Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 1

The paper "Bayesian Batch Active Learning as Sparse Subset Approximation" by Pinsler et al proposes a novel Bayesian Batch Active Learning approach that tries to deal with the issues faced by pre-existing methods. Most probabilistic active learning algorithms try to query new informative data points from the oracle that minimize the expected posterior entropy or posterior uncertainty. Such querying is done using greedy iterative methods like BALD and often results in spending resources on querying correlated or similar data points and retraining the model. Hence, is it not very useful in large scale settings.

- The batch AL selects the data $\mathcal{D}'$ which updates the log posterior such that it is a better approximation of the complete data log posterior. To do this, we use the expectation of parameters $\boldsymbol{\theta}$ given $\mathcal{D}_0$ and M (complete data pool size) data points $\mathcal{X}, \mathcal{Y}$ with respect the the current posterior predictive distribution of $\mathcal{Y}$ i.e. $p(\mathcal{Y}|\mathcal{X}, \mathcal{D}_0) = \int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}_0)$.

$$\mathop{\mathbb{E}}_{\mathcal{Y}}[\log p(\boldsymbol{\theta}|\mathcal{D}_0 \cup (\mathcal{X}, \mathcal{Y}))] = \log p(\boldsymbol{\theta}|\mathcal{D}_0) + \sum_{m=1}^{M} \underbrace{\mathop{\mathbb{E}}_{\mathcal{Y}}[\log p(y_m|x_m|\boldsymbol{\theta})] + \mathbb{H}[y_m|x_m, \mathcal{D}_0]}_{\mathcal{L}_m(\boldsymbol{\theta})}$$

  We have to choose the batch that best approximates $\mathcal{L} = \sum_m \mathcal{L}_m(\boldsymbol{\theta})$ (considering $\mathcal{L}_m$ as a vectors in function space) since the first term does not depend on the new data. The paper suggests an approach to do it as a sparse approximation. We assign a weight vector $\mathbf{w}$ to each $\mathcal{L}_m(\boldsymbol{\theta})$ for $m = 1$ to M and denote $\mathcal{L}(\mathbf{w}) = \sum_m w_m \mathcal{L}_m$. Now, our modified task is to find the optimal $\mathbf{w}$ as,

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{minimise}} \ ||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2 \ , w_m \in \{0, 1\} \ \forall \ m \ , \ \sum_m \mathbb{1}_m \leq b$$

  Intuitively this makes sense because the objective function mentioned above approximates $\mathcal{L}$ which means that the new posterior we obtain after querying the b data points with $w_m = 1$, would be closer to the expected value of the posterior if we had observed the entire data set.

- The sparse approximation objective is difficult to optimize and intractable. To solve the problem, a relaxed objective function is used and the problem is solved in a Hilbert space induced by the inner product $\langle \mathcal{L}_m, \mathcal{L}_m \rangle$. $\mathcal{L}_m$ is being considered as a vector in the function space. For the relaxed optimization problem, we remove the condition that all $w_m \in \{0, 1\}$ and replace with it a non-negative constraint. The cardinality constraint is also modified into a polytope constraint such that now the problem is,

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{minimise}} \ (\mathbf{1} - \mathbf{w})^T \mathbf{K} (\mathbf{1} - \mathbf{w}) \ , w_m \geq 0 \ \forall \ m \ , \ \sum_m w_m \sigma_m = \sigma$$

  where $\mathbf{K}$ is a $M \times M$ kernel matrix of the inner products with $K_{mn} = \langle \mathcal{L}_m, \mathcal{L}_m \rangle$, $\sigma_m = ||\mathcal{L}_m||$ and $\sigma = \sum_m w_m \sigma_m$. Here, the identity $||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2 = (\mathbf{1} - \mathbf{w})^T \mathbf{K} (\mathbf{1} - \mathbf{w})$ has been used to rewrite the original objective function. This optimization is solved approximately using the Frank-Wolfe algorithm mentioned in the paper.

- The acquisition function mentioned in the paper has a closed form expression for linear regression and probit regression. A closed form expression won't be available for models where the Fisher inner product doesn't have a closed form as well. The paper suggests using random feature projections to approximate the required quantities for models which have a tractable likelihood. The following projections are used in the algorithm,

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

  where $\hat{\mathcal{L}_n}$ is the J dimensional projection. This can be viewed as drawing J Monte Carlo samples from the posterior $\hat{\pi}$ and hence we get an unbiased estimator for the inner product $\langle \mathcal{L}_m, \mathcal{L}_m \rangle \approx \hat{\mathcal{L}_n^T} \hat{\mathcal{L}}_m$. The paper uses the weighted Euclidean inner product instead of the weighted Fisher product since it avoids the calculation of gradients of $\mathcal{L}_n$ which may be very difficult for complex models.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 2

By the idea of local conjugacy,

$$p(\mu|\mathbf{X},\beta) \propto p(\mathbf{X}|\mu,\beta)p(\mu)$$

$$\implies p(\mu|\mathbf{X},\beta) \propto \mathcal{N}(\mathbf{X}|\mu,\beta^{-1})\mathcal{N}(\mu|\mu_0,s_0)$$

$$\implies p(\mu|\mathbf{X},\beta) = \mathcal{N}(\mu|\mu_*,s_*) \quad \textbf{where,}$$

$$\mu_* = \frac{\mu_0 s_0^{-1} + \beta\sum_{i=1}^n x_i}{s_0^{-1} + n\beta} \quad \textbf{and} \quad s_* = (s_0^{-1} + n\beta)^{-1}$$

Similarly,

$$p(\beta|\mathbf{X},\mu) \propto p(\mathbf{X}|\mu,\beta)p(\beta)$$

$$\implies p(\beta|\mathbf{X},\mu) \propto \mathcal{N}(\mathbf{X}|\mu,\beta^{-1})Gamma(\beta|a,b)$$

$$\implies p(\beta|\mathbf{X},\mu) = Gamma(a + \frac{n}{2}, b + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2})$$

Steps for a Gibbs sampling algorithm to approximate the joint posterior $p(\mu,\beta|\mathbf{X})$:

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

When run long enough (for high values of S), the random samples $(\mu,\beta)_{s=1}^S$ represent the joint posterior $p(\mu,\beta|\mathbf{X})$.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 3

1. Since $v_0 > v_1$, this prior effectively attaches a lower variance to weights of some of the parameters, pushing them closer to 0. Hence it is able to create a sparse model where some parameters are more "relevant" than others.

2. We have $\Theta = (\gamma, \sigma^2, \theta)$ and $q(\boldsymbol{w}) = \prod_{d=1}^{D} p(w_d|\sigma, \gamma_d) = \mathcal{N}(\boldsymbol{w}|0, \sigma^2 \boldsymbol{K})$ where we have $\boldsymbol{K} = \boldsymbol{\gamma}^T v_1 + (\mathbf{I}_D - \boldsymbol{\gamma})^T v_0$. We can write the **posterior** on $\boldsymbol{w}$ as following,

$$p(\boldsymbol{w}|\Theta, \mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{w}, \mathbf{X}) p(\boldsymbol{w}|\Theta)$$

$$\implies p(\boldsymbol{w}|\Theta, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_w, \Sigma_w) \quad \text{[Standard Bayesian Linear Regression result]}$$

$$\text{where } \mu_w = (\mathbf{X}^T\mathbf{X} + (\boldsymbol{K})^{-1}\mathbf{I}_D)^{-1}\mathbf{X}^T\mathbf{y} \text{ and } \Sigma_w = ((\sigma^2)^{-1}\mathbf{X}^T\mathbf{X} + (\sigma^2\boldsymbol{K})^{-1}\mathbf{I}_D)^{-1}$$

Next, we calculate the Complete Data log-likelihood or CLL as follows,

$$\log p(\boldsymbol{w}, \mathbf{y}|\mathbf{X}, \Theta) = \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{w}, \Theta) + \log p(\boldsymbol{w}|\Theta)$$

$$= -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{w})^T(\mathbf{y} - \mathbf{X}\boldsymbol{w})}{2\sigma^2} - \frac{N+D}{2}\log(2\pi\sigma^2) - \frac{\boldsymbol{w}^T\boldsymbol{K}^{-1}\boldsymbol{w}}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma_d})$$

Expected value of CLL over the posterior of $\boldsymbol{w}$ is calculated as,

$$\mathbb{E}_{\boldsymbol{w}|\mathbf{y}}[\log p(\boldsymbol{w}, \mathbf{y}|\mathbf{X}, \Theta)] = -\frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbb{E}[\boldsymbol{w}] + \mathrm{Tr}(\mathbf{X}^T\mathbf{X}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}])}{2\sigma^2}$$

$$-\frac{N+D}{2}\log(2\pi\sigma^2) - \frac{\mathrm{Tr}(\boldsymbol{K}^{-1}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}])}{2\sigma^2} - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma_d})$$

where we can replace $\mathbb{E}[\boldsymbol{w}]$ with μ_w and $\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$ with $\Sigma_w + \mu_w\mu_w^T$.

We can get MAP estimates of Θ by maximizing their posterior. We will obtain the following expression,

$$\hat{\Theta} = \arg\max_{\Theta} \mathbb{E}_{\boldsymbol{w}|\mathbf{y}}[\log p(\boldsymbol{w}, \mathbf{y}|\mathbf{X}, \Theta)] + \log p(\Theta)$$

$$\log p(\Theta) = \log p(\sigma^2) + \log p(\theta) + \sum_{d=1}^{D}\log p(\gamma_d|\theta) \text{ where,}$$

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

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

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

- Update of σ^2:

$$\frac{\partial(\mathbb{E}_{\boldsymbol{w}|\mathbf{y}}[\log p(\boldsymbol{w},\mathbf{y}|\mathbf{X},\Theta)] + \log p(\Theta))}{\partial \sigma^2} = 0$$

$$\implies \sigma^2 = \frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbb{E}[\boldsymbol{w}] + \mathrm{Tr}((\mathbf{X}^T\mathbf{X} + \boldsymbol{K}^{-1})\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]) + \nu\lambda}{N + D + \nu + 2}$$

- Update of θ:

$$\frac{\partial(\mathbb{E}_{\boldsymbol{w}|\mathbf{y}}[\log p(\boldsymbol{w},\mathbf{y}|\mathbf{X},\Theta)] + \log p(\Theta))}{\partial \theta} = 0$$

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

- Update of $\boldsymbol{\gamma}$:
  Since $\gamma_d \in \{0,1\}$, we can directly calculate,

$$\gamma_d = \arg\max_{\gamma_d\in\{0,1\}} \mathbb{E}_{\boldsymbol{w}|\mathbf{y}}[\log p(\boldsymbol{w},\mathbf{y}|\mathbf{X},\Theta)] + \log p(\Theta)$$

  We perform this update for all $d \in [1, D]$.

EM Algorithm for Sparse Modeling:

1. Initialize Θ as Θ^0, set $t \leftarrow 1$
2. **E Step:** Compute posterior on $\boldsymbol{w}$ given current parameters

$$p(\boldsymbol{w}^{(t)}|\Theta^{(t-1)},\mathbf{X},\mathbf{y}) = \mathcal{N}(\mu_w,\Sigma_w)\ ,\ \text{recomputing values of } \mu_w,\Sigma_w \text{ mentioned above}$$

3. **M Step:** Update every parameter in $\Theta^{(t)} = (\boldsymbol{\gamma}^{(t)}, (\sigma^2)^{(t)}, \theta^{(t)})$ via the equations detailed above using the new posterior.
4. If not yet converged, set $t \leftarrow t + 1$ and go to step 2.
5. **Return** $\Theta^{(t)}$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 4

The expression for the GP posterior can be obtained as follows:

$$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) \;\; \textbf{where} \;\; f \sim \mathcal{GP}(0, \kappa)$$

$$\implies p(\mathbf{f}|\mathbf{y}) \propto \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2)\mathcal{N}(0, \mathbf{K}) = \mathcal{N}(0, \mathbf{K}) \prod_{i=1}^{N} \mathcal{N}(y_i|f(x_i), \sigma^2)$$

$$\implies log(p(\mathbf{f}|\mathbf{y}) = -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - f(x_i)^2}{\sigma^2} + \mathbf{const}$$

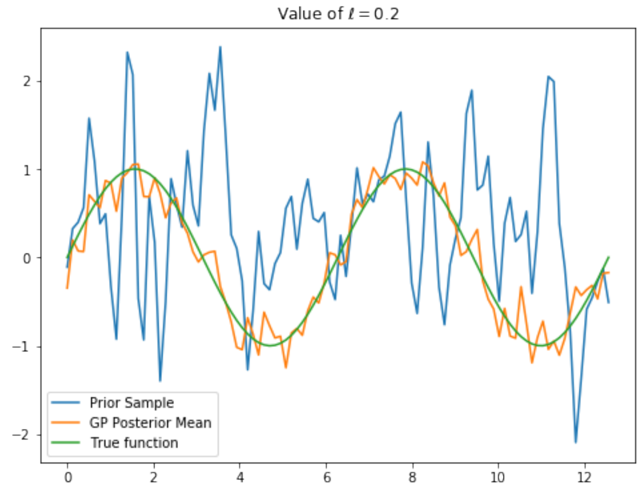
$$\implies log(p(\mathbf{f}|\mathbf{y}) = -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\frac{\sum_{i=1}^{N} f(x_i)^2}{\sigma^2} + \frac{y_i f(x_i)}{\sigma^2} + \mathbf{const}$$

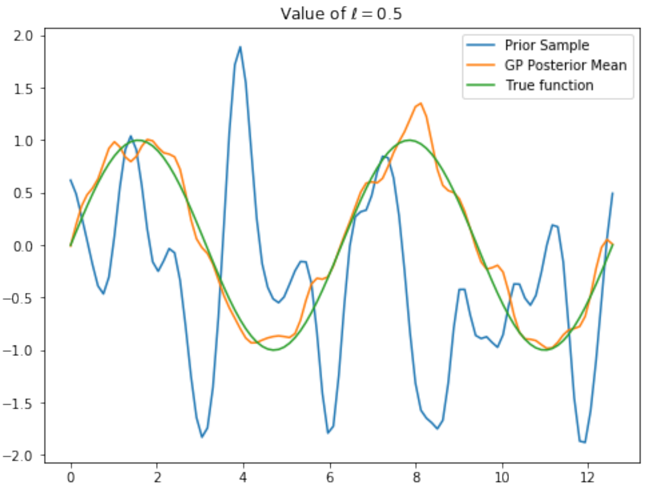
$$= -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\mathbf{f}^T(\sigma^2)^{-1}\mathbf{f} + \frac{\mathbf{f}^T\mathbf{y}}{\sigma^2} + \mathbf{const}$$

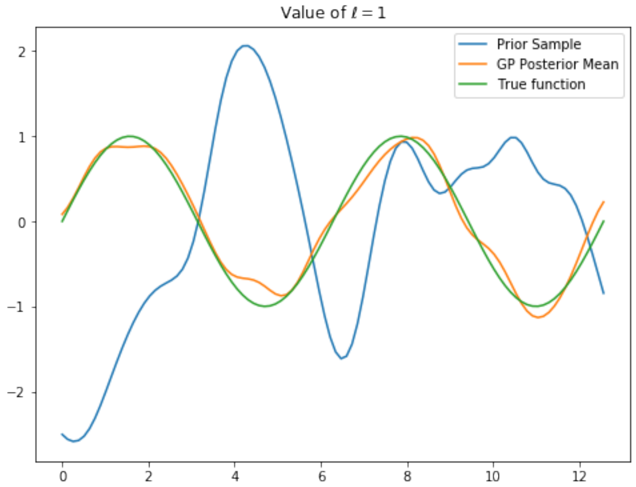
$$= -\frac{1}{2}\mathbf{f}^T(\mathbf{K}^{-1} + (\sigma^2)^{-1}I_N)\mathbf{f} + \frac{2\mathbf{f}^T\mathbf{y}}{2\sigma^2} + \mathbf{const}$$

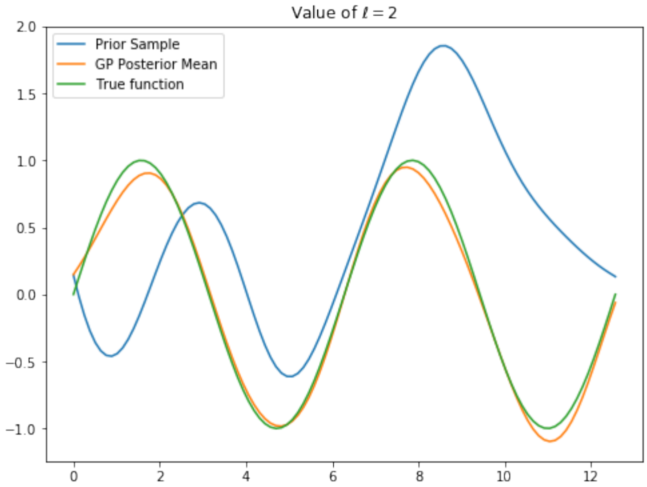
Collecting square terms and comparing the expression with the log-pdf of a multivariate normal, it is obvious that,

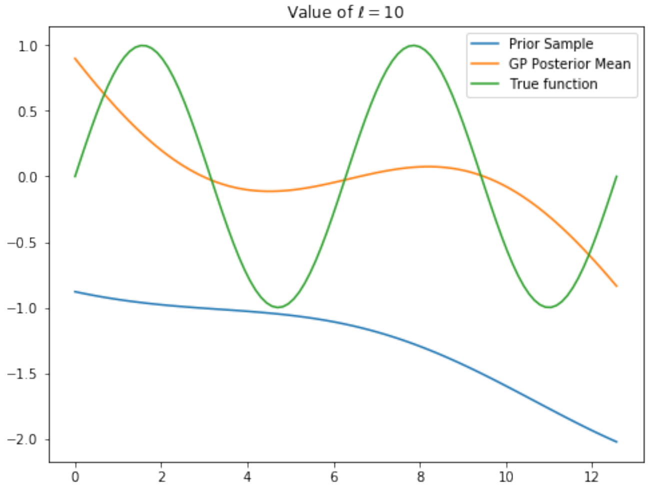
$$p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mu_*, \Sigma_*) \;\; \textbf{where,}$$
$$\mu_* = \mathbf{K}^T(\mathbf{K} + \sigma^2)^{-1}\mathbf{y} \;\; \textbf{and} \;\; \Sigma_* = \mathbf{K}^T(\mathbf{K} + \sigma^2)^{-1}\sigma^2$$











As the value of ℓ is increased from 0.2 to 10, the plots of the prior and posterior mean get a lot smoother. The plots for small values are very "spikey". However for $\ell = 10$, the GP posterior is unable to capture the variability in the true function and is too smooth. As a result, posterior mean is not accurate at all. For all other values, the posterior mean is more or less accurate. Hence, for smaller values of ℓ, it captures more complexity in the true function and is prone to overfitting as well. Hence is it more "spikey". Posterior mean seems to be the most accurate for $\ell = 1$ and $\ell = 2$ by plot observation.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION

5

1. The expression for the posterior predictive distribution for the output y_* of a new input x_* can be found as,

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

By Bayes' Rule,

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

It is given that,

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

where $\tilde{\mathbf{K}}^{-1}$ is the $M \times M$ kernel matrix of the pseudo inputs $\mathbf{Z}$ and $\tilde{\mathbf{k}}_n$ is the $M \times 1$ vector of kernel based similarities of x_n with pseudo-inputs.

$$\implies p(\mathbf{f}|\mathbf{X}, \mathbf{t}, \mathbf{Z}) = \prod_{n=1}^{N} p(f_n|x_n, \mathbf{Z}, \mathbf{t}) = \mathcal{N}(\mathbf{f}|\mathbf{A}\tilde{\mathbf{K}}^{-1}\mathbf{t}, \Sigma_f)$$

where $\mathbf{A}$ is a $N \times M$ matrix with $(\mathbf{A})_{nm} = \kappa(x_n, z_m)$ and Σ_f is a diagonal matrix with $(\Sigma_f)_{ii} = \kappa(x_i, x_i) - \tilde{\mathbf{k}}_i^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_i$. The pseudo points are sampled by the same noiseless GP and therefore, $p(\mathbf{t}|\mathbf{Z}) = \mathcal{N}(\mathbf{t}|0, \tilde{\mathbf{K}})$. Using Gaussian-Gaussian conjugacy, we can expand terms on the RHS of the proportionality above and compare expression with pdf of a multivariate normal to find the posterior on $\mathbf{t}$,

$$p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) = \mathcal{N}(\mu_{\mathbf{t}|\mathbf{f}}, \Sigma_{\mathbf{t}|\mathbf{f}}) \quad \text{where,}$$

$$\Sigma_{\mathbf{t}|\mathbf{f}} = ((\tilde{\mathbf{K}}^{-1})^T\mathbf{A}^T\Sigma_f^{-1}\mathbf{A}\tilde{\mathbf{K}}^{-1} + \tilde{\mathbf{K}}^{-1})^{-1} \quad \text{and} \quad \mu_{\mathbf{t}|\mathbf{f}} = \Sigma_{\mathbf{t}|\mathbf{f}}\tilde{\mathbf{K}}^{-1}\mathbf{A}^T\Sigma_f^{-1}\mathbf{f}$$

Since $\tilde{\mathbf{K}}^{-1} = (\tilde{\mathbf{K}}^{-1})^T$, the above two expressions can be simplified further to,

$$\Sigma_{\mathbf{t}|\mathbf{f}} = \tilde{\mathbf{K}}(\mathbf{A}^T\Sigma_f^{-1}\mathbf{A} + \tilde{\mathbf{K}})^{-1}\tilde{\mathbf{K}} \text{ and } \mu_{\mathbf{t}|\mathbf{f}} = \tilde{\mathbf{K}}(\mathbf{A}^T\Sigma_f^{-1}\mathbf{A} + \tilde{\mathbf{K}})^{-1}\mathbf{A}^T\Sigma_f^{-1}\mathbf{f}$$

Since we are considering the noiseless setting, $y_* = f_*$ and therefore,

$$p(y_*|x_*, \mathbf{Z}, \mathbf{t}) = \mathcal{N}(f_*|\tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\mathbf{t}, \kappa(x_*, x_*) - \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*)$$

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

Using the Linear Transformation property for Gaussian random vectors on the vector t, we get

$$p(y_*|x_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) = \int p(y_*|x_*, \mathbf{Z}, \mathbf{t})p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})d\mathbf{t} = \mathcal{N}(\mu_*, \Sigma_*)$$

$$\text{where } \mu_* = \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\mu_{\mathbf{t}|\mathbf{f}} \text{ and } \Sigma_* = \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\Sigma_{\mathbf{t}|\mathbf{f}}(\tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1})^T + \kappa(x_*, x_*) - \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*$$

$$\implies \mu_* = \tilde{\mathbf{k}}_*^T (\mathbf{A}^T \Sigma_f^{-1} \mathbf{A} + \tilde{\mathbf{K}})^{-1} \mathbf{A}^T \Sigma_f^{-1} \mathbf{f}$$

$$\Sigma_* = \tilde{\mathbf{k}}_*^T (\mathbf{A}^T \Sigma_f^{-1} \mathbf{A} + \tilde{\mathbf{K}})^{-1} \tilde{\mathbf{k}}_* + \kappa(x_*, x_*) - \tilde{\mathbf{k}}_*^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{k}}_*$$

In terms of computational cost, the dominant term here would be the computation of the term $\mathbf{A}^T \Sigma_f^{-1} \mathbf{A}$. Since $\mathbf{A}$ is a $N \times M$ matrix and Σ_f is a N dimensional diagonal matrix, the computational complexity for this posterior predictive would be $O(M^2 N)$ which is significantly lesser than $O(N^3)$ for $M << N$.

2. The marginal likelihood can be derived as follows,

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

$$= \mathcal{N}(\mathbf{f}|0, \mathbf{A}\tilde{\mathbf{K}}^{-1}\mathbf{A}^T + \Sigma_f) \ \ \text{[using the linear transform property]}$$

The expression for the above multivariate normal distribution is our MLE-II objective which can be maximised with respect to Z to find our pseudo inputs.