{q(\alpha_D)}[\alpha_D]\big)\right]\mathbf{w} - 2E_{q(\beta)}[\beta]\mathbf{w}^T\mathbf{X}^T\mathbf{y}\right) + \text{const}$$

Comparing it with log of gaussian distribution in its information form,we get:

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

$$\text{where } \ \boldsymbol{\Sigma}_N = \left(E_{q(\beta)}[\beta]\mathbf{X}^T\mathbf{X} + \text{diag}\big(E_{q(\alpha_1)}[\alpha_1], ..., E_{q(\alpha_D)}[\alpha_D]\big)\right)^{-1}$$

$$\text{and } \ \boldsymbol{\mu}_N = E_{q(\beta)}[\beta]\boldsymbol{\Sigma}_N\mathbf{X}^T\mathbf{y}$$

Note that all the expectations can be easily found using corresponding distributions.Hence,

$$E_{q(\beta)}[\beta] = \frac{\tau_1}{\tau_2}$$

$$E_{q(\alpha_d)}[\alpha_d] = \frac{e_d}{f_d} \ \ \forall\, d$$

$$E_{q(\mathbf{w})}\left[\mathbf{w}\right] = \boldsymbol{\mu}_N$$

$$E_{q(\mathbf{w})}\left[\mathbf{w}\mathbf{w}^T\right] = \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T$$

$$E_{q(\mathbf{w})}[w_d^2] = [\boldsymbol{\Sigma}_N]_{dd} + \boldsymbol{\mu}_{Nd}^2 \ \ \forall\, d$$

Note that updates of these distributions depend on each-other thus require cyclic updates:

The Mean Field VI Algorithm

1.Given $\theta, \mathbf{X}, \mathbf{y}$,compute the following as:

$$\tau_1 = a_0 + \frac{1}{2}$$

$$e_d = e_0 + \frac{1}{2}$$

2.Set t=0,Initialize $\boldsymbol{\Sigma}_N^{(0)} = \mathbf{I}_D$ and $\boldsymbol{\mu}_N^{(0)} = 0$ so that

$$f_d^{(0)} = f_0 + \frac{1}{2}$$

$$\tau_2^{(0)} = b_0 + \frac{\mathbf{y}^T\mathbf{y} + \text{Tr}\left(\mathbf{X}^T\mathbf{X}\right)}{2}$$

$$E_{q(\beta)}[\beta]^{(0)} = \frac{\tau_1}{\tau_2^{(0)}}$$

$$E_{q(\alpha_d)}[\alpha_d]^{(0)} = \frac{e_d}{f_d^{(0)}} \;\; \forall\, d$$

3.Set t=t+1;

$$\boldsymbol{\Sigma}_N^{(t)} = \Big(E_{q(\beta)}[\beta]^{(t-1)}\mathbf{X}^T\mathbf{X} + \mathrm{diag}\Big(E_{q(\alpha_1)}[\alpha_1]^{(t-1)}, ..., E_{q(\alpha_D)}[\alpha_D]^{(t-1)}\Big)\Big)^{-1}$$

$$\boldsymbol{\mu}_N^{(t)} = E_{q(\beta)}[\beta]^{(t-1)}\boldsymbol{\Sigma}_N^{(t)}\mathbf{X}^T\mathbf{y}$$

$$\tau_2^{(t)} = b_0 + \frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\right]^{(t)} + \mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}E_{q(\mathbf{w})}\left[\mathbf{w}\mathbf{w}^T\right]^{(t)}\right)}{2}$$

$$f_d^{(t)} = \left(f_0 + \frac{E_{q(\mathbf{w})}[w_d^2]^{(t)}}{2}\right) \;\; \forall\, d$$

$$E_{q(\beta)}[\beta]^{(t)} = \frac{\tau_1}{\tau_2^{(t)}}$$

$$E_{q(\alpha_d)}[\alpha_d]^{(t)} = \frac{e_d}{f_d^{(t)}} \;\; \forall\, d$$

4.If not converged,Go to step 3.

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

QUESTION 4

Student Name: Subham Kumar
Roll Number: 160707
Date: March 14, 2019

1.Score Function Gradient Method:

$$\nabla_\phi \mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\phi \log q(\mathbf{w_s}|\phi)\left(\log p(\mathbf{y}, \mathbf{w}_s|\mathbf{X}) - \log q(\mathbf{w}_s|\phi)\right)$$

Given that $q(\mathbf{w}|\phi) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$,we can rewrite $\mathbf{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$ where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{v} = \mathcal{N}(\mathbf{v}|0, \mathbf{I}_D)$

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

Using chain rule:

$$\nabla_{\boldsymbol{\mu}} \log q(\mathbf{w}|\phi) = (\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{w}-\boldsymbol{\mu})$$
$$\nabla_{\mathbf{L}} \log q(\mathbf{w}|\phi) = (\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L} - \mathbf{L}^{-T}$$

Also,

$$\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}) = \sum_{n=1}^{N} \log p(y_n|\mathbf{x}_n, \mathbf{w}) + \log p(\mathbf{w})$$

$$= \sum_{n=1}^{N}\left(y_n\mathbf{w}^T\mathbf{x}_n - \log\left(1+\exp(y_n\mathbf{w}^T\mathbf{x}_n)\right)\right) - \frac{\lambda\mathbf{w}^T\mathbf{w}}{2} + \frac{D}{2}\log\frac{\lambda}{2\pi}$$

Substitute $\mathbf{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$ in above expression to get $\log p(\mathbf{y}, \boldsymbol{\mu} + \mathbf{L}\mathbf{v}|\mathbf{X})$
Hence

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S} \Big((\mathbf{L}\mathbf{L}^T)^{-1}&(\mathbf{w}_s-\boldsymbol{\mu})(\mathbf{w}_s-\boldsymbol{\mu})^T(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L} \\ &- \mathbf{L}^{-T}\Big)\left(\sum_{n=1}^{N}\left(y_n(\mathbf{w}_s)^T\mathbf{x}_n - \log\left(1+\exp(y_n(\mathbf{w}_s)^T\mathbf{x}_n)\right)\right) - \frac{\lambda(\mathbf{w}_s)^T(\mathbf{w}_s)}{2} + \frac{D}{2}\log\lambda\right. \\ &\left. + \frac{1}{2}(\mathbf{w}_s-\boldsymbol{\mu})^T(\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{w}_s-\boldsymbol{\mu}) + \frac{1}{2}\log\left|\mathbf{L}\mathbf{L}^T\right|\right)\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left((\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{w}_s-\boldsymbol{\mu})\right)&\left(\sum_{n=1}^{N}\left(y_n(\mathbf{w}_s)^T\mathbf{x}_n - \log\left(1+\exp(y_n(\mathbf{w}_s)^T\mathbf{x}_n)\right)\right)\right. \\ &\left. - \frac{\lambda(\mathbf{w}_s)^T(\mathbf{w}_s)}{2} + \frac{D}{2}\log\lambda + \frac{1}{2}(\mathbf{w}_s-\boldsymbol{\mu})^T(\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{w}_s-\boldsymbol{\mu}) + \frac{1}{2}\log\left|\mathbf{L}\mathbf{L}^T\right|\right)\end{aligned}$$

The VI algorithm using B=1:

1.Initialize $\phi^{(0)} = \{\boldsymbol{\mu}, \mathbf{L}\}^{(0)}$,choose learning rate η,set t=1;
2.Draw S samples $\{\mathbf{w}_1, ..., \mathbf{w}_S\}^{(t)}$ from $q(\mathbf{w}|\phi^{(t-1)})$.

3.Pick a random example $(\mathbf{x}_n, y_n)$;
4.Using only this chosen example,update

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}(q)^{(t)} \approx \frac{1}{S}\sum_{s=1}^{S}(\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T})^{-1}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t-1)})\left(\left(y_n(\mathbf{w}_s^{(t)})^T\mathbf{x}_n - \log\left(1+\exp(y_n(\mathbf{w}_s^{(t)})^T\mathbf{x}_n)\right)\right) - \frac{\lambda(\mathbf{w}_s^{(t)})^T(\mathbf{w}_s^{(t)})}{2} + \frac{D}{2}\log\lambda + \frac{1}{2}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t-1)})^T(\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T})^{-1}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t-1)}) + \frac{1}{2}\log\left|\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T}\right|\right)$$

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

$$\nabla_{\mathbf{L}}\mathcal{L}(q)^{(t)} \approx \frac{1}{S}\sum_{s=1}^{S}\left((\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T})^{-1}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t)})(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t)})^T(\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T})^{-1}\mathbf{L}^{(t-1)} - \mathbf{L}^{(t-1)-T}\right)\left(\left(y_n(\mathbf{w}_s^{(t)})^T\mathbf{x}_n - \log\left(1+\exp(y_n(\mathbf{w}_s^{(t)})^T\mathbf{x}_n)\right)\right) - \frac{\lambda(\mathbf{w}_s^{(t)})^T(\mathbf{w}_s)^{(t)}}{2} + \frac{D}{2}\log\lambda + \frac{1}{2}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t)})^T(\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T})^{-1}(\mathbf{w}_s^{(t)} - \boldsymbol{\mu}^{(t)}) + \frac{1}{2}\log\left|\mathbf{L}^{(t-1)}\mathbf{L}^{(t-1)T}\right|\right)$$

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

5.If ELBO not converged,then set t=t+1,Go to step 2

2.Pathwise Gradient Method:

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

Note that the entropy form is integrabale given that $q(\mathbf{w}|\phi)$ is Gaussian and the value is:

$$-E_{q(\mathbf{w}|\phi)}\left[\log q(\mathbf{w}|\phi)\right] = \frac{1}{2}\log\left(2\pi|\mathbf{\Sigma}|e^D\right)$$

Reparmetrizing $\mathbf{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$ and finding the required gradients,we have:

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

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

Also Gradients of the entropy are:

$$\frac{1}{2}\nabla_{\boldsymbol{\mu}}\log\left(2\pi|\mathbf{L}\mathbf{L}^T|e^D\right) = 0$$

$$\frac{1}{2}\nabla_{\mathbf{L}}\log\left(2\pi|\mathbf{L}\mathbf{L}^T|e^D\right) = \mathbf{L}^{-T}$$

Summing these up the gradient of ELBO w.r.t. $\phi^{'}$ comes out to be:

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

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

The VI algorithm using B=1:
1.Generate S samples from $\mathcal{N}(\mathbf{v}|0,\mathbf{I}_D)$.
2.Initialize $\phi^{'} = \{\boldsymbol{\mu},\mathbf{L}\}$ as $\phi^{'(0)}$,choose learning rate η,set t=1;
3.pick a random example $(\mathbf{x}_n, y_n)$;
4.Using only this chosen example,update

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

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

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

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

5.If ELBO not converged,then set t=t+1,Go to step 3;