Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Areeb The Probabilistic Modeller
Roll Number: 180135
Date: October 23, 2022

QUESTION 1

$$
\begin{aligned}
p(\boldsymbol{f}|\boldsymbol{y}) &\propto \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, f)p(\boldsymbol{f}) \\
&\propto \prod_{n=1}^{N} \mathcal{N}(y_n|f(\boldsymbol{x}_n), \sigma^2)p(\boldsymbol{f}) \\
&\propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I})p(\boldsymbol{f}) \\
&\propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K})
\end{aligned}
$$

Hence, by using standard gaussian results, posterior will be

$$
\begin{aligned}
p(\boldsymbol{f}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\
\boldsymbol{\mu}_* &= (\boldsymbol{I} + \sigma^2\boldsymbol{K}^{-1})^{-1}\boldsymbol{y} \\
\boldsymbol{\Sigma}_* &= (\frac{\boldsymbol{I}}{\sigma^2} + \boldsymbol{K}^{-1})^{-1}
\end{aligned} \tag{2}
$$

With increasing l ,posterior means tries to match true $\sin(x)$ function, but if we keep increasing it , it will deviate.Also, the mean becomes smoother with increasing l, this is because of increasing covariance between terms.

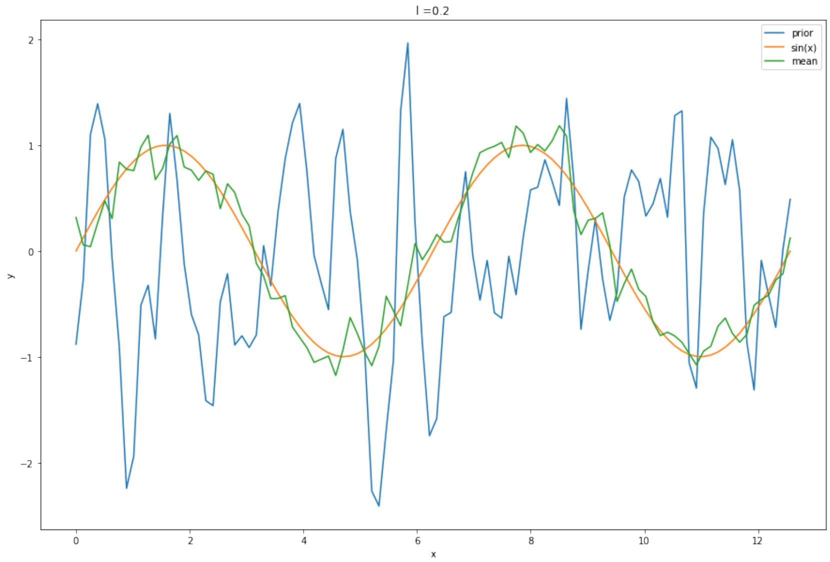

Figure 1: l=0.2

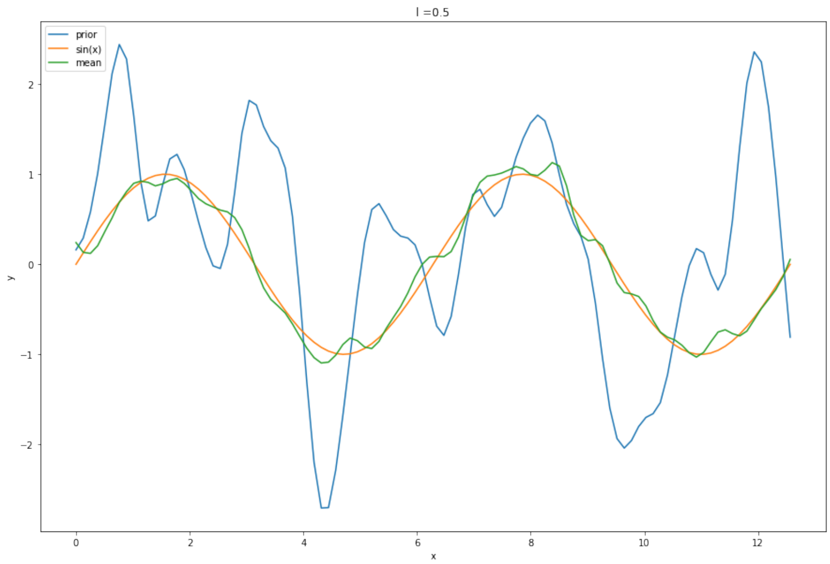

Figure 2: l=0.5

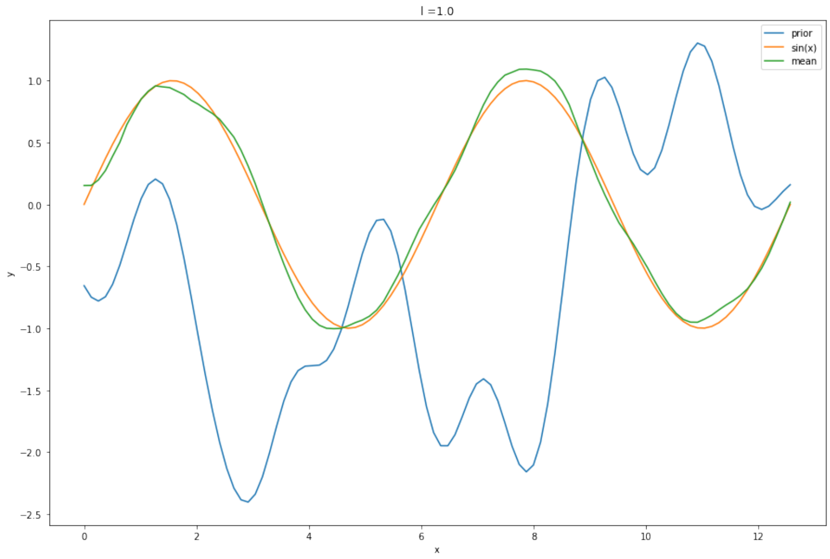

Figure 3: l=1

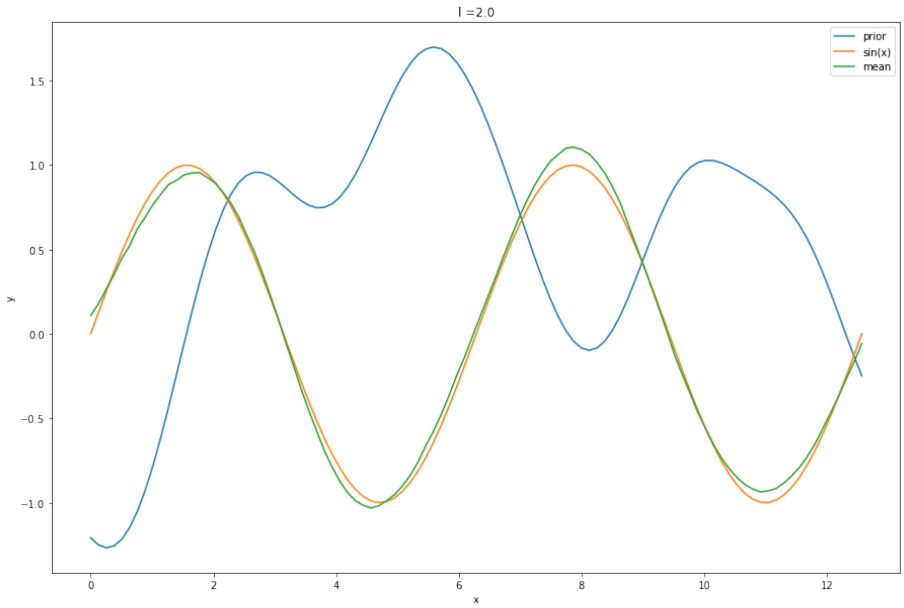

Figure 4: l=2

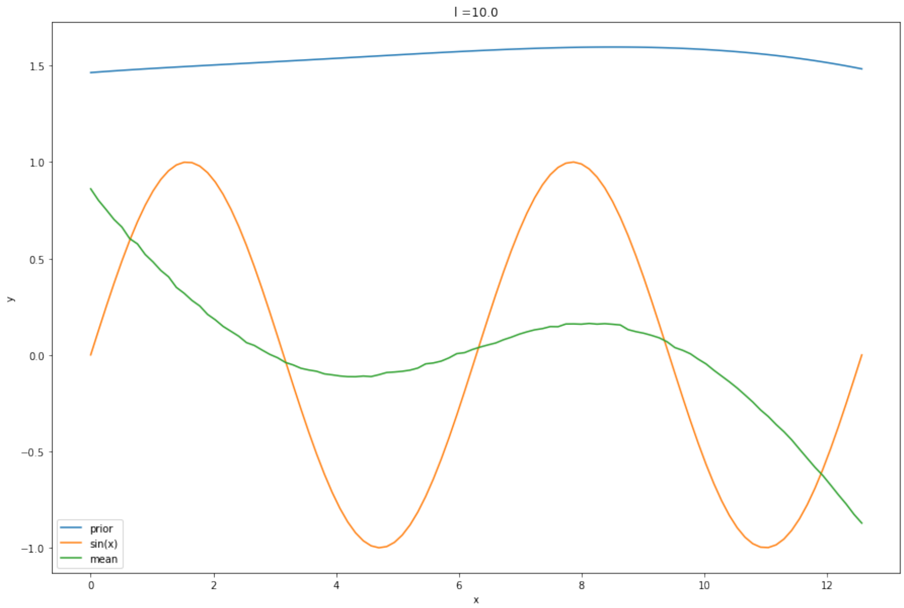

Figure 5: l=10

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Areeb The Probabilistic Modeller
Roll Number: 180135
Date: October 23, 2022

QUESTION 2

My solution to problem 2.
Posterior predictive for new input $\boldsymbol{x_*}$ is :

$$p(\boldsymbol{f_*}|\boldsymbol{x_*}, \boldsymbol{X}, \boldsymbol{f}) = \mathcal{N}(\boldsymbol{f_*}|\boldsymbol{k_*}^T \boldsymbol{K}^{-1}\boldsymbol{f}, \kappa(\boldsymbol{x_*}, \boldsymbol{x_*}) - k_*^T \boldsymbol{K^{-1}} \boldsymbol{k_*})$$

Now we have pseudo-training inputs $\boldsymbol{Z} = z_1, z_2, ..., z_M$ along with their pseudo noiseless outputs $\boldsymbol{t} = t_1, t_2, ..., t_M$ where $M << N$.For this we have:

$$p(f_n|x_n, \boldsymbol{Z}, \boldsymbol{t}) = \mathcal{N}(f_n|\tilde{\boldsymbol{k}}^T \tilde{\boldsymbol{K}}^{-1}\boldsymbol{t}, \kappa(x_n, x_n) - \tilde{\boldsymbol{k}}^T \tilde{\boldsymbol{K}}^{-1}\tilde{\boldsymbol{k}}) \tag{3}$$

When $\tilde{\boldsymbol{K}}$ is $M \times M$ kernel matrix of pseudo inputs $\boldsymbol{Z}$ and $\tilde{\boldsymbol{k}_n}$ if the $M \times 1$ vector of kernel based similarities of x_n with each of the pseudo inputs $\boldsymbol{z}_1, ..., \boldsymbol{z}_M$

$$\begin{aligned} p(f_n|x_n, \boldsymbol{Z}, \boldsymbol{t}) &= \mathbf{N}(f_n|\boldsymbol{k}_n^T \boldsymbol{K}_M^{-1}\boldsymbol{t}, \kappa(x_n, x_n) - \boldsymbol{k}_n^T \boldsymbol{K}_n^{-1}\boldsymbol{k}_n) \\ p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{t}) &= \prod_{n=1}^{N} p(f_n|x_n, \boldsymbol{Z}, \boldsymbol{t}) \\ p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{t}) &= \mathcal{N}(\boldsymbol{f}|\boldsymbol{P}\boldsymbol{K}_M^{-1}\boldsymbol{t}, \boldsymbol{\Lambda}) \end{aligned} \tag{4}$$

Where $\boldsymbol{P}_{nm} = \kappa(x_n, z_m)$ and $\boldsymbol{K}_M$ is $M \times M$ matrix with $(K_M)_{nm} = \kappa(z_n, z_m)$ and $\boldsymbol{\Lambda_{ii}} = \boldsymbol{\kappa(x_i, x_i) - k_n^T K_M^{-1} k_n}$.

- Posterior Predictive Distribution of output y_* of new input x_*

$$\begin{aligned} p(y_*|x_*, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{f}) &= \int p((y_*|x_*, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{f}, \boldsymbol{t})p(\boldsymbol{t}|(y_*|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{f})d\boldsymbol{t} \\ p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{f}) &\propto p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{t})p(\boldsymbol{t}|\boldsymbol{Z}) \end{aligned} \tag{5}$$

We have $p(\boldsymbol{t}|\boldsymbol{Z}) = \mathcal{N}(\boldsymbol{t}|0, \boldsymbol{K}_M)$.Using the results of gaussian process we get

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{f}) &= \mathcal{N}(\boldsymbol{t}|\boldsymbol{\mu_{t|f}}, \boldsymbol{\Sigma_{t|f}}) \\ \boldsymbol{\mu_{t|f}} &= \boldsymbol{\Sigma_{t|f} K_M^{-1} P^T \Lambda^{-1} f} \\ \boldsymbol{\Sigma_{t|f}} &= (\boldsymbol{K_M}^{-1}\boldsymbol{P^T \Lambda^{-1} P K_M^{-1}})^{-1} \end{aligned} \tag{6}$$

We have $\boldsymbol{f_*} = \boldsymbol{k_*^T K_M^{-1} t} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \kappa(x_*, x_*))$.Using results of the gaussian process we have

$$\begin{aligned} p(y_*|\boldsymbol{x_*, X, f, Z}) &= \mathcal{N}(y_*|\boldsymbol{\mu_*}.\boldsymbol{\Sigma_*}) \\ \boldsymbol{\mu_*} &= \boldsymbol{k_*^T K_M^{-1} \Sigma_{t|f} K_M^{-1} P^T \Lambda^{-1} f} \\ \boldsymbol{\Sigma_*} &= \boldsymbol{k_*^T K_M^{-1} \Sigma_{t|f} K_M^{-1} k_*} + \kappa(x_*, x_*) - \boldsymbol{k_*^T K_M^{-1} k_*} \end{aligned} \tag{7}$$

How does this posterior predictive for y_* compare with the usual GP's posterior predictive for y_* in terms of computational cost?
The computation cost is now $O(M^2N)$, which is due to covariance matrix $\boldsymbol{\Sigma_{t|f}}$, since $M << N$ is much better than the previous $O(N^3)$.

- The Marginal Likelihood is:

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

We can directly use the properties of gaussian models:

$$\begin{aligned} p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{Z}) &= \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ \boldsymbol{\mu} &= 0 \\ \boldsymbol{\Sigma} &= \boldsymbol{P}\boldsymbol{K_M^{-1}}\boldsymbol{P^T} + \boldsymbol{\Lambda} \end{aligned} \tag{8}$$

MLE-II objective is

$$\begin{aligned} \hat{\boldsymbol{Z}} &= \underset{\boldsymbol{Z}}{argmax}\; p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{Z}) \\ &= \underset{\boldsymbol{Z}}{argmax}(-\log|\boldsymbol{\Sigma}| - \boldsymbol{f^T}\boldsymbol{\Sigma^{-1}}\boldsymbol{f}) \end{aligned} \tag{9}$$

This can be solved using a gradient accent.

In the case of arguments model

$$p(y_n, z_n|\boldsymbol{w}, \boldsymbol{x_n}, \sigma^2, \nu) = \mathcal{N}(y_n|\boldsymbol{w^T x_n}, \frac{\sigma^2}{z_n}) * Gamma(z_n|\frac{v}{2}, \frac{v}{2})$$

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \rho^2, \sigma^2) &\propto \prod_{n=1}^{N} p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X}, \boldsymbol{z}, \sigma^2) p(\boldsymbol{w}|\rho^2) \\ &\propto \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w^T x_n}, \frac{\sigma^2}{z_n}) \mathcal{N}(\boldsymbol{w}|0, \rho^2 \boldsymbol{I_D}) \\ &\propto \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, diag\left[\frac{\sigma^2}{z_1}, ..., \frac{\sigma^2}{z_N}\right]\right) \mathcal{N}(\boldsymbol{w}|0, \rho^2 \boldsymbol{I_D}) \end{aligned} \tag{10}$$

Using slides , gaussian posterior can be written as $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \rho^2, \sigma^2) = \mathcal{N}(\boldsymbol{\mu_*}, \boldsymbol{\Sigma_*})$

$$\begin{aligned} \boldsymbol{\mu_*} &= \boldsymbol{\Sigma_* X^T \Sigma^{-1} y} \\ \boldsymbol{\Sigma_*} &= \left(\boldsymbol{X^T \Sigma^{-1} X} + \frac{\boldsymbol{I_D}}{\rho^2}\right)^{-1} \end{aligned} \tag{11}$$

where $\boldsymbol{\Sigma^{-1}} = diag\left[\frac{z_1}{\sigma^2}, ..., \frac{z_N}{\sigma^2}\right]$
conditional posterior of z_n

$$\begin{aligned} p(z_n|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{z}_{-n}.\nu, \sigma^2) &\propto p(y_n|z_n, \boldsymbol{w}, x_n, \sigma^2) p(z_n|\nu) \\ &\propto \mathcal{N}\left(y_n|\boldsymbol{w^T x_n}, \frac{\sigma^2}{z_n}\right) * Gamma\left(z_n|\frac{\nu}{2}, \frac{\nu}{2}\right) \\ &\propto z_n^{\frac{\nu+1}{2}-1} \boldsymbol{exp}\left[-z_n\left(\frac{(y_n - w^T x_n)^2}{2\sigma^2} + \frac{\nu}{2}\right)\right] \end{aligned} \tag{12}$$

Hence , $p(z_n|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{z}_{-n}.\nu, \sigma^2) = Gamma\left(\frac{\nu+1}{2}, \frac{(y_n - w^T x_n)^2}{2\sigma^2} + \frac{\nu}{2}\right)$

The Gibbs Sampling Algorithm:

1. Initialize $w = w^{(0)}$
2. for $t = 0, 1, ..., T$

$$(i)\ z_n^{(t)} \sim Gamma\left(\frac{\nu+1}{2}, \frac{(y_n - w^T x_n)^2}{2\sigma^2} + \frac{\nu}{2}\right) \tag{13}$$

$$(ii)\ \boldsymbol{w^{(t)}} \sim \mathcal{N}(\boldsymbol{\mu_*}, \boldsymbol{\Sigma_*}) \tag{14}$$

Repeat till convergence or threshold.

EM algorithm

- **E step**
  for all $z_1, z_2, ..., z_n$

$$z_n \sim p(z_n|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{z}_{-n}.\nu, \sigma^2) \tag{15}$$

  Expectation will be

$$\mathbb{E}\left[p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \rho^2, \sigma^2)\right]$$

  As mentioned in slides

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \mathbb{E}\left[\boldsymbol{z}\right], \rho^2, \sigma^2)$$

- **Maximization step**

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{argmax}\;\; p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \mathbb{E}\left[\boldsymbol{z}\right], \rho^2, \sigma^2) \tag{16}$$

  It is basically MAP estimate.By using first order optimality condition i.e $\frac{\partial p}{\partial \boldsymbol{w}} = \boldsymbol{0}$,we will get

$$\hat{\boldsymbol{w}} = \left[\boldsymbol{X^T X} + \frac{\sigma^2}{\rho^2}\boldsymbol{I}_D\right]^{-1} \boldsymbol{X^T}\mathbb{E}\left[Diag[\boldsymbol{z}]\right]\boldsymbol{y} \tag{17}$$

Probabilistic Machine Learning (CS772A), Fall 2022
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Areeb The Probabilistic Modeller
Roll Number: 180135
Date: October 23, 2022

QUESTION 4

Given,

$$\begin{aligned} p(\gamma_d) &= Bernoulli(\theta) \\ p(\theta) &= Beta(a_o, b_o) \\ p(\sigma^2) &= IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}) \\ p(w_d|\sigma, \gamma_d) &= \mathcal{N}(0, \sigma^2\kappa_{\gamma_d}) \\ where \quad \kappa_{\gamma_d} &= \gamma_d v_1 + (1-\gamma_d)v_0 \end{aligned} \tag{18}$$

- The given weight prior is dividing the features into two types based on their importance. It also does sparse learning for weight parameters of two types: the precision is high for one type while lower for another.We can see this as an automatic feature division.
- Posterior over the latent variables:

$$\begin{aligned} \mathcal{P}(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}) &\propto \mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2)\mathcal{P}(\mathbf{w}|\sigma^2, \boldsymbol{\gamma}) \\ &\propto \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\boldsymbol{I_N})\mathcal{N}(\mathbf{w}|0, \sigma^2\boldsymbol{K}) \end{aligned} \tag{19}$$

$$K = diag(\kappa_{\gamma_1}, \kappa_{\gamma_2}, ..., \kappa_{\gamma_D})$$

Now, from using results from slides

$$\mathcal{P}(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu_w}, \boldsymbol{\Sigma_w}) \tag{20}$$

$$\Sigma_w = \sigma^2(\boldsymbol{X^T X} + \boldsymbol{K}^{-1})^{-1} \tag{21}$$

$$\boldsymbol{\mu_w} = \frac{1}{\sigma^2}\boldsymbol{\Sigma_w X^T y} \tag{22}$$

The complete data log-likelihood(CLL) will be

$$\begin{aligned} log\mathcal{P}(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \boldsymbol{\gamma}) &= log\mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) + log\mathcal{P}(\mathbf{w}|\sigma^2, \boldsymbol{\gamma}) \\ &= -\frac{N+D}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Xw})^T(\boldsymbol{y} - \boldsymbol{Xw}) \\ &- \frac{1}{2\sigma^2}\boldsymbol{w}^T\boldsymbol{K^{-1}}\boldsymbol{w} - \frac{1}{2}\sum_{d=1}^{D} log(\kappa_{\gamma_d}) \end{aligned} \tag{23}$$

$$\begin{aligned} \mathbb{E}\left[CLL\right] = -\frac{1}{2\sigma^2}(\boldsymbol{y^T y} - 2\boldsymbol{y^T X}\mathbb{E}\left[\mathbf{w}\right]) + Tr((\boldsymbol{X^T X} + \boldsymbol{K^{-1}}\mathbb{E}\left[\boldsymbol{w^T w}\right])) \\ - \frac{N+D}{2}log(2\pi\sigma^2) - \frac{1}{2}\sum_{d=1}^{D} log(\kappa_{\gamma_d}) \end{aligned} \tag{24}$$

from slides

$$\mathbb{E}\left[\boldsymbol{w}\right] = \boldsymbol{\mu_w} \tag{25}$$

$$\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w^T}\right] = \boldsymbol{\Sigma_w} + \boldsymbol{\mu_w}\boldsymbol{\mu_w^T} \tag{26}$$

Maximization step:

$$\log[\mathcal{P}(\sigma^2)] = -\left(\frac{\nu}{2}+1\right)\log(\sigma^2) - \frac{\nu\gamma}{2\sigma^2} + constant \tag{27}$$

$$\log[\mathcal{P}(\theta)] = (a_0-1)\log(\theta) + (b_0-1)\log(1-\theta) \tag{28}$$

$$\log[\mathcal{P}(\gamma_d|\theta)] = \gamma_d\log(\theta) + (1-\gamma_d)\log(1-\theta) \tag{29}$$

The MAP estimate can be written as follows:

$$\begin{aligned}\{\sigma^2,\gamma,\theta\}_{MAP} &= \underset{\sigma^2,\theta,\gamma}{arg\ max}\ \mathbb{E}\left[CLL\right] + \log\mathcal{P}(\sigma^2,\theta,\gamma) \\ &= \underset{\sigma^2,\theta,\gamma}{arg\ max}\ \mathbb{E}\left[CLL\right] + \log\mathcal{P}(\sigma^2) + \log\mathcal{P}(\theta) + \sum_{d=1}^{D}\log\mathcal{P}(\gamma_d|\theta)\end{aligned} \tag{30}$$

Update of $\gamma_d|\theta$ **:**

$$\begin{aligned}\gamma_d &= \underset{\gamma_d\in\{0,1\}}{arg\ max}\ \mathbb{E}\left[CLL\right] + \log\mathcal{P}(\sigma^2,\theta,\gamma) \\ &= \underset{\gamma_d\in\{0,1\}}{arg\ max}\ -\frac{1}{2\sigma^2\kappa_{\gamma_d}}\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w^T}\right] - \frac{1}{2}\log(\kappa_{\gamma_d}) + \gamma_d\log(\theta) + (1-\gamma_d)\log(1-\theta)\end{aligned} \tag{31}$$

Update of σ^2 **:**

$$\frac{\partial(\mathbb{E}\left[CLL\right] + \log\mathcal{P}(\sigma^2,\theta,\gamma))}{\partial(\sigma^2)} = 0 \tag{32}$$

$$\frac{1}{2\sigma^4}(\boldsymbol{y^T}\boldsymbol{y} - 2\boldsymbol{y^T}\boldsymbol{X}\mathbb{E}\left[\boldsymbol{w}\right] + Tr((\boldsymbol{X^T}\boldsymbol{X} + \boldsymbol{K^{-1}}\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w^T}\right]) - \frac{N+D}{2\sigma^2} - \frac{1}{\sigma^2}\left(\frac{\nu}{2}+1\right) + \frac{\nu\gamma}{2\sigma^4} = 0 \tag{33}$$

$$\sigma^2 = \frac{\boldsymbol{y^T}\boldsymbol{y} - 2\boldsymbol{y^T}\boldsymbol{X}\mathbb{E}\left[\boldsymbol{w}\right] + Tr((\boldsymbol{X^T}\boldsymbol{X} + \boldsymbol{K^{-1}}\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w^T}\right])) + \nu\gamma}{N+D+\nu+2} \tag{34}$$

Update of θ **:**

$$\frac{\partial(\mathbb{E}\left[CLL\right] + \log\mathcal{P}(\sigma^2,\theta,\gamma))}{\partial(\theta)} = 0 \tag{35}$$

$$\frac{1}{\theta}\left(\sum_{d=1}^{D}\gamma_d + a_0 - 1\right) - \frac{1}{1-\theta}\left(\sum_{d=1}^{D}(1-\gamma_d) + b_0 - 1\right) = 0 \tag{36}$$

$$\theta = \frac{\sum_{d=1}^{D}\gamma_d + a_0 - 1}{D + a_0 + b_0 - 2} \tag{37}$$

EM algorithm:

1.$(\sigma^2,\gamma,\theta) = (\sigma^2,\gamma,\theta)^0$

2.for $t = 0, 1, .., T$

- E step:
  updating the posterior:

$$\mathcal{P}(\boldsymbol{w}^{t+1}|\boldsymbol{y},\boldsymbol{X},\sigma^{2(t)},\gamma^{(t)}) = \mathcal{N}(\boldsymbol{w}^{(t)}|\boldsymbol{\mu_w}^{(t+1)},\boldsymbol{\Sigma_w}^{(t+1)}) \tag{38}$$

$$\boldsymbol{\Sigma_w^{(t+1)}} = \sigma^{2(t)}\left[\boldsymbol{X^T X} + \boldsymbol{(K^{-1})}^{(t)}\right]^{-1} \tag{39}$$

$$\boldsymbol{\mu_w}^{(t+1)} = \frac{1}{\sigma^{2(t)}}\left[\boldsymbol{\Sigma_w}^{(t+1)}\boldsymbol{X^T y}\right] \tag{40}$$

$$\mathbb{E}\left[\boldsymbol{w}\right]^{(t+1)} = \boldsymbol{\mu_w}^{(t+1)} \tag{41}$$

$$\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w^T}\right] = \boldsymbol{\Sigma_w}^{(t+1)} + \boldsymbol{\mu_w}^{(t+1)}\boldsymbol{(\mu_w^T)}^{(t+1)} \tag{42}$$

- M step: update the parameters:

  1. $\gamma_d|\theta$ using eq.14.

  2.σ^2 using eq.17

  3.θ using eq.20.

  Return $(\sigma^2,\gamma,\theta)^T$ and $\mathcal{P}(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X},\sigma^{2(T-1)},\gamma^{(T-1)})$ until convergence.