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

Student Name: Shashank Shailabh
Roll Number: 170655
Date: April 21, 2021

QUESTION 1

1. The equation 4 defines the expected complete data log posterior as

$$\mathbb{E}_{\mathcal{Y}_p}[\log p(\boldsymbol{\theta}|\mathcal{D}_0 \cup (\mathcal{X}_p, \mathcal{Y}_p))]$$

The batch $\mathcal{D}'$ is find such that it covers whole data variation which can be $\mathcal{D}_0 \cup \mathcal{D}_p$. The unavailability of complete data is removed using posterior expectation as.

$$\begin{aligned}\mathbb{E}_{\mathcal{Y}_p}[\log p(\boldsymbol{\theta}|\mathcal{D}_0 \cup \mathcal{D}')] &\approx \mathbb{E}_{\mathcal{Y}_p}[\log p(\boldsymbol{\theta}|\mathcal{D}_0 \cup \mathcal{D}_p)] \\ &\approx \log p(\boldsymbol{\theta}|\mathcal{D}_0) + \sum_{m=1}^{M} \mathcal{L}_m\end{aligned}$$

The sparse approximation is selecting B points from M points and with objective function to minimize $||\sum_{m=1}^{M} \mathcal{L}_m - \sum_{m=1}^{M} w_m \mathcal{L}_m||^2$ subject to $w_m \in \{0, 1\}$ and $\sum_{m=1}^{M} w_m \leq B$. The idea above is true as sparse approximation where B points approximates the updated expected data log posterior to expected complete data log posterior.

2. The relaxed objective is

$$\min_{w} (1 - \mathbf{w})^T \mathbf{K} (1 - \mathbf{w})$$

such that w_m is non-negative and $\sum_{m=1}^{M} w_m ||\mathcal{L}_m|| = \sum_{m=1}^{M} ||\mathcal{L}_m||$.
The original objective function constraint is NP-hard so the relaxed objective uses trick of constrain optimization. The original constrain is changed to non-negative constrain where greedily B points are chosen in one by one fashion followed by projecting to the original constrain space. The selection is performed such that the point chosen takes closest to the complete posterior.

3. The acquisition function proposed in the paper gave closed form solution for linear supervised model. The non-linear supervised learning models can use random feature projection using MC approximation since acquisition function won't be available. The Equation(16) in the paper describes the approximate projection.

$$\hat{\mathcal{L}_n} = \frac{1}{\sqrt{J}} [\mathcal{L}_n(\boldsymbol{\theta}_1), ..., \mathcal{L}_n(\boldsymbol{\theta}_J)]^T$$

where $\boldsymbol{\theta}_j \sim \hat{\pi}$ and $\hat{\mathcal{L}_n}$ is J-dimensional projection of $\mathcal{L}_n$.

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

QUESTION 2

Student Name: Shashank Shailabh
Roll Number: 170655
Date: April 21, 2021

Given the input $x_1, x_2, .., x_N \sim \mathcal{N}(x, |\mu, \beta^{-1})$ and prior on $\mu \sim \mathcal{N}(\mu|\mu_0, s_0)$ and $\beta \sim Gamma(\beta|a, b)$.

Conditional Posterior
Using Bayes rule for conditional posterior calculation.

$$p(\mu|x, \beta^{-1}) = \frac{p(x|\mu, \beta^{-1})p(\mu|\mu_0, s_0)}{\int p(x|\mu, \beta^{-1})p(\mu)d\mu}$$

Both likelihood and prior are gaussian, using squares trick and assume mean $= \mu_N$, variance $= \sigma_N^2$ and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ where

$$\mu_N = \frac{\mu\beta^{-1}}{Ns_0 + \beta^{-1}} + \frac{Ns_0\bar{x}}{Ns_0 + \beta^{-1}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{s_0} + N\beta$$

Now again using Bayes rule for β posterior calculation.

$$p(\beta|x, \mu) = \frac{p(x|\mu, \beta^{-1})p(\beta|a, b)}{\int p(x|\mu, \beta^{-1})p(\beta)d\beta}$$

Now the conditional on β is defined by $Gamma(\beta|a_N, b_N)$ where

$$a_N = a + \frac{N}{2}$$

and

$$b_N = b + \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2}$$

Gibbs Sampling
The random samples are iteratively drawn from the conditional posteriors. When the sampling is long run enough (convergence achieved), then the samples are from the joint posterior. These are the steps for Gibbs sampler -

- Initialize $\mu^{(0)}$
- For $s = 1, 2, ..., S$
  - Draw a random sample for β as $\beta^{(s)} \sim p(\beta|x, \mu^{(s-1)})$
  - Draw a random sample for μ as $\mu^{(s)} \sim p(\mu|x, \beta^{(s)})$

These pairs of samples $(\mu^s, \beta^s)_{s=1}^{S}$ represent joint posterior of μ and β.

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

QUESTION 3

Student Name: Shashank Shailabh
Roll Number: 170655
Date: April 21, 2021

---

1. The above prior on $\mathbf{w}$ penalises differently for each weight and forces the weights towards zero. Intuitively, the prior is classifying the features in two groups.

2. The conditional posterior on $\mathbf{w}$ is calculated as $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$ and $p(\mathbf{w}|\sigma^2, \gamma) = \mathcal{N}(\mathbf{w}|\sigma^2\mathbf{K})$ where $\mathbf{K}$ is $D \times 1$ diagonal matrix withe $\mathbf{K} = diag(\kappa_{\gamma 1}, .., \kappa_{\gamma D})$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \gamma) \propto p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\sigma^2, \gamma)$$

Now, $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \gamma) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ where $\boldsymbol{\Sigma}_N = \sigma^2(\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}$ and $\boldsymbol{\mu}_N = (\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}\mathbf{X}^T\mathbf{y}$.

The EM algorithm -
Expectation Step:

$$\begin{aligned}
\log p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma) &= \log(p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2)) + \log(p(\mathbf{w}|\sigma^2, \gamma)) \\
&= \frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{(D+N)\log(2\pi\sigma^2)}{2} \\
&\quad - \frac{\mathbf{w}^T\mathbf{K}\mathbf{w}}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma d}) \\
&= \frac{-1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X} + \mathbf{w}^T\mathbf{K}^{-1}\mathbf{w}) - \frac{(D+N)\log(2\pi\sigma^2)}{2} \\
&\quad - \frac{\mathbf{w}^T\mathbf{K}\mathbf{w}}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma d})
\end{aligned}$$

Taking Expectation on both sides of CLL,

$$\begin{aligned}
\mathbb{E}[\log(p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma))] &= \frac{-1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}] + Tr((\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})\mathbb{E}[\mathbf{w}\mathbf{w}^T])) - \frac{(D+N)\log(2\pi\sigma^2)}{2} \\
&\quad - \frac{\mathbf{w}^T\mathbf{K}\mathbf{w}}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma d})
\end{aligned}$$

Using posterior calculation, $\mathbb{E}[\mathbf{w}] = \boldsymbol{\mu}_N = (\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^T] = \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T = \sigma^2(\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1} + (\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}$.
Therefore,

$$\begin{aligned}
\mathbb{E}[\log(p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma))] &= \frac{-1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{K}^{-1})^{-1}\mathbf{X}^T\mathbf{y}) - \frac{(D+N)\log(2\pi\sigma^2)}{2} \\
&\quad - \frac{\mathbf{w}^T\mathbf{K}\mathbf{w}}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma d})
\end{aligned}$$

Maximization step:
The expectation of $p(\sigma^2, \gamma, \theta|\mathbf{wy}, \mathbf{X})$ is maximised for estimating the σ^2, γ, θ.

$$\mathbb{E}[\log\{p(\sigma^2, \gamma, \theta|\mathbf{wy}, \mathbf{X})\}] = \mathbb{E}[\log(p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma))] + \log(p(\sigma^2, \gamma, \theta)) + constant$$

$$\begin{aligned}\{\sigma^2, \gamma, \theta\}_{MAP} &= \arg\max_{\{\sigma^2,\gamma,\theta\}} \mathbb{E}[\log(p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma))] + \log(p(\sigma^2, \gamma, \theta)) \\ &= \arg\max_{\{\sigma^2,\gamma,\theta\}} \mathbb{E}[\log(p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma^2, \gamma))] + \log(p(\sigma^2)) + \log(p(\gamma|\theta)) + \log(p(\theta))\end{aligned}$$

where

$$\log(p(\sigma^2)) = -(\frac{\nu}{2} + 1)\log\sigma^2 - \frac{\nu\lambda}{2\sigma^2} + constant$$

$$\log(p(\gamma_d|\theta)) = \gamma_d \log\theta + (1 - \gamma_d)\log(1 - \theta)$$

$$\log(p(\theta)) = (a_0 - 1)\log\theta + (b_0 - 1)\log(1 - \theta)$$

- Update step for σ^2 on differentiating and solving yield

$$\sigma_t^2 = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{K}_{t-1}^{-1})^{-1}\mathbf{X}^T\mathbf{y} + \nu\lambda}{N + D + \nu + 2}$$

- Update step for γ_d on solving yield

$$\gamma_d = \arg\max_{\gamma_d \in \{0,1\}} \frac{-1}{2\sigma_t^2\kappa_{\gamma d}}\mathbb{E}_{t-1}[\mathbf{w}\mathbf{w}^T]_{d,d} - \frac{\log(\kappa_{\gamma d})}{2} + \gamma_d \log\theta_{t-1} + (1 - \gamma_d)\log(1 - \theta_{t-1})$$

- Update step for θ on differentiating and solving yield

$$\theta_t = \frac{a_0 - 1 + \sum_{d=1}^{D}(\gamma_d)_t}{a_0 + b_0 + D - 2}$$

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

QUESTION

4

Student Name: Shashank Shailabh
Roll Number: 170655
Date: April 21, 2021

Part 1:

Given that $p(\mathbf{f}) = \mathcal{N}(0, \mathbf{K})$ and $p(y_n|\mathbf{x}_n, f) = \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2)$ where $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N)]^T$ is $N \times 1$ vector and $\mathbf{K}$ is $N \times N$ kernel matrix with $K_{nm} = \kappa(\mathbf{x}_n, \mathbf{x}_m)$
Using bayes rule to calculate posterior,

$$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f})$$

Applying linear Gaussian model result where $A = \mathbf{I}$,$b = 0$,$L^{-1} = \sigma^2\mathbf{I}$ which is $N \times N$ diagonal matrix,$\mu = 0$ and $\Lambda^{-1} = \mathbf{K}$

$$\begin{aligned} p(\mathbf{f}|\mathbf{y}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|0, \mathbf{K}) \\ &= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\ &= \mathcal{N}\left(\mathbf{f}|\left(\mathbf{K}^{-1} + \frac{\mathbf{I}}{\sigma^2}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}, \left(\mathbf{K}^{-1} + \frac{\mathbf{I}}{\sigma^2}\right)^{-1}\right) \end{aligned}$$

Hence, posterior $\boldsymbol{\mu}_N = \left(\mathbf{K}^{-1} + \frac{\mathbf{I}}{\sigma^2}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}$ and $\boldsymbol{\Sigma}_N = \left(\mathbf{K}^{-1} + \frac{\mathbf{I}}{\sigma^2}\right)^{-1}$

Part 2:

The required Python notebook shows all the code. Prior samples become smooth (useless) on increasing the value of l from 0.2 to 10 i.e, it do not change according to the change in the samples with increase in l. The mean of the GP posterior fluctuates for 0.2 and becomes smooth (useless) with increase in l. The mean of the GP posterior matches the true sine function for $l = 2$. For $l = 10$, the posterior mean differs a lot than true sine function.

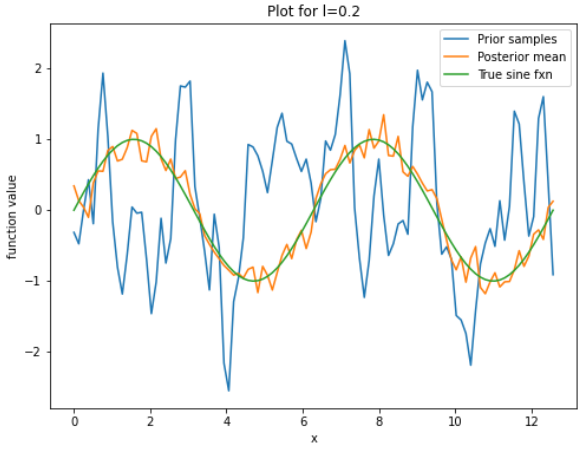

Figure 1: Plot for $l = 0.2$.

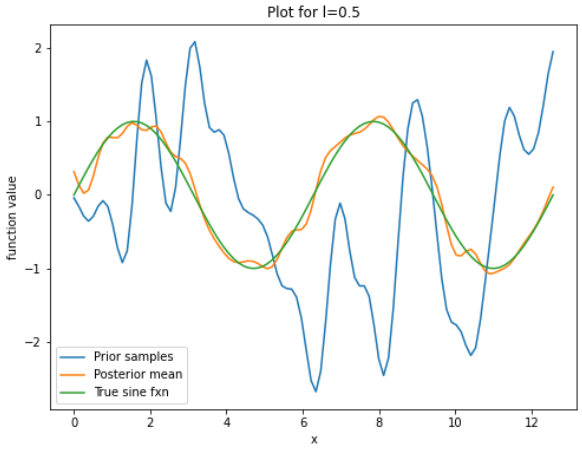

Figure 2: Plot for $l = 0.5$.

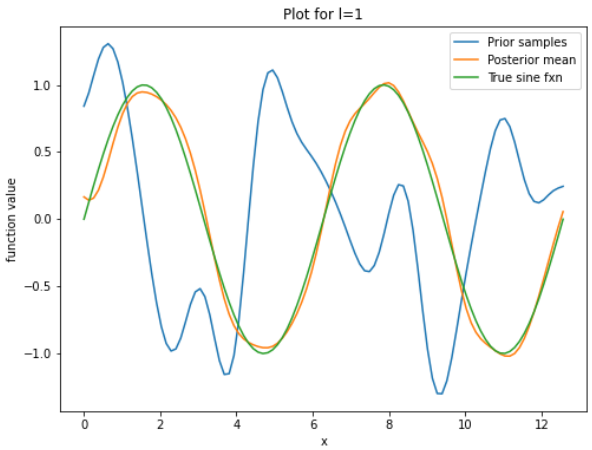

Figure 3: Plot for $l = 1.0$.

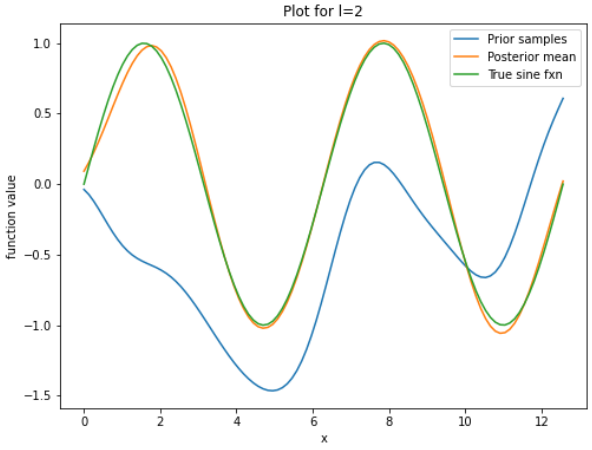

Figure 4: Plot for $l = 2.0$.

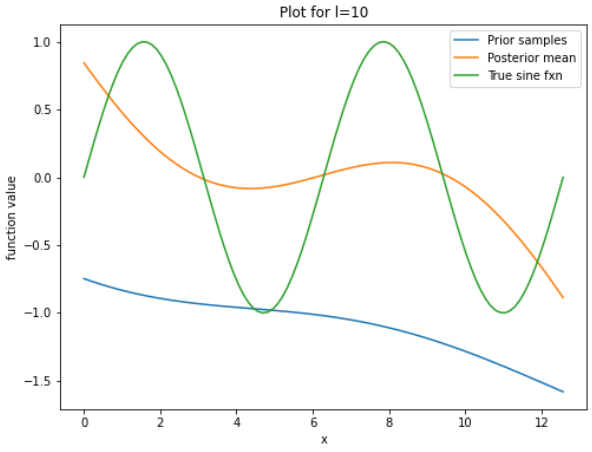

Figure 5: Plot for $l = 10$.

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

QUESTION 5

Student Name: Shashank Shailabh
Roll Number: 170655
Date: April 21, 2021

Given that the likelihood of each training output f_n has same form as GP regression's posterior predictive with $\mathbf{Z}, \mathbf{t}$ as pseudo training data.

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

where $\tilde{\mathbf{K}}$ is $M \times M$ kernel matrix, $\tilde{\mathbf{k}}_n$ is the $M \times 1$ vector of kernel based similarities of $\mathbf{x}_n$ with pseudo inputs $\mathbf{z}_1, \mathbf{z}_2, .., \mathbf{z}_M$.

Posterior Predictive distribution

Posterior predictive distribution for the output y_* of a new input $\mathbf{x}_*$

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

and

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

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

$$p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) = \prod_{n=1}^{N} p(f_n|\mathbf{x}_n, \mathbf{Z}, \mathbf{t}) = \prod_{n=1}^{N} \mathcal{N}(f_n|\tilde{\mathbf{k}}_n\tilde{\mathbf{K}}^{-1}\mathbf{t}, \kappa(\mathbf{x}_n, \mathbf{x}_n) - \tilde{\mathbf{k}}_n^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_n) = \mathcal{N}(\mathbf{f}|\mathbf{R}\grave{\mathbf{K}}^{-1}\mathbf{t}, \mathbf{\Lambda})$$

where $\mathbf{R}_{nm} = \kappa(\mathbf{x}_n, \mathbf{z}_m)$, shape of $\mathbf{R}$ is $N \times M$, $\grave{\mathbf{K}}$ is $M \times M$ matrix where each element $\grave{\mathbf{K}}_{nm} = \kappa(\mathbf{z}_n, \mathbf{z}_m)$, $\mathbf{\Lambda}$ is $N \times N$ diagonal matrix with diagonal element $\mathbf{\Lambda}_{nn} = \kappa(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{k}_n^T\grave{\mathbf{K}}^{-1}\mathbf{k}_n$. $p(\mathbf{t}|\mathbf{Z})$ is gaussian with mean= 0 and co-variance matrix = $\grave{\mathbf{K}}$. Using the conjugate property of gaussian for $p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})$.

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

where $\boldsymbol{\mu}_t = \mathbf{\Sigma}_t\grave{\mathbf{K}}^{-1}\mathbf{R}^T\mathbf{\Lambda}^{-1}\mathbf{f}$ and $\mathbf{\Sigma}_t = \grave{\mathbf{K}}\mathbf{R}^{-1}\mathbf{\Lambda}\mathbf{R}^{T^{-1}}\grave{\mathbf{K}}$.

Replacing $y_* = f_*$, $f(\mathbf{x}_*) = \mathbf{k}_*^T\grave{\mathbf{K}}^{-1}\mathbf{k}_*$ and $\epsilon \sim \mathcal{N}(0, \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T\grave{\mathbf{K}}^{-1}\mathbf{k}_*)$ in GP regression formula where $\mathbf{k}_*$ is $M \times 1$ vector with element $(\mathbf{k}_*)_m = \kappa(\mathbf{x}_*, \mathbf{z}_m)$. Therefore,

$$\begin{aligned} p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) &= \mathcal{N}(f_*|\mu_*, \sigma_*) \\ &= \mathcal{N}(f_*|\mathbf{k}_*^T\grave{\mathbf{K}}^{-1}\mathbf{\Sigma}_t\grave{\mathbf{K}}^{-1}\mathbf{R}^T\mathbf{\Lambda}^{-1}\mathbf{f}, \mathbf{k}_*^T\grave{\mathbf{K}}^{-1}\mathbf{\Sigma}_t\grave{\mathbf{K}}^{-1}\mathbf{k}_* + \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T\grave{\mathbf{K}}^{-1}\mathbf{k}_*) \end{aligned}$$

Now, this is evident that new computation cost is $O(M^2N)$ due to the $\mathbf{\Sigma}_t$ computation in predictive posterior.

Marginal Likelihood

Given $\mathbf{f} = \mathbf{R}\grave{\mathbf{K}}^{-1}\mathbf{t} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ and linear Gaussian model.

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

where $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{R}\grave{\mathbf{K}}^{-1}\mathbf{R}^T + \boldsymbol{\Lambda}$
Using MLE-II for $\mathbf{Z}$ computation and objective function,

$$\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} P(\mathbf{f}|\mathbf{X}, \mathbf{Z}) = \arg\max_{\mathbf{Z}} \left[\frac{-1}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{f}^T\boldsymbol{\Sigma}^{-1}\mathbf{f} \right] = \arg\max_{\mathbf{Z}} \left[\log|\boldsymbol{\Sigma}| + \mathbf{f}^T\boldsymbol{\Sigma}^{-1}\mathbf{f} \right]$$