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

QUESTION 1

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: April 20, 2021

(1.) Equation (4) in the paper defines "expected complete data log posterior". This quantity means the expected value of the posterior obtained if we had the complete data.

$$\begin{aligned}\mathbb{E}_{\mathcal{Y}_p}\left[\log p\left(\theta|\mathcal{D}_0 \cup (\mathcal{X}_p, \mathcal{Y}_p)\right)\right] = \log p(\theta|\mathcal{D}_0) + \sum_{m=1}^{M} \mathcal{L}_m(\theta) \\ \text{where, } \mathcal{L}_m(\theta) = \left(\mathbb{E}_{y_m}\left[\log p(y_m|x_m,\theta)\right] + \mathbb{H}[y_m|x_m,\mathcal{D}_0]\right)\end{aligned} \tag{1}$$

Here, $\mathcal{D}_0$ is the labeled dataset, $\mathcal{X}_p = \{x_m\}_{m=1}^{M}$ is the unlabeled pool set and $\mathcal{Y}_p = \{y_m\}_{m=1}^{M}$ are the corresponding labels obtained using an oracle labeling mechanism.

Hence, the complete dataset would be $\mathcal{D}_0 \cup (\mathcal{X}_p, \mathcal{Y}_p)$ and the complete data log posterior would be $\log p\left(\theta|\mathcal{D}_0 \cup (\mathcal{X}_p, \mathcal{Y}_p)\right)$. The complete data posterior is optimal for Bayesian learning, however it would be costly to obtain it. The above equation tries to find the "expected" value of the complete data log posterior.

A batch $\mathcal{D}'$ is obtained such that the updated log posterior, i.e. $p\left(\theta|\mathcal{D}_0 \cup \mathcal{D}'\right)$ best approximates the "expected complete data log posterior" given in the equation 1.

The sparse approximation based objective function is

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

Here, $\mathcal{L} = \sum_m \mathcal{L}_m$ and $\mathcal{L}(\mathbf{w}) = \sum_m w_m \mathcal{L}_m$, b is a query budget and $w_m \in \{0,1\}^M$ is a weight vector indicating which points to include in our batch $\mathcal{D}'$.

Since the first term in equation 1 only depends on $\mathcal{D}_0$, hence we need to approximate $\sum_{m=1}^{M} \mathcal{L}_m(\theta)$ such that the resulting posterior is close to the "expected complete data log posterior". Since $\mathcal{L}$ represents the corresponding $\mathcal{L}_m(\theta)$ term for the complete data, hence finding an $\mathcal{L}(\mathbf{w})$ close to $\mathcal{L}$ would be a good approximation. This is ensured by minimizing $||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2$, i.e. the equation 2.

(2.) The sparse approximation based objective (equation 2) is difficult to optimize. The following relaxed objective is minimized instead in the paper:

$$\mathbf{w}^* = \min_{\mathbf{w}} (\mathbf{1} - \mathbf{w})^\top \mathbf{K} (\mathbf{1} - \mathbf{w}) \quad \text{subject to} \quad w_m \geq 0 \ \ \forall m, \sum_m w_m \sigma_m = \sigma \tag{3}$$

Here, $\sigma_m = ||\mathcal{L}_m||, \sigma = \sum_m \sigma_m$ and $\mathbf{K} \in \mathbb{R}^M$ is a Kernel matrix with $(\mathbf{K})_{mn} = \langle \mathcal{L}_m, \mathcal{L}_n \rangle$.

Equation 3 is different from equation 2 in the respect that the cardinality constraint $||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2$ is replaced with the polytope constraint $(\mathbf{1} - \mathbf{w})^\top \mathbf{K} (\mathbf{1} - \mathbf{w})$ and the constraint of $w_m \in \{0,1\}$ is relaxed to $w_m \geq 0$. This new relaxed optimization is solved using the Frank-Wolfe algorithm[1].

[1] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. Naval Research Logistics Quarterly, 3(1-2):95–110, 1956.

The main computation step in their algorithm is

$$\left\langle \mathcal{L} - \mathcal{L}(\mathbf{w}), \frac{1}{\sigma_n}\mathcal{L}_n \right\rangle = \frac{1}{\sigma_n}\sum_{m=1}^{N}(1 - w_m)\langle \mathcal{L}_m, \mathcal{L}_n \rangle$$

At each iteration, the algorithm uses greedy method to find the vector $\mathcal{L}_p$ which minimizes the residual error $(\mathcal{L} - \mathcal{L}(\mathbf{w}))$. The weights are then updated accordingly. After b iterations the algorithm returns the optimal weights $\mathbf{w}^*$. Since our original weights were binary, i.e. 0 or 1, hence we set $\tilde{w}_m^* = 1$ if $w_m^* > 0$ and 0 otherwise. Hence, the final AL batch becomes $\mathcal{D}' = \{x_m \in \mathcal{X}_p | w_m^* > 0\}$.

The paper employs different choices of the inner product $\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi}} = \mathbb{E}_{\hat{\pi}}[\langle \mathcal{L}_n, \mathcal{L}_m \rangle]$ where $\hat{\pi}$ is the current posterior $p(\theta|\mathcal{D}_0)$.

They have defined the weighted Fisher inner product which requires taking gradient of the expected log-likelihood terms w.r.t. the parameters:

$$\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},\mathcal{F}} = \mathbb{E}_{\hat{\pi}}\left[\nabla_\theta \mathcal{L}_n(\theta)^\top \nabla_\theta \mathcal{L}_m(\theta)\right]$$

Another type of inner product, the weighted Euclidean inner product, is based on the marginal likelihood of the data points:

$$\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},2} = \mathbb{E}_{\hat{\pi}}\left[\mathcal{L}_n(\theta)\mathcal{L}_m(\theta)\right]$$

(3.) The acquisition function proposed in the paper has a closed form expression for two types of models: (a) Bayesian linear regression and (b) Probit regression.

For other type of models where the acquisition function won't be available, the paper proposes to use random feature projections. They have considered projection for the weighted Euclidean inner product:

$$\hat{\mathcal{L}}_n = \frac{1}{\sqrt{J}}[\mathcal{L}_n(\theta_1), \cdots, \mathcal{L}_n(\theta_J)]^\top, \quad \theta_j \sim \hat{\pi}$$

Here, $\hat{\mathcal{L}}_n$ represents the J-dimensional projection of $\mathcal{L}_n$ in Euclidean space. Using this projection the paper approximates the inner products as

$$\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},2} \approx \hat{\mathcal{L}}_n^\top \hat{\mathcal{L}}_m$$

By using this approximation, we can get closed form expressions for models which don't have acquisition function.

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

QUESTION 2

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: April 20, 2021

We are given N scalar observations $x_1, x_2, .., x_N$ drawn i.i.d. from $\mathcal{N}(x|\mu, \beta^{-1})$, where $\mu \sim \mathcal{N}(\mu|\mu_0, s_0)$, and $\beta \sim Gamma(\beta|a, b)$.
We can write the conditional posterior for μ as

$$\begin{aligned} p(\mu|\mathbf{X}, \beta) &= \frac{p(\mathbf{X}|\mu, \beta)p(\mu)}{\int p(\mathbf{X}|\mu, \beta)p(\mu)d\mu} \\ &\propto p(\mathbf{X}|\mu, \beta)p(\mu) \\ &= \mathcal{N}(x|\mu, \beta^{-1})\mathcal{N}(\mu|\mu_0, s_0) \end{aligned}$$

Using completing the squares trick, we can write the posterior for μ as

$$\begin{aligned} p(\mu|\mathbf{X}, \beta) &= \mathcal{N}(\mu|\mu_N, s_N^{-1}) \\ \text{where, } \frac{1}{s_N} &= N\beta + \frac{1}{s_0} \\ \mu_N &= \frac{1}{1 + Ns_0\beta}\mu_0 + \frac{Ns_0\beta}{Ns_0\beta + 1}\bar{x} \end{aligned} \qquad (\bar{x} = \tfrac{\sum_{n=1}^{N} x_n}{N})$$

Similarly, we can write the conditional posterior for β as

$$\begin{aligned} p(\beta|\mathbf{X}, \mu) &\propto p(\mathbf{X}|\mu, \beta)p(\beta) \\ &= \mathcal{N}(x|\mu, \beta^{-1})Gamma(\beta|a, b) \end{aligned}$$

Using the properties of Gaussian model and the conjugacy of Gaussian and Gamma functions, we can write the posterior for β as

$$p(\beta|\mathbf{X}, \mu) = Gamma\left(\beta \middle| a + \frac{N}{2}, \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{2} + b\right)$$

As we obtained the conditional posteriors for μ and β in closed forms, we can use Gibbs sampling algorithm to approximate the joint posterior of μ and β by continuously drawing samples from their conditional posteriors. Following is the algorithm:

Algorithm 1: Gibbs sampling

```
Initialize $\mu_0$;
for $s = 1, 2, \cdots, S$ do
    Sample $\beta^{(s)}$ from the distribution $p(\beta|\mathbf{X}, \mu^{(s-1)})$;
    Sample $\mu^{(s)}$ from the distribution $p(\mu|\mathbf{X}, \beta^{(s)})$;
end
```

The samples $\{\mu^{(s)}, \beta^{(s)}\}_{s=1}^{S}$ will form the joint posterior $p(\mu, \beta|\mathbf{X})$.

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: April 20, 2021

QUESTION 3

(1.) The given prior on w helps in sparse learning of the weight parameters. However, the level of sparsity induced is of two types depending on γ_d, one with higher precision (when $\gamma_d = 0$) while the other with lower precision (when $\gamma_d = 1$).

(2.) Finding the **conditional posterior over the latent variable** (w).
Using Baye's rule, we can write

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

Since $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, we can write the likelihood as $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N)$. Also given the prior $p(\mathbf{w}|\sigma, \gamma) = \mathcal{N}(0, \sigma^2\mathbf{K})$, where $\mathbf{K}$ is a $D \times D$ diagonal matrix with $(\mathbf{K})_{ii} = \kappa_{\gamma_i}$ and $\kappa_{\gamma_d} = \gamma_d v_1 + (1 - \gamma_d) v_0$. Using completing the square trick, we can write the conditional posterior as

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma, \gamma) &= \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \\ \text{where, } \boldsymbol{\Sigma}_{\boldsymbol{w}} &= \sigma^2[\mathbf{K}^{-1} + \mathbf{X}^\top\mathbf{X}]^{-1} \\ \boldsymbol{\mu}_w &= \frac{1}{\sigma^2}\boldsymbol{\Sigma}_w\mathbf{X}^\top\mathbf{y} \end{aligned} \tag{4}$$

Finding the **Expectation of CLL**.
The complete data log likelihood (CLL) can be written as

$$\begin{aligned} \log p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma, \gamma) &= \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma) + \log p(\mathbf{w}|\sigma, \gamma) \\ &= -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{N+D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbf{w}^\top\mathbf{K}^{-1}\mathbf{w} \\ &\quad - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma_d}) \end{aligned} \tag{5}$$

The expected CLL can be written as

$$\mathbb{E}[CLL] = \mathbb{E}[\log p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma, \gamma)]$$

Using equation 5, it can be written as

$$\begin{aligned} \mathbb{E}[CLL] = &-\frac{1}{2\sigma^2}\left(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\mathbb{E}[\mathbf{w}] + \operatorname{Tr}\left[(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1})\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\right]\right) \\ &- \frac{N+D}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{d=1}^{D}\log(\kappa_{\gamma_d}) \end{aligned} \tag{6}$$

Since the posterior of $\mathbf{w}$ is Gaussian, we can directly write the expectations $\mathbb{E}[\mathbf{w}]$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^\top]$ as

$$\begin{aligned} \mathbb{E}[\mathbf{w}] = \boldsymbol{\mu}_w &= \left(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w\boldsymbol{\mu}_w^\top &= \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1}\right)^{-1} + \mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1}\right)^{-2}\mathbf{X}^\top\mathbf{y} \end{aligned} \tag{7}$$

Finding the **MAP estimates of the parameters** σ^2, γ and θ

$$\sigma^2, \gamma, \theta = \underset{\sigma^2,\theta,\gamma}{\arg\max} \left\{ \mathbb{E}\left[\log\left(p\left(\sigma^2, \gamma, \theta | \mathbf{y}, \mathbf{w}, \mathbf{X}\right)\right)\right] \right\}$$

Since the posterior over the parameters $p(\sigma^2, \gamma, \theta|\mathbf{y}, \mathbf{w}, \mathbf{X}) \propto p(\mathbf{w}, \mathbf{y}|\mathbf{X}, \sigma, \gamma).p(\sigma^2, \gamma, \theta)$, we can write the expectation of the log posterior as $\mathbb{E}[\log(p(\sigma^2, \gamma, \theta|\mathbf{y}, \mathbf{w}, \mathbf{X}))] = \mathbb{E}[CLL] + \log p(\sigma^2, \gamma, \theta)$, after ignoring the constants. Hence, the MAP estimate becomes

$$\sigma^2, \gamma, \theta = \underset{\sigma^2,\theta,\gamma}{\arg\max} \left\{ \mathbb{E}[CLL] + \log\left(p(\sigma^2, \gamma, \theta)\right) \right\} \tag{8}$$

Given the prior over the parameters $p(\sigma^2) = IG\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$, $p(\theta) = Beta(a_0, b_0)$ and $p(\gamma_d|\theta) = Bernoulli(\theta)$. We can write the joint prior over the parameters as

$$\begin{aligned} p(\sigma^2, \gamma, \theta) &= p(\sigma^2) \left\{ \prod_{d=1}^{D} p(\gamma_d|\theta) \right\} p(\theta) \\ \log p\left(\sigma^2, \gamma, \theta\right) &= \log\left(p(\sigma^2)\right) + \sum_{d=1}^{D} \log\left(p(\gamma_d|\theta)\right) + \log\left(p(\theta)\right) \end{aligned} \tag{9}$$

Plugging the values from equations 6, 9 in equation 8 and differentiating w.r.t. each parameter, we can find the MAP estimate of the parameters as follows:
Taking partial derivative w.r.t. σ^2 and ignoring other terms:

$$\begin{aligned} 0 &= \frac{\partial\left(\mathbb{E}[CLL] + \log\left(p(\sigma^2, \gamma, \theta)\right)\right)}{\partial(\sigma^2)} \\ 0 &= \frac{1}{2\sigma^4}\left(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\mathbb{E}[\mathbf{w}] + \mathrm{Tr}\left[(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1})\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\right]\right) - \frac{N+D}{2\sigma^2} - \frac{1}{\sigma^2}\left(\frac{\nu}{2} + 1\right) + \frac{\nu\lambda}{2\sigma^4} \\ \sigma^2_{\mathrm{MAP}} &= \frac{\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\mathbb{E}[\mathbf{w}] + \mathrm{Tr}\left[(\mathbf{X}^\top\mathbf{X} + \mathbf{K}^{-1})\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\right] + \nu\lambda}{N + D + \nu + 2} \end{aligned}$$

We don't need to take partial derivative w.r.t γ_d since it takes only binary values. Collecting γ_d terms and maximizing w.r.t. it:

$$\begin{aligned} \gamma_d &= \underset{\gamma_d\in\{0,1\}}{\arg\max} \left\{ \mathbb{E}[CLL] + \log p(\sigma^2, \gamma, \theta) \right\} \\ \gamma_{d_{\mathrm{MAP}}} &= \underset{\gamma_d\in\{0,1\}}{\arg\max} \left\{ -\frac{1}{2\sigma^2\kappa_{\gamma_d}}(\mathbb{E}[\mathbf{w}\mathbf{w}^\top])_{d,d} - \frac{1}{2}\log\left(\kappa_{\gamma_d}\right) + \gamma_d\log\theta + (1-\gamma_d)\log(1-\theta) \right\} \\ \gamma_{\mathrm{MAP}} &= [\gamma_0, \gamma_1, \cdots, \gamma_D] \end{aligned}$$

Taking partial derivative w.r.t. θ and ignoring other terms:

$$\begin{aligned} 0 &= \frac{\partial\left(\mathbb{E}[CLL] + \log\left(p(\sigma^2, \gamma, \theta)\right)\right)}{\partial(\theta)} \\ &= \frac{1}{\theta}\left(\sum_{d=1}^{D}\gamma_d + a_0 - 1\right) - \frac{1}{1-\theta}\left(\sum_{d=1}^{D}(1-\gamma_d) + b_0 - 1\right) \\ \theta_{\mathrm{MAP}} &= \frac{\sum_{d=1}^{D}\gamma_d + a_0 - 1}{D + a_0 + b_0 - 2} \end{aligned}$$

The final EM algorithm is given in Algorithm 2.

```
Algorithm 2: EM Algorithm
Initialize $\{\sigma^2, \gamma, \theta\} \leftarrow \{\sigma^{2^{(0)}}, \gamma^{(0)}, \theta^{(0)}\}$
$\mathbf{K}^{(0)} \leftarrow$ diagonal matrix with $(\mathbf{K}^{(0)})_{ii} = \kappa^{(0)}_{\gamma_i}$ and $\kappa^{(0)}_{\gamma_d} = \gamma^{(0)}_d v_1 + (1 - \gamma^{(0)}_d) v_0$
for $t = 0, 1, \cdots, T-1$ do
    (a) Update the posterior of latent variable $\mathbf{w}$ as:
        $p\left(\mathbf{w}^{(t+1)} \middle| \mathbf{y}, \mathbf{X}, \sigma^{(t)}, \gamma^{(t)}\right) = \mathcal{N}\left(\mathbf{w}^{(t+1)} \middle| \boldsymbol{\mu}^{(t)}_w, \boldsymbol{\Sigma}^{(t)}_w\right)$
        $\boldsymbol{\Sigma}_{\boldsymbol{w}}{}^{(t+1)} = (\sigma^2)^{(t)} \left((\mathbf{K}^{(t)})^{-1} + \mathbf{X}^\top \mathbf{X}\right)^{-1}$
        $\boldsymbol{\mu}^{(t+1)}_w = \frac{1}{(\sigma^2)^{(t)}} \boldsymbol{\Sigma}^{(t+1)}_w \mathbf{X}^\top \mathbf{y}$
    (b) Update the expectations as:
        $\mathbb{E}[\mathbf{w}]^{(t+1)} = \boldsymbol{\mu}^{(t+1)}_w$
        $\mathbb{E}[\mathbf{w}\mathbf{w}^\top]^{(t+1)} = \boldsymbol{\Sigma}^{(t+1)}_w + \boldsymbol{\mu}^{(t+1)}_w \boldsymbol{\mu}^{(t+1)^\top}_w$
    (c) Update the parameters as:
        $(\sigma^2)^{(t+1)} = \frac{\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X} \mathbb{E}[\mathbf{w}]^{(t+1)} + \mathrm{Tr}\left[\left(\mathbf{X}^\top \mathbf{X} + \left(\mathbf{K}^{(t)}\right)^{-1}\right) \mathbb{E}[\mathbf{w}\mathbf{w}^\top]^{(t+1)}\right] + \nu\lambda}{N + D + \nu + 2}$
        $\theta^{(t+1)} = \frac{\sum_{d=1}^{D} \gamma^{(t)}_d + a_0 - 1}{D + a_0 + b_0 - 2}$
        $\gamma^{(t+1)}_d = \underset{\gamma_d \in \{0,1\}}{\arg\max} \{ -\frac{1}{2(\sigma^2)^{(t+1)} \kappa^{(t)}_{\gamma_d}} \mathbb{E}[\mathbf{w}\mathbf{w}^\top]^{(t+1)}_{(d,d)} - \frac{1}{2} \log\left(\kappa^{(t)}_{\gamma_d}\right) + \gamma^{(t)}_d \log\left(\theta^{(t+1)}\right)$
            $+ (1 - \gamma^{(t)}_d) \log\left(1 - \theta^{(t+1)}\right) \}$
    (d) Update the Kernel matrix $\mathbf{K}$ as:
        $(\mathbf{K}^{(t+1)})_{ii} = \kappa^{(t+1)}_{\gamma_i}$ and $\kappa^{(t+1)}_{\gamma_d} = \gamma^{(t+1)}_d v_1 + (1 - \gamma^{(t+1)}_d) v_0$
end
Return $p\left(\mathbf{w} | \mathbf{y}, \mathbf{X}, \sigma^{(T-1)}, \gamma^{(T-1)}\right)$ and $\{\gamma, \sigma^2, \theta\} = \{\gamma^{(T)}, (\sigma^2)^{(T)}, \theta^{(T)}\}$
```

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: April 20, 2021

QUESTION 4

(1.) Given the likelihood model $p(y_n|\mathbf{x}_n, \mathbf{f}) = \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2)$ for each observation. Assuming i.i.d. observations, the likelihood model can be written as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{f}) &= \prod_{n=1}^{N} \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2) \\ &= \mathcal{N}\left(\mathbf{y}\middle|\mathbf{f}, \sigma^2\mathbf{I}_N\right) \end{aligned}$$

Also given the prior

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$$

where $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2)..., f(\mathbf{x}_N)]^T$ and $\mathbf{K}$ is a $N \times N$ matrix with entries $\mathbf{K}_{nm} = \kappa(\mathbf{x}_n, \mathbf{x}_m)$ Using Baye's rule, we can write the GP posterior as

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

Using the standard properties of Gaussian model, we can write the GP posterior as

$$\begin{aligned} p(\mathbf{f}|\mathbf{y}) &= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (10) \\ \text{where, } \boldsymbol{\mu} &= (\sigma^2\mathbf{K}^{-1} + \mathbf{I}_N)^{-1}\mathbf{y} \\ \boldsymbol{\Sigma} &= (\sigma^2\mathbf{K}^{-1} + \mathbf{I}_N)^{-1}\sigma^2 \end{aligned}$$

(2.) Figure 1 contains the required plots.

We observe that as the value of l increases, the random sample from prior becomes more smooth. As we see that for smaller values of l, i.e. $l = 0.2, 0.5$, the mean of the GP posterior is noisy but close to the true function $\sin(x)$. For $l = 1, 2$, the mean of the posterior is almost close to the true function and fits well. For large values of l, i.e. $l = 10$, the mean of the posterior varies significantly from the true function.

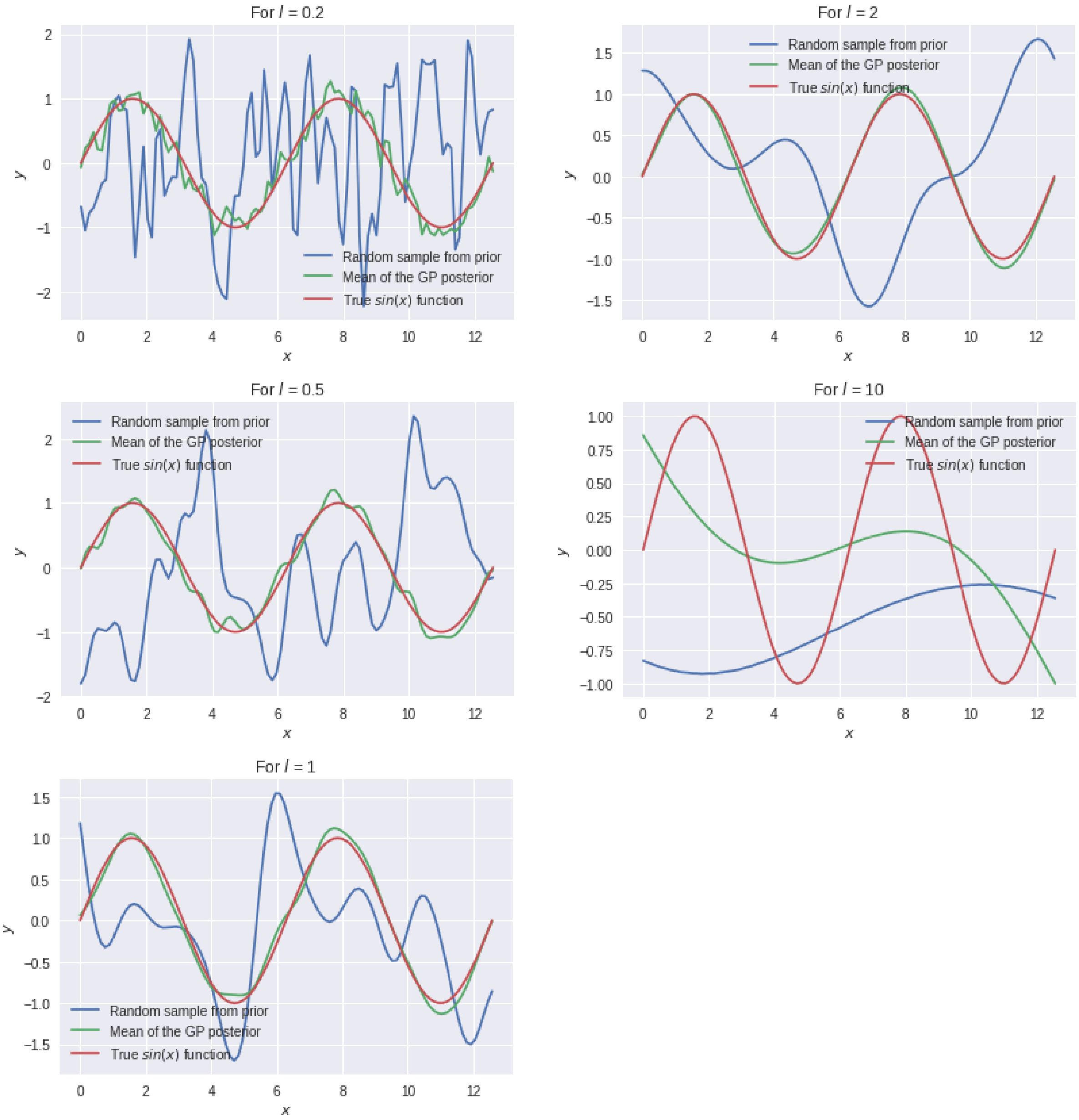

Figure 1: Plots for (a) random sample from the GP prior, (b) mean of the GP posterior and (c) true function $\sin(x)$, for different values of l.

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

Student Name: Bholeshwar Khurana
Roll Number: 170214
Date: April 20, 2021

QUESTION 5

(1.) Given the likelihood with pseudo training data $(\mathbf{Z}, \mathbf{t})$,

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

Here, $\tilde{\mathbf{K}}$ 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 each of the pseudo inputs $z_1, \cdots, z_M$.
Assuming i.i.d. observations, we can write the likelihood as

$$\begin{aligned} p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) &= \prod_{n=1}^{N} p(f_n|\mathbf{x}_n, \mathbf{Z}, \mathbf{t}) \\ &= \mathcal{N}(\mathbf{f}|\mathbf{P}\tilde{\mathbf{K}}^{-1}\mathbf{t}, \mathbf{\Lambda}) \end{aligned} \tag{11}$$

Here, $\mathbf{P}$ is a $N \times M$ matrix with $(\mathbf{P})_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$ and $\mathbf{\Lambda}$ is a $N \times N$ diagonal matrix with $(\mathbf{\Lambda})_{ii} = \kappa(\mathbf{x}_i, \mathbf{x}_i) - \tilde{\mathbf{k}}_n^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_n$.
We can write the posterior predictive distribution for the output y_* as:

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

Using Baye's rule, we can write the posterior over $\mathbf{t}$ as

$$p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) \propto p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) p(\mathbf{t}|\mathbf{Z}) \tag{12}$$

Since the pseudo training data is generated using Gaussian process, hence we can write

$$p(\mathbf{t}|\mathbf{Z}) = \mathcal{N}(\mathbf{t}|0, \tilde{\mathbf{K}}) \tag{13}$$

Using equations 11, 12, 13 and the knowledge of Gaussians, we can write the posterior over t as

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) &= \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_{\boldsymbol{t}|\boldsymbol{f}}, \mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}}) \\ \text{where, } \mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}} &= (\tilde{\mathbf{K}}^{-1}\mathbf{P}^\top \mathbf{\Lambda}^{-1}\mathbf{P}\tilde{\mathbf{K}}^{-1})^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{t}|\boldsymbol{f}} &= \mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}} \tilde{\mathbf{K}}^{-1}\mathbf{P}^\top \mathbf{\Lambda}^{-1}\mathbf{f} \end{aligned}$$

Since we have a noiseless setting, i.e. $y_* = f_*$, we can write $f_* = \tilde{\mathbf{k}}_*^\top \tilde{\mathbf{K}}^{-1}\mathbf{t} + \epsilon$ where $\tilde{\mathbf{k}}_*$ is $M \times 1$ matrix with $(\tilde{\mathbf{k}}_*)_i = \kappa(\mathbf{x}_*, \mathbf{z}_i)$ and $\epsilon \sim \mathcal{N}(0, \kappa(\mathbf{x}_*, \mathbf{x}_*) - \tilde{\mathbf{k}}_*^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*)$. Using the properties of Linear Gaussian model, we can write the posterior predictive distribution as

$$\begin{aligned} p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) &= \mathcal{N}(f_*|\boldsymbol{\mu}_*, \mathbf{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \tilde{\mathbf{k}}_*^\top \tilde{\mathbf{K}}^{-1}\mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}} \tilde{\mathbf{K}}^{-1}\mathbf{P}^\top \mathbf{\Lambda}^{-1}\mathbf{f} \\ \mathbf{\Sigma}_* &= \tilde{\mathbf{k}}_*^\top \tilde{\mathbf{K}}^{-1}\mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}} \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_* + \kappa(\mathbf{x}_*, \mathbf{x}_*) - \tilde{\mathbf{k}}_*^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_* \end{aligned}$$

The computation cost of $\tilde{\mathbf{K}}^{-1}$ is $\mathcal{O}(M^3)$ while that for $\mathbf{\Sigma}_{\boldsymbol{t}|\boldsymbol{f}}$ is $\mathcal{O}(M^2N)$. Since $M << N$, hence the time complexity for computing this posterior predictive is $\mathcal{O}(M^2N)$ which is significantly less than the earlier complexity of $\mathcal{O}(N^3)$.

(2.) We can write the marginal likelihood as

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

Also, $\mathbf{f} = \mathbf{P}\tilde{\mathbf{K}}^{-1}\mathbf{t} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\boldsymbol{\Lambda})$ and $p(\mathbf{t}|\mathbf{Z}) = \mathcal{N}(\mathbf{t}|0,\tilde{\mathbf{K}})$ (Equation 13). Using the properties of linear Gaussian model, we can write

$$\begin{aligned} p(\mathbf{f}|\mathbf{X},\mathbf{Z}) &= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}',\boldsymbol{\Sigma}') \\ \boldsymbol{\mu}' &= \mathbf{P}\tilde{\mathbf{K}}^{-1}\mathbf{0} = \mathbf{0} \\ \boldsymbol{\Sigma}' &= \mathbf{P}\tilde{\mathbf{K}}^{-1}\mathbf{P}^{\top} + \boldsymbol{\Lambda} \end{aligned}$$

The MLE-II objective to obtain $\mathbf{Z}$ is to maximize the marginal likelihood.

$$\begin{aligned} \hat{\mathbf{Z}}_{\text{MLE-II}} &= \arg\max_{\mathbf{Z}} p(\mathbf{f}|\mathbf{X},\mathbf{Z}) \\ &= \arg\min_{\mathbf{Z}}(\log|\boldsymbol{\Sigma'}| + \mathbf{f}^{\top}\boldsymbol{\Sigma'}^{-1}\mathbf{f}) \end{aligned}$$