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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: April 20, 2021

QUESTION 1

- The complete data log posterior refers to the posterior distribution $P(\boldsymbol{\theta}|\mathcal{D}_o, \mathcal{D}_p)$, which assumes access to the labels of the entire unlabelled pool set.
  Since the labels of the entire pool are not known to us, we try to approximate it by inferring these labels using the current predictive posterior distribution, i.e. $p(\mathcal{Y}_p|\mathcal{X}_p, \mathcal{D}_o)$, which can be computed with the given information. This PPD is calculated by marginalizing the likelihood over the current posterior, i.e. $p(\boldsymbol{\theta}|\mathcal{D}_o)$.
  By taking an expectation over this PPD, we compute approximate complete data log posterior $p(\boldsymbol{\theta}|\mathcal{D}_o, \mathcal{D}_p)$(Expected CDL) as well as the approximate posterior obtained after querying the labels for the next batch(Updated Posterior), i.e. $p(\boldsymbol{\theta}|\mathcal{D}_o, \mathcal{D}')$.
  The objective function in the paper tries to minimize the difference in these inferred posteriors, i.e. expected complete data log posterior and the posterior obtained after selecting the next b samples out of $\mathcal{D}_p$. The reason behind this matching is that we want the next b samples to have the same effect on model as by training it on the entire data $\mathcal{D}_p$.

  The above condition reduces to a sparse approximation based objective function in the steps discussed below:

  - ◇ Using Bayes Rule, expand the Expected CDL to obtain Eq[4].
  - ◇ A similar form can be obtained for the updated posterior.
  - ◇ The first term (prior) is same for both. Hence, matching the posteriors is equivalent to matching the second, i.e.$\mathcal{L}$ term.
  - ◇ The $\mathcal{L}$ term in takes a sum over M training examples for Expected CDL, and at most b samples in case of the Updated Posterior. However, the individual terms, i.e.$\mathcal{L}_m(\theta)$ will be the same in both the cases since they depend only on the likelihood and the PPD(marginal likelihood) term.
  - ◇ Hence, a parameter $\mathbf{w}$, a M-dimensional binary vector, is introduced to represent the $\mathcal{L}$ term of the updated posterior as $\mathcal{L}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathcal{L}_m$
  - ◇ This vector $\mathbf{w}$ can have at most b 1's in it and this reduces the problem to a constrained optimization problem on the parameter $\mathbf{w}$ as below:

  $$\arg\min_{\mathbf{w}} ||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2 \quad s.t. \;\; \mathbf{w}_m \in \{0,1\}, \; \sum_{m=1}^{M} w_m \leq b \tag{1}$$

  where $\mathcal{L} = \sum_{m=1}^{M} \mathcal{L}_m(\theta)$, $\mathcal{L}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathcal{L}_m(\theta)$

  On solving the objective with the given constraints, the entries of $\mathbf{w}$ determine whether the example should be chosen for the next batch ($w_m = 1$) or not ($w_m = 0$).
  Since the cardinality of $\mathbf{w}$ can be at most b, it ensures that the next batch of samples queried from the oracle contain at most b data points.

- In the relaxed objective, the weight vector is assumed to be a vector over non-negative reals. In addition, $\sigma_m = ||\mathcal{L}_m||$ is introduced such that the polytope constraint $\sum_{m=1}^{M} w_m \sigma_m = \sum_{m=1}^{M} \sigma_m$ replaces the cardinality constraint.
  The polytope(convex hull) is constructed over an M-Dimensional Hilbert Space with vertices $\frac{\sum_{m=1}^{M} \sigma_m}{\sigma_m} \check{1}_m$, where $\check{1}_m$ is the unit basis vector along m.
  The optimization objective is thus, expressed in the form below:
  $$\mathbf{w}^* = \arg\min_{\mathbf{w}} (1 - \mathbf{w})^T \mathbf{K} (1 - \mathbf{w}) \quad s.t. \;\; \mathbf{w}_m \geq 0, \; \sum_{m=1}^{M} w_m \sigma_m = \sum_{m=1}^{M} \sigma_m \qquad (2)$$
  The matrix $\mathbf{K}$ is defined as $\mathbf{K}_{mn} =< \mathcal{L}_m, \mathcal{L}_n >$, where $<>$ can be a Fischer or a weighted-Euclidean inner product.
  The objective in Equation 1 translates to Eq 2 because:
  $$\begin{aligned} ||\mathcal{L} - \mathcal{L}(\mathbf{w})||^2 &= (\sum_{m=1}^{M} \mathcal{L}_m - \sum_{m=1}^{M} \mathbf{w}_m \mathcal{L}_m)^2 \\ &= (\sum_{m=1}^{M} (1 - \mathbf{w}_m) \mathcal{L}_m)^2 = \prod_{m,m'} (1 - \mathbf{w}_m) \mathcal{L}_m \mathcal{L}'_m (1 - \mathbf{w}'_m) \\ &= (1 - \mathbf{w})^T \mathbf{K} (1 - \mathbf{w}) \end{aligned}$$
  The objective 2 is efficiently solved using b iterations of **Frank-Wolfe algorithm** as discussed in the paper. This algorithm computes the inner product $< cL - \mathcal{L}(\mathbf{w}), \frac{\mathcal{L}_m}{\sigma_m} >$ for each m and selects the sample f whose $\mathcal{L}_f$ is most aligned with the vector $cL - \mathcal{L}(\mathbf{w})$ in the M-dimensional Hilbert Space in each iteration. The resulting solution for $\mathbf{w}$ will thus have at most $\leq b$ non-zero entries.

  Finally, as the solution obtained for $\mathbf{w}$ contains non-negative real entries, we then use the **projection step** to obtain $\mathbf{w}$ in the desired form(as a binary vector). The projection step:
  $$w_m = \begin{cases} 1 & w_m^* \geq 0 \\ 0 & otherwise \end{cases}$$
  gives us the solution for $\mathbf{w}$,i.e., the samples to be queried for the next batch.
- The objective(acquisition function) has a closed form solution in case of **Bayesian Linear regression** and **Probit Regression**, when using Fischer Inner Product.
  Further, for the case of Bayesian Linear regression, the form of the inner product can be directly compared with the acquisition function used in BALD.

  For cases where these inner products are intractable, the paper proposes an approximation technique to compute the inner products using **random feature projections**. This technique allows to extend the batch construction method to any non-linear model for which the likelihood is tractable.
  The paper defines the notion of projections using which the inner products are approximated. The J-dimensional projection of $\mathcal{L}_n$ is represented as $\hat{\mathcal{L}}_n$, and is defined for the case of Euclidean inner products as:
  $$\begin{aligned} \hat{\mathcal{L}}_n &= \frac{1}{\sqrt{J}} [\mathcal{L}_n(\boldsymbol{\theta}_1) \ldots \mathcal{L}_n(\boldsymbol{\theta}_J)], \\ \implies < \mathcal{L}_n, \mathcal{L}_m >_{\pi,2} &\approx \hat{\mathcal{L}}_n^T \hat{\mathcal{L}}_m \end{aligned}$$

where the Weighted Euclidean inner product is defined as:

$$< \mathcal{L}_n, \mathcal{L}_m >_{\pi,2} \quad = \quad \mathbb{E}_\pi[\mathcal{L}_n(\boldsymbol{\theta})\mathcal{L}_m(\boldsymbol{\theta}_1)], \quad where \quad \pi = p(\boldsymbol{\theta}|\mathcal{D}_o) \tag{3}$$

This approximate inner products are then used during the Frank-Wolfe optimization procedure.

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: April 20, 2021

QUESTION 2

For mean:

$$p(\mu|x_1, x_2 \dots x_N, \beta) = \frac{\mathcal{N}(\mu|\mu_o, s_o) \prod_{i=1}^{N} \mathcal{N}(x|\mu, \beta^{-1})}{\int \mathcal{N}(\mu|\mu_o, s_o) \prod_{i=1}^{N} \mathcal{N}(x|\mu, \beta^{-1}) d\mu}$$

Here, we assume the variance β^{-1} is given. Hence, both the prior and likelihood are gaussians and conjugate to each other. On computing the conditional posterior using the results of Bayesian Inference for Mean of a Univariate Gaussian, we get:

$$p(\mu|x_1, x_2 \dots x_N, \beta) = \mathcal{N}\left(\mu| \frac{\mu_o}{N s_o \beta + 1} + \frac{N s_o \beta \bar{x}}{N s_o \beta + 1}, (so^{-1} + N\beta)^{-1}\right)$$
$$where \quad \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

For variance:

$$p(\beta|x_1, x_2 \dots x_N, \mu) = \frac{Gamma(\beta|a, b) \prod_{i=1}^{N} \mathcal{N}(x|\mu, \beta^{-1})}{\int Gamma(\beta|a, b) \prod_{i=1}^{N} \mathcal{N}(x|\mu, \beta^{-1}) d\beta}$$

Here, we assume the mean μ is given. Hence, both the prior(Gamma) and likelihood(Gaussian) are conjugate to each other. On computing the conditional posterior using the results of Bayesian Inference for Precision of a Univariate Gaussian, we get:

$$p(\beta|x_1, x_2 \dots x_N, \mu) = Gamma\left(a + \frac{N}{2}, b + \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2}\right)$$

Gibbs sampling algorithm for joint posterior:

1. Initialize $\beta = \beta^{(0)}$
2. For $s = 1, 2, \dots S$
    - Draw a random sample for μ as $\mu^{(s)} \sim p(\mu|x_1, x_2 \dots x_N, \beta^{(s-1)})$
    - Draw a random sample for β as $\beta^{(s)} \sim p(\beta|x_1, x_2 \dots x_N, \mu^{(s)})$
3. The S samples $\left(\mu^{(s)}, \beta^{(s)}\right)_{s=1}^{S}$ collectively represent the joint posterior $p(\mu, \beta|x_1, x_2 \dots x_N)$

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: April 20, 2021

QUESTION 3

- This prior is similar to a Spike-and-slab prior but instead of using a Dirac function for promoting sparsity (with a particular probability) along any dimension, it uses 2 Gaussian distributions with different variances for a sparsity-inducing effect.
  The variance term in the prior distribution of $\mathbf{w}$ depends on the entries of $\boldsymbol{\gamma}$, a D dimensional binary vector. The entries 1 correspond to the directions with higher variance in the prior, while the entries 0 represent the directions with a lower variance (promoting w_d to be close to 0 apriori)

- **Computing CP**

$$\begin{aligned} \textit{Prior}: \quad & p(\mathbf{w} \mid \sigma, \boldsymbol{\gamma}) = \mathcal{N}(0, \sigma^2 \mathbf{K}), \qquad \textit{where} \;\; \mathbf{K} = \begin{bmatrix} \kappa_{\gamma_1} & \dots & \kappa_{\gamma_D} \end{bmatrix} \mathbf{I}_D \\ \textit{Likelihood}: \quad & p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N) \end{aligned}$$

Since the prior and likelihood are gaussian, we can directly use the properties of Linear Gaussian Models to obtain conditional posterior over $\mathbf{w}$ as:

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma, \boldsymbol{\gamma}) = \mathcal{N}(\mu_{\mathbf{N}}, \mathbf{\Sigma_N}) \quad & \textit{where} \;\; \mathbf{\Sigma} = \sigma^2 (\mathbf{K}^{-1} + \mathbf{X^T X})^{-1} \\ & \textit{and} \quad \mu_{\mathbf{N}} = (\sigma^2)^{-1} \mathbf{\Sigma X^T y} \;\; , \; \mathbf{\Sigma_N} = \mathbf{\Sigma} \end{aligned}$$

Computing CLL
The CLL is given as:

$$\begin{aligned} \log(\mathbf{y}, \mathbf{w} \mid \mathbf{X}, \boldsymbol{\gamma}, \theta, \sigma^2) &= \log(p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}, \sigma)) + \log(p(\mathbf{w} \mid \sigma, \boldsymbol{\gamma})) \\ &= \log(\mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N)) + \log(\mathcal{N}(0, \sigma^2 \mathbf{K})) \\ &= -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{N}{2} \log 2\pi\sigma^2 \\ &\qquad - \frac{1}{2\sigma^2} \mathbf{w}^T \mathbf{K}^{-1} \mathbf{w} - \frac{D}{2} \log 2\pi\sigma^2 - \frac{1}{2} \sum_{d=1}^{D} \log \kappa_{\gamma_d} \\ &= -\frac{1}{2\sigma^2} (\mathbf{y}^T \mathbf{y} - \mathbf{w}^T \mathbf{X}^T y - \mathbf{y}^T \mathbf{X}\mathbf{w}) - \frac{1}{2\sigma^2} \mathbf{w}^T (\mathbf{X}^T \mathbf{X} + \mathbf{K}^{-1}) \mathbf{w} \\ &\qquad - \frac{N+D}{2} \log 2\pi\sigma^2 - \frac{1}{2} \sum_{d=1}^{D} \log \kappa_{\gamma_d} \end{aligned}$$

E-step

Taking the expectation of the CLL calculated above wrt Conditional posterior of $\mathbf{w}$, we get:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{w}|\mathbf{y},\boldsymbol{\sigma},\boldsymbol{\gamma})}[\log(\mathbf{y},\mathbf{w}\mid\mathbf{X},\boldsymbol{\gamma},\theta,\sigma^2)] = &-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-\mathbb{E}[\mathbf{w}^T\mathbf{X}^Ty]-\mathbb{E}[\mathbf{y}^T\mathbf{X}\mathbf{w}])\\ &-\frac{1}{2\sigma^2}\mathbb{E}[\mathbf{w}^T(\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\mathbf{w}]\\ &-\frac{N+D}{2}\log 2\pi\sigma^2-\frac{1}{2}\sum_{d=1}^{D}\log\kappa_{\gamma_d} \qquad (4)\end{aligned}$$

Using results (374), (377), (378) of Matrix cookbook and the relation $\mathbb{E}[x^Ta]=\mathbb{E}[a^Tx]$, we get the following expectations:

$$\begin{aligned}\mathbb{E}[\mathbf{w}] &= \boldsymbol{\mu}_N\\ \mathbb{E}[\mathbf{w}\mathbf{w}^T] &= \boldsymbol{\Sigma}_N+\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T\\ \mathbb{E}[\mathbf{w}^T\mathbf{X}^Ty] &= \mathbb{E}[\mathbf{y}^T\mathbf{X}\mathbf{w}] = \mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}]\\ \mathbb{E}[\mathbf{w}^T(\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\mathbf{w}] &= Tr((\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\boldsymbol{\Sigma}_N)+\boldsymbol{\mu}_N^T(\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\boldsymbol{\mu}_N\end{aligned}$$

Using properties of trace [refer to link], the last relation can also be expressed as:

$$\mathbb{E}[\mathbf{w}^T(\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\mathbf{w}] = Tr\left(\ (\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\ \mathbb{E}[\mathbf{w}\mathbf{w}^T]\ \right)$$

On putting these values in (4), we get the final objective (Expected CLL) as:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{w}|\mathbf{y},\boldsymbol{\sigma},\boldsymbol{\gamma})}[\log(\mathbf{y},\mathbf{w}\mid\mathbf{X},\boldsymbol{\gamma},\theta,\sigma^2)] = &-\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y}-2\mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}]+Tr\left(\ (\mathbf{X}^T\mathbf{X}+\mathbf{K}^{-1})\ \mathbb{E}[\mathbf{w}\mathbf{w}^T]\ \right)\right]\\ &-\frac{N+D}{2}\log 2\pi\sigma^2-\frac{1}{2}\sum_{d=1}^{D}\log\kappa_{\gamma_d} \qquad (5)\end{aligned}$$

M-step

The objective we need wrt $\{\sigma^2,\boldsymbol{\gamma},\theta\}$ for the MAP solution is obtained by adding the logarithm of prior over $\{\sigma^2,\boldsymbol{\gamma},\theta\}$ to $\mathbb{E}$[CLL].

The prior $p(\sigma^2,\boldsymbol{\gamma},\theta)$ is given as:

$$\begin{aligned}p(\sigma^2,\boldsymbol{\gamma},\theta) &= p(\sigma^2|\boldsymbol{\gamma},\theta)p(\boldsymbol{\gamma}|\theta)p(\theta)\\ where, \qquad p(\sigma^2|\boldsymbol{\gamma},\theta) \ &= \ p(\sigma^2)\propto(\sigma^2)^{-(\frac{\nu}{2}+1)}\exp(-\frac{\nu\lambda}{2\sigma^2})\\ p(\boldsymbol{\gamma}|\theta) &\propto \prod_{d=1}^{D}\theta^{\gamma_d}(1-\theta)^{1-\gamma_d}\\ p(\theta) &\propto \theta^{a_o-1}(1-\theta)^{b_o-1}\end{aligned}$$

Let $\boldsymbol{\Theta}=\{\sigma^2,\boldsymbol{\gamma},\theta\}$. Maximizing the objective below wrt $\boldsymbol{\Theta}$ gives us the updates for the M-step.

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

Differentiating $\mathcal{L}_{MAP}$ wrt σ^2, we get:

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

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

Since γ_d is a discrete random variable with support $\{0,1\}$, we can directly substitute it in $\mathcal{L}_{MAP}$ and take the value for which $\mathcal{L}_{MAP}$ is larger. The terms in $\mathcal{L}_{MAP}$ that depend on γ_d are:

$$\mathcal{F}(\gamma_d) = -\frac{1}{2\sigma^2}\frac{\mathbb{E}[\mathbf{w}\mathbf{w}^T]_{d,d}}{\kappa_{\gamma_d}} - \frac{1}{2}\log\kappa_{\gamma_d} + \gamma_d\log\theta + (1-\gamma_d)\log(1-\theta)$$

$$\gamma_d = \max\{\mathcal{F}(0), \mathcal{F}(1)\}$$

Differentiating $\mathcal{L}_{MAP}$ wrt θ, we get:

$$\frac{(\sum_{d=1}^{D}\gamma_d) + (a_o - 1)}{\theta} - \frac{\sum_{d=1}^{D}(1-\gamma_d) + (b_o - 1)}{1-\theta} = 0$$

$$\implies \theta = \frac{(\sum_{d=1}^{D}\gamma_d) + a_o - 1}{D + a_o + b_o - 2}$$

EM Algorithm

1. Initialize $\boldsymbol{\Theta} = \{\sigma^{2^{(0)}}, \boldsymbol{\gamma}^{(0)}, \theta^{(0)}\}$. Set $t = 1$
2. **E-step:**
   Compute Conditional Posterior of w as:

$$p(\mathbf{w}^{(t)} \mid \mathbf{X}, \mathbf{y}, \sigma^{2^{(t-1)}}, \boldsymbol{\gamma}^{(t-1)}) = \mathcal{N}(\mu_{\mathbf{N}}^{(\mathbf{t})}, \mathbf{\Sigma}_{\mathbf{N}}^{(\mathbf{t})})$$

$$where\ \ \mu_{\mathbf{N}}^{(\mathbf{t})} = (\mathbf{K^{(t-1)}}^{-1} + \mathbf{X^TX})^{-1}\mathbf{X^Ty}$$

$$and\ \ \ \mathbf{\Sigma}_{\mathbf{N}}^{(\mathbf{t})} = \sigma^{2^{(t-1)}}(\mathbf{K^{(t-1)}}^{-1} + \mathbf{X^TX})^{-1}$$

   Update the expectations as:

$$\mathbb{E}[\mathbf{w}]^{(t)} = \boldsymbol{\mu}_N^{(t)}$$

$$\mathbb{E}[\mathbf{w}\mathbf{w}^T]^{(t)} = \mathbf{\Sigma}_N^{(t)} + \boldsymbol{\mu}_N^{(t)}\boldsymbol{\mu}_N^{T\,(t)}$$

3. **M-step:**
   Maximize $\mathcal{L}_{MAP}(\boldsymbol{\Theta})$ and update $\boldsymbol{\Theta}$ as:

$$\forall d \in [1, D],\ \gamma_d^{(t)} = \underset{\gamma_d\in\{0,1\}}{\arg\max} -\frac{1}{2\sigma^{2(t-1)}}\frac{\mathbb{E}[\mathbf{w}\mathbf{w}^T]_{d,d}^{(t)}}{\kappa_{\gamma_d}} - \frac{1}{2}\log\kappa_{\gamma_d} + \gamma_d\log\theta^{(t-1)} + (1-\gamma_d)\log(1-\theta^{(t-1)})$$

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

$$\theta^{(t)} = \frac{(\sum_{d=1}^{D}\gamma_d^{(t)}) + a_o - 1}{D + a_o + b_o - 2}$$

4. If not yet converged, set $t = t + 1$, and go to step 2.
5. Else return $\{\gamma^{(t)}, \sigma^{2^{(t)}}, \theta^{(t)}\}$ and $p(\mathbf{w}^{(t)} \mid \mathbf{X}, \mathbf{y}, \sigma^{2^{(t-1)}}, \boldsymbol{\gamma}^{(t-1)})$

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: April 20, 2021

QUESTION 4

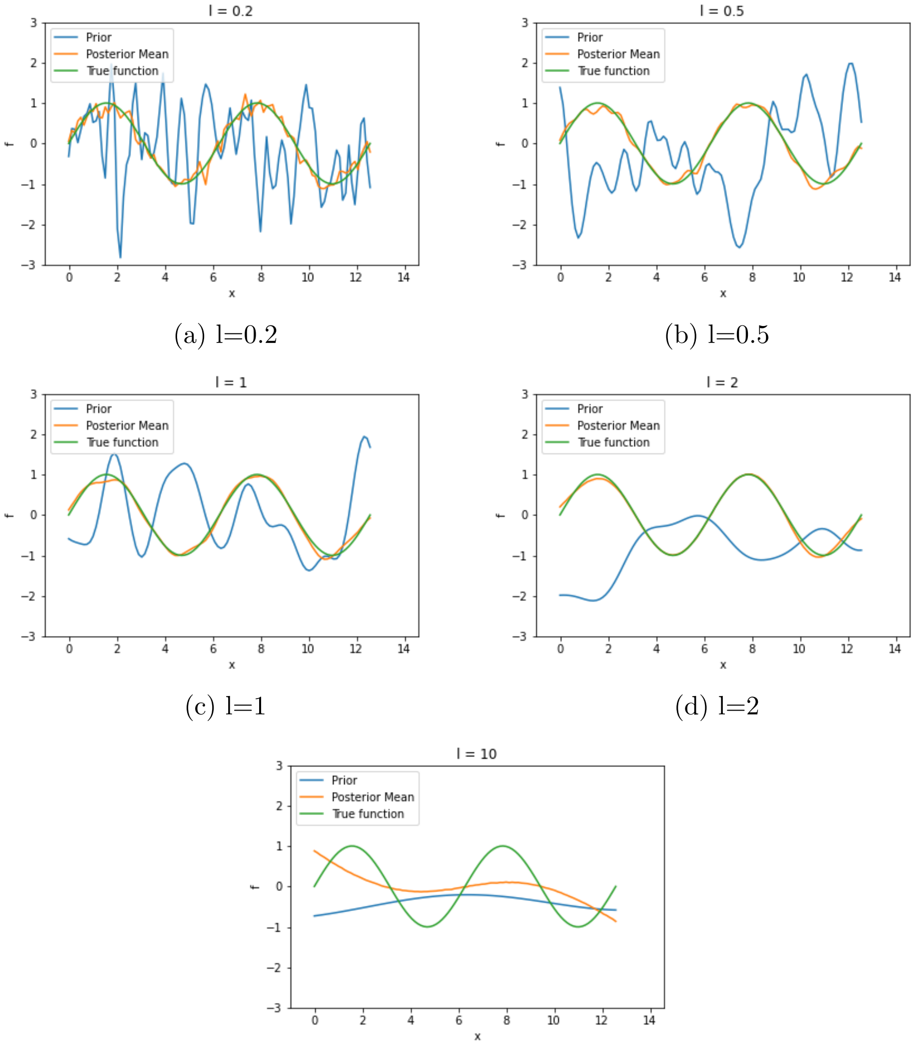
The prior and the likelihood(assuming all y_n are i.i.d.) for this probelm are given as :

$$p(\mathbf{f}) = \mathcal{N}(0, \mathbf{K})$$
$$p(\mathbf{y}|\mathbf{f}) = \prod_{n=1}^{N} \mathcal{N}(f(x_n), \sigma^2) = \mathcal{N}(\mathbf{f}(\mathbf{x}), \sigma^2\mathbf{I})$$

This can be expressed in the form of a Linear gaussian model as $y = Af + b + \epsilon$, where $\epsilon = \mathcal{N}(0, L)$ with $L = (\sigma^2)^{-1}\mathbf{I}$ and $A = I,\ b = 0$. The parameters of $\mathcal{N}(f|\mu, \Lambda)$ are $\mu = 0,\ \Lambda = \mathbf{K}^{-1}$.
Using the results of Linear gaussian model, we get the GP posterior as:

$$p(\mathbf{f}|\mathbf{y}) = \mathcal{N}\left((\sigma^2\mathbf{K}^{-1} + \mathbf{I})^{-1}\mathbf{y} \mid \sigma^2(\sigma^2\mathbf{K}^{-1} + \mathbf{I})^{-1}\right)$$

Observation regarding plots:
As l increases, the random samples from the prior distribution represent smoother(less noisy) curves.
For the mean of the posterior, as l increases, the noise reduces and the curve becomes more and more like the sin curve . The value $l = 2$ fits the sin curve best.
At large values of l (for $l = 10$ case), the mean of the posterior resembles the prior more than the data.



(a) l=0.2

(b) l=0.5

(c) l=1

(d) l=2

(e) l=10

Figure 1: Plots for Posterior, prior and true function with different values of l

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: April 20, 2021

QUESTION 5

1. The Total likelihood for the model is :

$$p(\mathbf{f}|\mathbf{X},\mathbf{Z},\mathbf{t}) = \prod_{n=1}^{N} p(\mathbf{f}_n|\mathbf{x}_n,\mathbf{Z},\mathbf{t}) = \prod_{n=1}^{N} \mathcal{N}\left(f_n|\tilde{\mathbf{k}}_n^T\tilde{\mathbf{K}}^{-1}\mathbf{t}, \kappa(\mathbf{x}_n,\mathbf{x}_n)) - \tilde{\mathbf{k}}_n^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_n\right)$$
$$= \mathcal{N}\left(\mathbf{f}|\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L\right),\ where$$
$$\boldsymbol{\mu}_L = \mathbf{G}\tilde{\mathbf{K}}^{-1}\mathbf{t};\quad Matrix \underset{N\times M}{\mathbf{G}}\ \ s.t.\ \ (\mathbf{G})_{nm} = \kappa(\mathbf{x}_n,\mathbf{z}_m)$$
$$\boldsymbol{\Sigma}_L = diagonal(S_1, S_2 \dots S_N);\quad S_n = \kappa(\mathbf{x}_n,\mathbf{x}_n)) - \tilde{\mathbf{k}}_n^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_n$$

The PPD whose expression we need to derive can be written as:

$$p(y_*|\mathbf{x}_*,\mathbf{f},\mathbf{X},\mathbf{Z}) = \int p(y_*|\mathbf{t},\mathbf{x}_*,\mathbf{f},\mathbf{X},\mathbf{Z})p(\mathbf{t}|\mathbf{f},\mathbf{X},\mathbf{Z})d\mathbf{t} \tag{6}$$

Thus we need to find the posterior wrt $\mathbf{t}$.
It is given that $\mathbf{Z},\mathbf{t}$ are modelled by the same gaussian process. Since it is a noiseless setting, we may assume $f_m = t_m$ for the m outputs. Thus, the prior over $\mathbf{t}$ will be: $p(\mathbf{t}|\mathbf{Z}) = \mathcal{N}(\mathbf{t}|\mathbf{0},\tilde{\mathbf{K}})$
The posterior is given by Bayes rule as:

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

Since the likelihood and priors are gaussians, we can directly derive the result for posterior using the properties of Gaussian Linear model as $\mathbf{f} = \mathbf{G}\tilde{\mathbf{K}}^{-1}\mathbf{t} + \epsilon$, where $\epsilon = \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_L)$:

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

$$\boldsymbol{\Sigma}_T = \left(\tilde{\mathbf{K}}^{-1} + \left((\mathbf{G}\tilde{\mathbf{K}}^{-1})^T diag(\frac{1}{S_1},\frac{1}{S_2}\dots\frac{1}{S_N})(\mathbf{G}\tilde{\mathbf{K}}^{-1})\right)\right)^{-1}$$
$$\boldsymbol{\mu}_T = \boldsymbol{\Sigma}_T\left((\mathbf{G}\tilde{\mathbf{K}}^{-1})^T diag(\frac{1}{S_1},\frac{1}{S_2}\dots\frac{1}{S_N})\mathbf{f}\right)$$

The expression for the first term in 6 is of the form $p(f_*|f)$ and thus, will always be a Gaussian:

$$p(y_*|\mathbf{t},\mathbf{x}_*,\mathbf{f},\mathbf{X},\mathbf{Z}) = \mathcal{N}\left(y_*|\tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\mathbf{t}, \kappa(\mathbf{x}_*,\mathbf{x}_*)) - \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*\right)$$

Since both the terms in Eq 6 are gaussians, we again use the results of a Linear Gaussian Model of the form $y* = \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\mathbf{t} + \epsilon$, where $\epsilon = \mathcal{N}(\mathbf{0}, \kappa(\mathbf{x}_*,\mathbf{x}_*)) - \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*)$ to solve for $p(y_*|\mathbf{x}_*,\mathbf{f},\mathbf{X},\mathbf{Z}) = \mathcal{N}(\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F)$ where,

$$\boldsymbol{\mu}_F = \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\boldsymbol{\mu}_T$$
$$\boldsymbol{\Sigma}_F = \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\boldsymbol{\Sigma}_T(\tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1})^T + \kappa(\mathbf{x}_*,\mathbf{x}_*)) - \tilde{\mathbf{k}}_*^T\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*$$

The computational cost of the above model is $O(M^3 + M^2N) \approx O(M^2N)$ as compared to $O(N^3)$ cost in case of normal Gaussian process. The main cost incurred during the whole process is computation of the inverse of $M \times M$ matrix $\tilde{\mathbf{K}}$ in $O(M^3)$ time and multiplication with an $N \times M$ matrix in $O(M^2N)$ time.
Thus this improved model is much more time efficient.

2. The expression for marginal likelihood for this model is given as:

$$p(\mathbf{f}|\mathbf{X},\mathbf{Z}) = \int p(\mathbf{f}|\mathbf{X},\mathbf{Z},\mathbf{t})p(\mathbf{t}|\mathbf{Z})d\mathbf{t} \tag{7}$$

The expressions for both the terms on RHS are gaussians and have been calculated in Part 1. Thus, we simply use the results of a Linear Gaussian Model as above:

$$\begin{aligned}
&\mathbf{f} = \mathbf{G}\tilde{\mathbf{K}}^{-1}\mathbf{t} + \epsilon,\ \ \textit{where}\quad \epsilon = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_L)\\
&\quad\therefore\ p(\mathbf{f}|\mathbf{X},\mathbf{Z}) = \mathcal{N}(\boldsymbol{\mu}_M, \mathbf{\Sigma}_M),\ \textit{where}\\
&\qquad\qquad\qquad\qquad \boldsymbol{\mu}_M = 0\\
&\qquad\qquad\qquad\qquad \mathbf{\Sigma}_M = \mathbf{G}\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{K}}(\mathbf{G}\tilde{\mathbf{K}}^{-1})^T + \mathbf{\Sigma}_L = \mathbf{G}(\tilde{\mathbf{K}}^{-1})^T\mathbf{G}^T + \mathbf{\Sigma}_L
\end{aligned}$$

The objective for MLE-II can be computed by taking log of the marginal likelihood derived above as:

$$\begin{aligned}
\hat{\mathbf{Z}} &= \arg\max_{\mathbf{Z}} \left(-\frac{1}{2}\mathbf{f}^T\mathbf{\Sigma}_M^{-1}\mathbf{f} - \frac{1}{2}\log(|\mathbf{\Sigma}_M|) - \frac{N}{2}\log(2\pi)\right)\\
&= \arg\min_{\mathbf{Z}} \left(\mathbf{f}^T\mathbf{\Sigma}_M^{-1}\mathbf{f} + \log(|\mathbf{\Sigma}_M|)\right)
\end{aligned}$$

Solving this objective, we get the MLE-II estimate for $\mathbf{Z}$.