Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

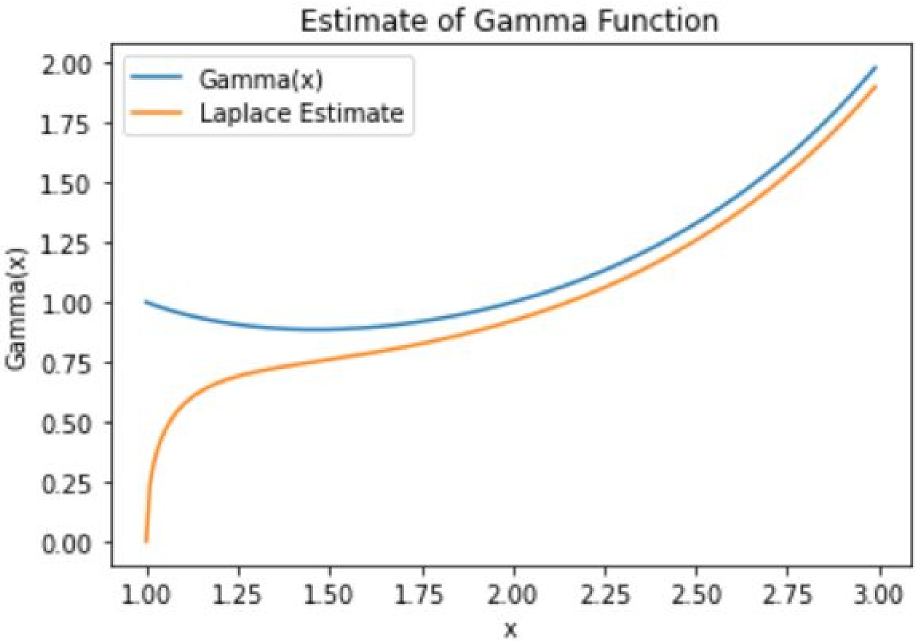
Student Name: Dishay Mehta
Roll Number: 200341
Date: March 30, 2023

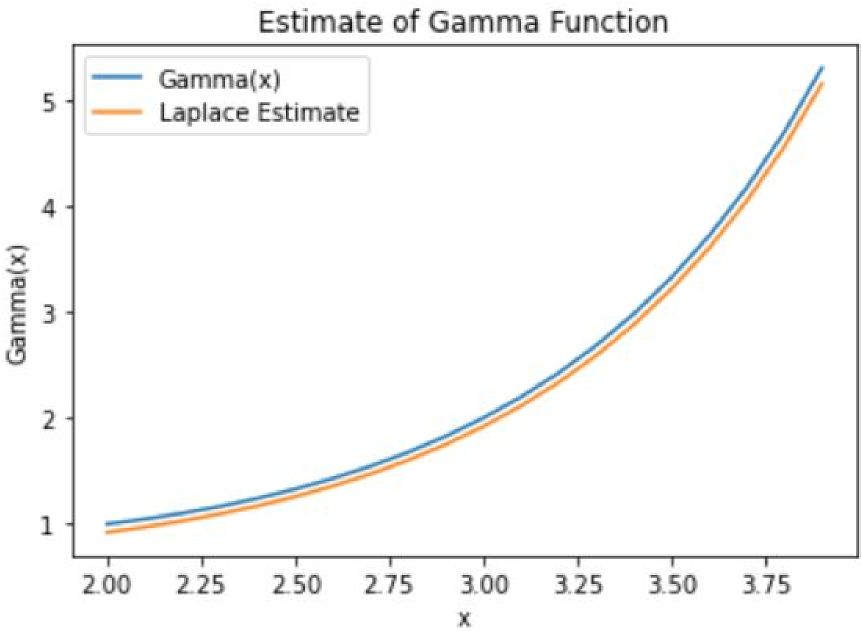
---

a)Gamma$(x|a,b)$=$\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$
Let us approximate Gamma distribution to a Gaussian using Laplace's Approximation.
Gamma$(x|a,b) \approx \mathcal{N}(x_{map}, H^{-1})$ where x_{map} is the value of mode and H is the Hessian matrix.
From slides,
x_{map}=$\underset{x}{\text{argmax}}$ Gamma$(x|a,b)$
=$\underset{x}{\text{argmax}}\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$
Differentiation wrt x and equating to 0,
$0 = (a-1)x^{a-2}e^{-bx} + x^{a-1}(-b)e^{-bx}$
$\implies x_{map} = \frac{a-1}{b}$
$H = -\nabla^2 \text{log Gamma}(x|a,b)\Big|_{x=\frac{a-1}{b}}$
$=-\nabla^2(\text{a log b - log }\Gamma(a) + \text{(a-1)log x - bx})\Big|_{x=\frac{a-1}{b}}$
$=-\nabla(\frac{a-1}{x} - b)\Big|_{x=\frac{a-1}{b}}$
$=(\dfrac{a-1}{x^2})\Big|_{x=\frac{a-1}{b}}$
$H = \frac{b^2}{a-1}$
$\implies H^{-1} = \frac{a-1}{b^2}$

Thus, Gamma$(x|a,b) \approx \mathcal{N}(\frac{a-1}{b}, \frac{a-1}{b^2})$
From the prob stats refresher slides,
Mean of Gamma distribution Gamma$(x|a,b) = \frac{a}{b}$
Variance of Gamma distribution Gamma$(x|a,b) = \frac{a}{b^2}$
Thus approximating Gamma$(x|a,b)$ by a Gaussian whose mean and variance are equal to the mean and variance of Gamma$(x|a,b)$ is Gamma$(x|a,b) \approx \mathcal{N}(\frac{a}{b}, \frac{a}{b^2})$
Thus $\mathcal{N}(\frac{a-1}{b}, \frac{a-1}{b^2})$ and $\mathcal{N}(\frac{a}{b}, \frac{a}{b^2})$ will behave same if a tends to ∞, since
$\frac{a-1}{b} \approx \frac{a}{b}$ and $\frac{a-1}{b^2} \approx \frac{a}{b^2}$
Hence, when a is large, then the approximation will be the same.
b) Usign Laplaces Approximation, we get
$\mathcal{N}(\frac{a-1}{b}, \frac{a-1}{b^2}) \approx \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$
$\Gamma(a) \approx \dfrac{b^a x^{a-1}e^{-bx}}{\mathcal{N}(\frac{a-1}{b}, \frac{a-1}{b^2})}$
Solving this gets us,
$\Gamma(a) = \sqrt{2\pi(a-1)}b^{a-1}x^{a-1}e^{-\frac{(bx-(a-1))^2}{2(a-1)} - bx}$
I have put the value of x_{map} from above since the gaussian and gamma are maximum at x_{map} and we are approximating both, hence we equate their max values.
Putting the value of $x = \frac{a-1}{b}$, we get

$\Gamma(a) = \sqrt{2\pi(a-1)}(\frac{a-1}{e})^{a-1}$





This is the plot of $\Gamma(a)$ using a standard library function (blue) and the approximated form (orange).
The first plot has values of x bounded by [1,3] and a step of 0.01 and the second plot has values of x bounded by [2,4] and a step of 0.1.
As we can see here the Laplace approximation is very close to the actual plot made by the some standard library function, indicating that laplace approximation is **highly accurate** in nature.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

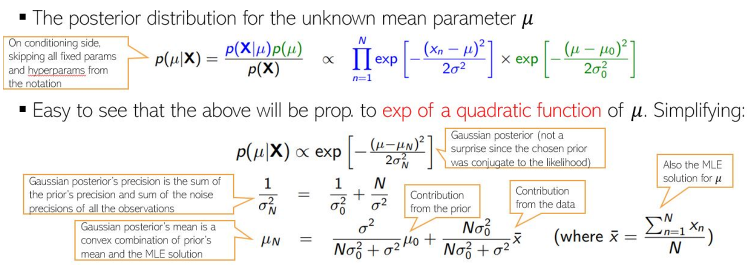
Student Name: Dishay Mehta
Roll Number: 200341
Date: March 30, 2023

QUESTION 2

We have N scalar iid observations, $x_1, x_2, x_3 ... x_N$
$\mu \sim \mathcal{N}(\mu|\mu_0, s_0)$
$\beta \sim \text{Gamma}(\beta|a, b)$
Finding the conditional posterior, (Assuming $p(\mathbf{X})$ to be a constant $\frac{1}{A}$)

- $p(\mu|\mathbf{X}, \beta) = \frac{p(\mathbf{X}|\mu,\beta)p(\mu)}{p(X)}$

$=A \prod_{n=1}^{N} e^{-\frac{1}{2}\beta(x_n-\mu)^2} e^{-\frac{(\mu-\mu_0)^2}{2s_0}}$
$=Ae^{-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}}$
Using the results from slides,



here the expressions for σ_N^2 and μ_N are

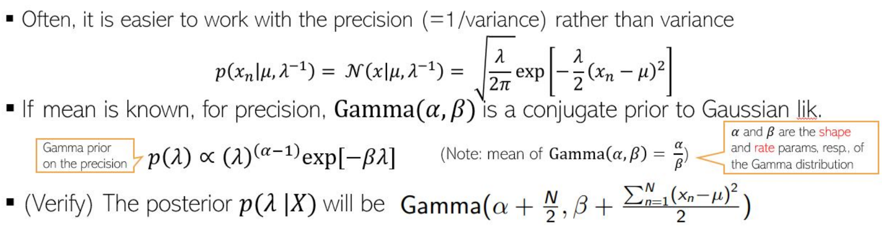
$\frac{1}{\sigma_N^2} = N\beta + \frac{1}{s_0}$

$\mu_N = \frac{1}{N\beta s_0+1}\mu_0 + \frac{N\beta s_0}{N\beta s_0+1}\bar{x}$ where$(\bar{x} = \frac{\Sigma_{n=1}^{N} x_n}{N})$
Thus,

$$\boxed{p(\mu|\mathbf{X}, \beta)=\mathcal{N}(\mu_N, \sigma_N^2)}$$

- $p(\beta|\mathbf{X}, \mu) = \frac{p(\mathbf{X}|\mu,\beta)p(\beta)}{p(X)}$

$=A \prod_{n=1}^{N} e^{-\frac{1}{2}\beta(x_n-\mu)^2} \frac{b^a}{\Gamma(a)} \beta^{a-1} e^{-b\beta}$
Using the results from slides,



Thus,

$$\boxed{p(\beta|\mathbf{X},\mu) = \text{Gamma}(\beta|a+\tfrac{N}{2}, b+\frac{\Sigma_{n=1}^{N}(x_n-\mu)^2}{2})}$$

Gibbs sampling can be used to get a sampling-based approximation of a multiparameter posterior. Gibbs sampler iteratively draws random samples from CPs. When run long enough, the sampler produces samples from the joint posterior.

This is the case of two parameter (μ, β), the Gibbs Sampler looks like this
1) Initialize $\beta^{(0)}$
2) For s=1,2,3....S

- Draw a random sample for μ as $\mu^{(s)} \sim p(\mu|\mathbf{X}, \beta^{(s-1)})$
- Draw a random sample for μ as $\beta^{(s)} \sim p(\beta|\mathbf{X}, \mu^{(s)})$

These S samples $(\mu^s, \beta^s)_{s=1}^{S}$ represent the joint posterior $p(\mu, \beta|\mathbf{X})$

Student Name: Dishay Mehta
Roll Number: 200341
Date: March 30, 2023

QUESTION 3

We have the following distributions,

$$\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta \sim \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{x}_n, \mathbf{w}, \beta^{-1})$$

$$\mathbf{w}|\lambda \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)$$

$$\mathbf{w}|\mathbf{y}, \mathbf{X}, \lambda, \beta \sim \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\mu}$ and Σ are defined as follows
$\Sigma = (\beta\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_D)^{-1}$ $\qquad$ $\boldsymbol{\mu} = (\mathbf{X}^\intercal\mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\intercal\mathbf{y}$
The model parameter is λ and β so it will be a global variable that can be calculated by maximizing $\mathbb{E}[CLL]$ in the M step. $\mathbf{w}$ is a latent variable whose conditional parameter will be computed in the E step. Therefore, we will use EM Algorithm to alternatively model the parameters and the hyperparameters. With EM, can treat $\mathbf{w}$ as "latent var", and λ, β as "parameters".

EM algorithm
This is the whole algorithm with the working also shown, the algorithmic sketch is enclosed inside the box

Step 1:

Initialize λ as λ_0 and β as β_0 and set t=1

Step 2: (Expectation step)
• Compute the posterior of $\mathbf{w}_t$ given the current parameters, λ_{t-1} and β_{t-1}

$p(\mathbf{w}_t|\mathbf{y}, \mathbf{X}, \lambda_{t-1}, \beta_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$
$\Sigma_{t-1} = (\beta_{t-1}\mathbf{X}^\intercal\mathbf{X} + \lambda_{t-1}\mathbf{I}_D)^{-1}$
$\boldsymbol{\mu}_{t-1} = (\mathbf{X}^\intercal\mathbf{X} + \frac{\lambda_{t-1}}{\beta_{t-1}}\mathbf{I}_D)^{-1}\mathbf{X}^\intercal\mathbf{y}$

• Now we will calculate the complete log likelihood - log(p($\mathbf{y}, \mathbf{w}|\mathbf{X}, \beta, \lambda$))
Using chain rule of probability,
$p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \beta, \lambda)$=$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\lambda)$
$\therefore$ log(p($\mathbf{y}, \mathbf{w}|\mathbf{X}, \beta, \lambda$))=log(p($\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta$))+log(p($\mathbf{w}|\lambda$))
$\implies$ CLL $= \frac{1}{2}[Nlog\beta + Dlog\lambda - \beta(\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal(\mathbf{y} - \mathbf{X}\mathbf{w}) - \lambda\mathbf{w}^\intercal\mathbf{w} - (N + D)log2\pi]$

Since the posterior of $\mathbf{w}$ is a normal distribution, we can directly write $\mathbb{E}[\mathbf{w}] = \boldsymbol{\mu}$

• Now we will compute the expectations as follows
$\mathbb{E}[\mathbf{w}^\intercal] = \mathbb{E}[\mathbf{w}]^\intercal = \boldsymbol{\mu}^\intercal$
$\mathbb{E}[\mathbf{w}\mathbf{w}^\intercal] = Cov(w) + \mathbb{E}[\mathbf{w}]\,\mathbb{E}[\mathbf{w}]^\intercal = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$
$\mathbb{E}[\mathbf{w}^\intercal R\mathbf{w}] = \mathrm{Tr}(\mathrm{R}\ \mathbb{E}[\mathbf{w}\mathbf{w}^\intercal]) = Tr(R(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal))$

• Now we will calculate the expectation of CLL

Using the results from above,
$\mathbb{E}[CLL] = \frac{1}{2}[Nlog\beta_{t-1} + Dlog\lambda_{t-1} - \beta_{t-1}(\mathbf{y}^\intercal\mathbf{y} - \mathbb{E}[\mathbf{w}^\intercal]\mathbf{X}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\,\mathbb{E}[\mathbf{w}] + \mathbb{E}[\mathbf{w}^\intercal(\mathbf{X}^\intercal\mathbf{X} + \frac{\lambda_{t-1}}{\beta_{t-1}}\mathbf{I}_D)\mathbf{w}]) -$ const]

$$\boxed{\mathbb{E}[CLL] = \frac{1}{2}[Nlog\beta_{t-1} + Dlog\lambda_{t-1} - \beta_{t-1}(\mathbf{y}^\intercal\mathbf{y} - \boldsymbol{\mu}_{t-1}^\intercal\mathbf{X}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\boldsymbol{\mu}_{t-1} + Tr(\mathbf{X}^\intercal\mathbf{X}(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal))) - \lambda_{t-1}Tr(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal) - \text{const}]}$$

Step 3: Maximization step
Now, we can do MLE for the parameters by maximising the $\mathbb{E}[CLL]$ as in the standard EM algorithm.
$(\lambda_t, \beta_t) = \underset{\lambda,\beta}{\operatorname{argmax}}\, \mathbb{E}[CLL]$
$= \underset{\lambda,\beta}{\operatorname{argmax}} \frac{1}{2}[Nlog\beta_{t-1} + Dlog\lambda_{t-1} - \beta_{t-1}(\mathbf{y}^\intercal\mathbf{y} - \boldsymbol{\mu}_{t-1}^\intercal\mathbf{X}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\boldsymbol{\mu}_{t-1} + Tr(\mathbf{X}^\intercal\mathbf{X}(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal))) - \lambda_{t-1}Tr(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal) -$ const]

We need the point estimate of β_t and λ_t
MLE Estimate of β
We simply require to maximize the Expectation of CLL wrt β

$$\boxed{\beta_t^{-1} = \frac{1}{N}(\mathbf{y}^\intercal\mathbf{y} - \boldsymbol{\mu}_{t-1}^\intercal\mathbf{X}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\boldsymbol{\mu}_{t-1} + Tr(\mathbf{X}^\intercal\mathbf{X}(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal)))}$$

MLE Estimate of λ
We simply require to maximize the Exceptation of CLL wrt λ

$$\boxed{\lambda_t^{-1} = \frac{1}{D}(Tr(\Sigma_{t-1} + \boldsymbol{\mu}_{t-1}\boldsymbol{\mu}_{t-1}^\intercal))}$$

where D is the dimension of the vector space

Step 4:

$\boxed{\text{If } \lambda \text{ and } \beta \text{ are not yet converged then set t=t+1 and repeat from step 2.}}$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 4

Student Name: Dishay Mehta
Roll Number: 200341
Date: March 30, 2023

We have the following distributions,

$$y_n|z_n \sim \mathbb{I}(z_n > 0)$$

$$z_n|\mathbf{w}, \mathbf{x}_n \sim \mathcal{N}(z_n|\mathbf{w}^\intercal\mathbf{x}_n, 1) \implies \mathbf{z}|\mathbf{w}, \mathbf{X} \sim \mathcal{N}(\mathbf{z}|\mathbf{X}\mathbf{w}, \mathbf{I}_N)$$

The model parameter is $\mathbf{w}$ so it will be a global variable that can be calculated by maximizing the $\mathbb{E}[CLL]$ in the M step. z_n is a latent variable whose conditional posterior will be computed in the E step. Therefore, we will use EM Algorithm to alternatively model the parameters and the hyperparameters. With EM, can treat z_n as "latent var", and $\mathbf{w}$ as "parameter".

EM algorithm
This is the whole algorithm with the working also shown, the algorithmic sketch is enclosed inside the box

Step 1:
Initialize $\mathbf{w}$ as $\mathbf{w}^0$ and set t=1

Step 2: (Expectation step)
• Compute the posterior of z_n^t given the current parameters, $\mathbf{w}^{t-1}$

$p(z_n|x_n, y_n, \mathbf{w}) \propto p(z_n|x_n, \mathbf{w})p(y_n|z_n) \propto$ prior*likelihood

$$p(z_n^t|\mathbf{w}^{t-1}, y_n, \mathbf{x}_n) = \begin{cases} \mathcal{N}(z_n^t|\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n, 1)\, \mathbb{I}(z_n^t > 0) & \text{if } y_n = 1 \\ \mathcal{N}(z_n^t|\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n, 1)\, \mathbb{I}(z_n^t < 0) & \text{if } y_n = 0 \end{cases}$$

Since we constrained the variable of gaussian to be within certain limits, it is called as **Truncated Gaussian.**

• Now we will calculate the complete log likelihood
$log(p(\mathbf{z}^t, \mathbf{y}|\mathbf{X}, \mathbf{w}^{t-1}) = log(p(\mathbf{y}|\mathbf{z}^t)) + log(p(\mathbf{z}^t|\mathbf{X}, \mathbf{w}^{t-1}))$
$=\sum_{n=1}^{N} log(p(y_n|z_n^t)) - \frac{1}{2}(\mathbf{z}^t - \mathbf{X}\mathbf{w}^{t-1})^\intercal(\mathbf{z}^t - \mathbf{X}\mathbf{w}^{t-1})$ + const

• Now we will calculate the expectation of CLL
Since CLL depends linearly on $\mathbf{z}^t$, we just used to calculate $\mathbb{E}[z_n^t|\mathbf{w}^{t-1}, y_n, \mathbf{x}_n]$ to calculate $\mathbb{E}[CLL]$. Since $p(z_n^t|\mathbf{w}^{t-1}, y_n, \mathbf{x}_n)$ is a truncated gaussian, so we just need to compute the posterior mean as $\mathbf{w}^{t-1}$ and z_n^t share a linear relation and hence we need to compute the expectation of $z_n^t|\mathbf{x}_n, y_n, \mathbf{w}^{t-1}$

$$\mathbb{E}[\mathbf{z}_n^t|\mathbf{w}^{t-1}, y_n, \mathbf{x}_n] = \begin{cases} \mathbf{w}^{(t-1)\intercal}\mathbf{x}_n + \dfrac{\phi(\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n)}{1 - \Phi(-\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n)} & \text{if } y_n = 1 \\ \mathbf{w}^{(t-1)\intercal}\mathbf{x}_n + \dfrac{\phi(\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n)}{\Phi(-\mathbf{w}^{(t-1)\intercal}\mathbf{x}_n)} & \text{if } y_n = 0 \end{cases}$$

where $\phi(.)$ is the standard normal probability distribution function and $\Phi(.)$ is the Probit function, which is given on the wikipedia link given in the assignment.

• Now we calculate the expected complete log likelihood

$\mathbb{E}[CLL]$ = const - $\mathbb{E}[\frac{(\mathbf{z}^t-\mathbf{X}\mathbf{w}^{t-1})^\intercal(\mathbf{z}^t-\mathbf{X}\mathbf{w}^{t-1})}{2}]$ (the const term is indep of $\mathbf{w}^{t-1}$)

- Expected CLL in EM is given by (assume observations are i.i.d.)

$$\mathcal{Q}(\Theta, \Theta^{old}) = \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)]$$
$$= \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta) + \log p(\boldsymbol{z}_n|\Theta)]$$

- If $p(\mathbf{z}_n|\Theta)$ and $p(\boldsymbol{x}_n|\mathbf{z}_n, \Theta)$ are exp-family distributions, $\mathcal{Q}(\Theta, \Theta^{\mathrm{old}})$ has a very simple form
- In resulting expressions, replace terms containing $\boldsymbol{z}_n$'s by their respective expectations, e.g.,
  - $\boldsymbol{z}_n$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\hat{\Theta})}[\boldsymbol{z}_n]$
  - $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\hat{\Theta})}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$

Thus from the above results,

$$\boxed{\mathbb{E}[CLL] = \text{const - } \frac{||\,\mathbb{E}[\mathbf{z}^t] - \mathbf{X}\mathbf{w}^{t-1}||^2}{2}}$$

Step 3: Maximization Step

Now, we can do MLE for the parameters by maximising the $\mathbb{E}[CLL]$ as in the standard EM algorithm.

$$\mathbf{w}^t = \underset{\mathbf{w}}{\operatorname{argmax}}\, \mathbb{E}[CLL]$$
$$= \underset{\mathbf{w}}{\operatorname{argmax}}(const - \frac{||\,\mathbb{E}[\mathbf{z}^t] - \mathbf{X}\mathbf{w}^{t-1}||^2}{2})$$

MLE Estimate of w

We require to maximize the Expectation of CLL wrt $\mathbf{w}$

$$\boxed{\mathbf{w}^t = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\,\mathbb{E}[\mathbf{z}^t]}$$

where $\mathbb{E}[\mathbf{z}^t] = \Bigg[\mathbb{E}[z_1^t], \mathbb{E}[z_2^t], ... \,\mathbb{E}[z_N^t]\Bigg]^\intercal$

Step 4:

$\boxed{\text{If } \mathbf{w}^t \text{ are not yet converged then set t=t+1 and repeat from step 2.}}$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Dishay Mehta
Roll Number: 200341
Date: March 30, 2023

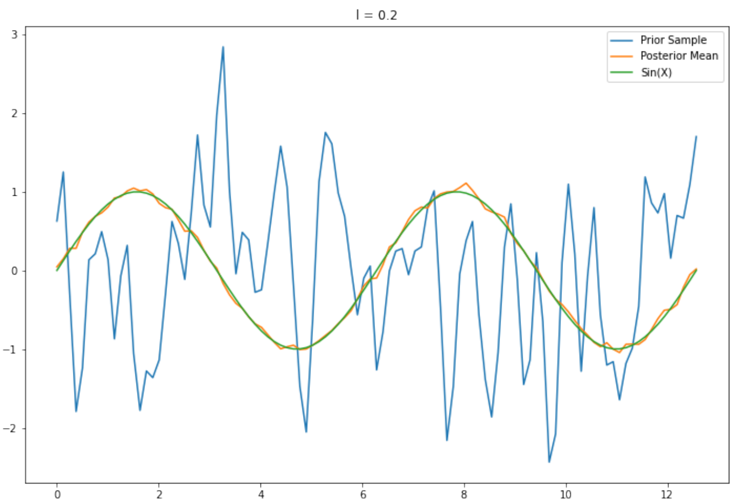
QUESTION 5

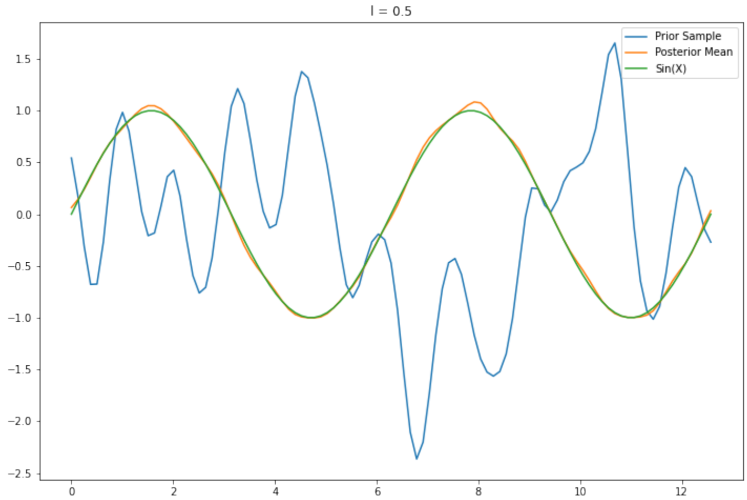
Part 1

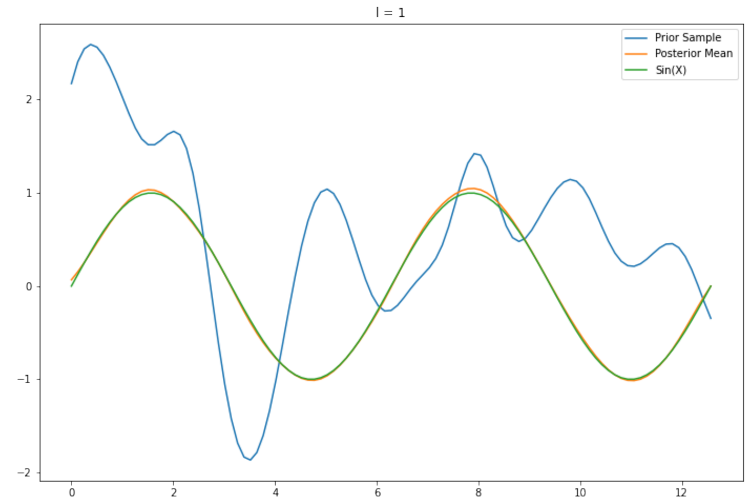
Given:
p($\mathbf{f}$)=GP($0, \boldsymbol{\kappa}$)=$\mathcal{N}(0, \boldsymbol{\kappa})$
p($y_n|x_n, f) = \mathcal{N}(y_n|f(x_n, \sigma^2)$
For posterior,
p($\mathbf{f}|\mathbf{y}) = \dfrac{p(\mathbf{f})p(\mathbf{y}|\mathbf{f})}{(\mathbf{y})}$
here $p(\mathbf{y})$ is independent of $\mathbf{f}$, so the posterior becomes,
p($\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f})p(\mathbf{y}|\mathbf{f})$
$\propto \mathcal{N}(0, \boldsymbol{\kappa}) \prod_{n=1}^{N} \mathcal{N}(y_n|f(x_n), \sigma^2)$
$\propto \mathcal{N}(0, \boldsymbol{\kappa})\mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}_N)$
$p(\mathbf{f}|\mathbf{y}) \propto exp[-\frac{1}{2\sigma^2}(\mathbf{f}-\mathbf{y})^\intercal(\mathbf{f}-\mathbf{y}) - \frac{1}{2}\mathbf{f}^\intercal\boldsymbol{\kappa}^{-1}\mathbf{f}] \therefore p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\mu, \boldsymbol{\Sigma})$
where $\boldsymbol{\Sigma} = \sigma^2(\sigma^2\mathbf{I}_N + \boldsymbol{\kappa})^{-1}\boldsymbol{\kappa}$ and $\mu = \boldsymbol{\kappa}(\sigma^2\mathbf{I}_N + \boldsymbol{\kappa})^{-1}\mathbf{y}$

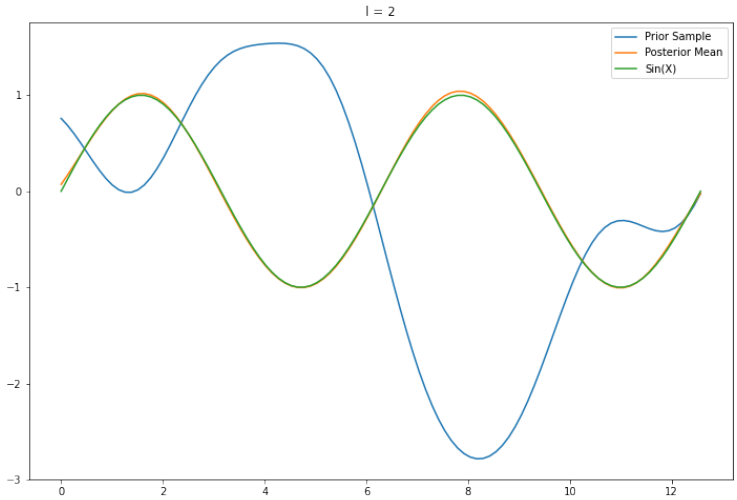
Part 2

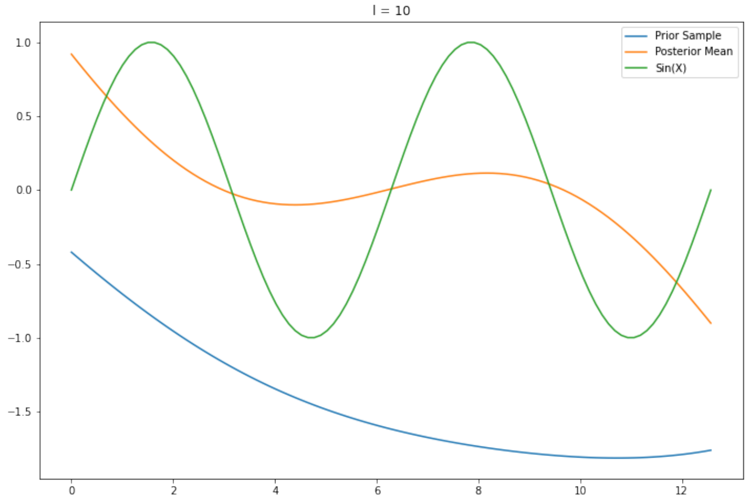
Visualising the GP Priors and Posteriors for Regression




l = 0.5
Prior Sample
Posterior Mean
Sin(X)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0
2
4
6
8
10
12


l = 1
Prior Sample
Posterior Mean
Sin(X)
2
1
0
−1
−2
0
2
4
6
8
10
12


l = 2
Prior Sample
Posterior Mean
Sin(X)
1
0
−1
−2
−3
0
2
4
6
8
10
12


l = 10
Prior Sample
Posterior Mean
Sin(X)
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
2
4
6
8
10
12

Inference:

The difference between the plots generated using different values of the parameter l is mainly in the smoothness and amplitude of the GP prior and posterior functions, as well as in the uncertainty estimates. Here are some specific observations:

- For smaller values of l, the GP functions tend to be wiggly and have higher frequency variations and for larger values of l, the GP functions tend to be smoother and have lower frequency variations.
- The GP posterior functions tend to be smoother than the GP prior functions because they are conditioned on the observed data and thus have reduced uncertainty.
- As l increases the number of peaks and valleys in prior decreases. The posterior however is very close to the true function as we increase l from 0.2 to 2 but then for l=10, there is a significant difference between the posterior mean and true function, since the kernel becomes large as l increases.