Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 1

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: March 30, 2023

Given the p.d.f $Gamma(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$, its Laplace approximation will be $\mathcal{N}(x_{MAP}, H^{-1})$
$x_{MAP} = arg\,max_x Gamma(x|a,b)$
Let's take the derivative of the pdf w.r.t x
$\Rightarrow G' = (a-1-bx)e^{-bx}x^{a-2}$
Therefore $x_{MAP} = \frac{a-1}{b}$
$\Rightarrow H = -\nabla^2 log(Gamma(x|a,b))$
$\Rightarrow -\nabla^2(alog(b) - log(\Gamma(a)) + (a-1)log(x) - bx)$
$\Rightarrow H = \frac{a-1}{x^2}$
Plugging in the value of x_{MAP}:
$\Rightarrow H^{-1} = \frac{a-1}{b^2}$
Therefore, the Laplace approximation is

$$Gamma(x|a,b) \approx \mathcal{N}\left(x|\frac{a-1}{b}, \frac{a-1}{b^2}\right)$$

Using the results given in the prop-stats refresher slides, for a Gamma distribution, its $mean = k\theta$ and $variance = k\theta^2$. Here $\theta = \frac{1}{b}$ and $k = a$. So $mean = \frac{a}{b}$ and $variance = \frac{a}{b^2}$.
Therefore the Gaussian whose mean and variance are equal to the mean and variance, respectively, of $Gamma(x|a,b)$ is

$$Gamma(x|a,b) \approx \mathcal{N}\left(x|\frac{a}{b}, \frac{a}{b^2}\right)$$

The only difference between the two approximations is the term $\frac{1}{b}$ and $\frac{1}{b^2}$ for mean and variance respectively. This difference will tend to zero as b increases, hence taking a large value of b will ensure that the two approximations are same.

Now using our Laplace approximation of $Gamma(x|a,b)$, we cget the following equation:
$\Rightarrow \mathcal{N}\left(x|\frac{a-1}{b}, \frac{a-1}{b^2}\right) \approx \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$
$\Rightarrow \Gamma(a) \approx \frac{b^a x^{a-1} e^{-bx}}{\mathcal{N}\left(x|\frac{a-1}{b}, \frac{a-1}{b^2}\right)}$
Now, plugging in the value of x_{MAP} in the above equation, we get:

$$\Rightarrow \Gamma(a) \approx \sqrt{2\pi(a-1)}\left(\frac{a-1}{e}\right)^{a-1}$$

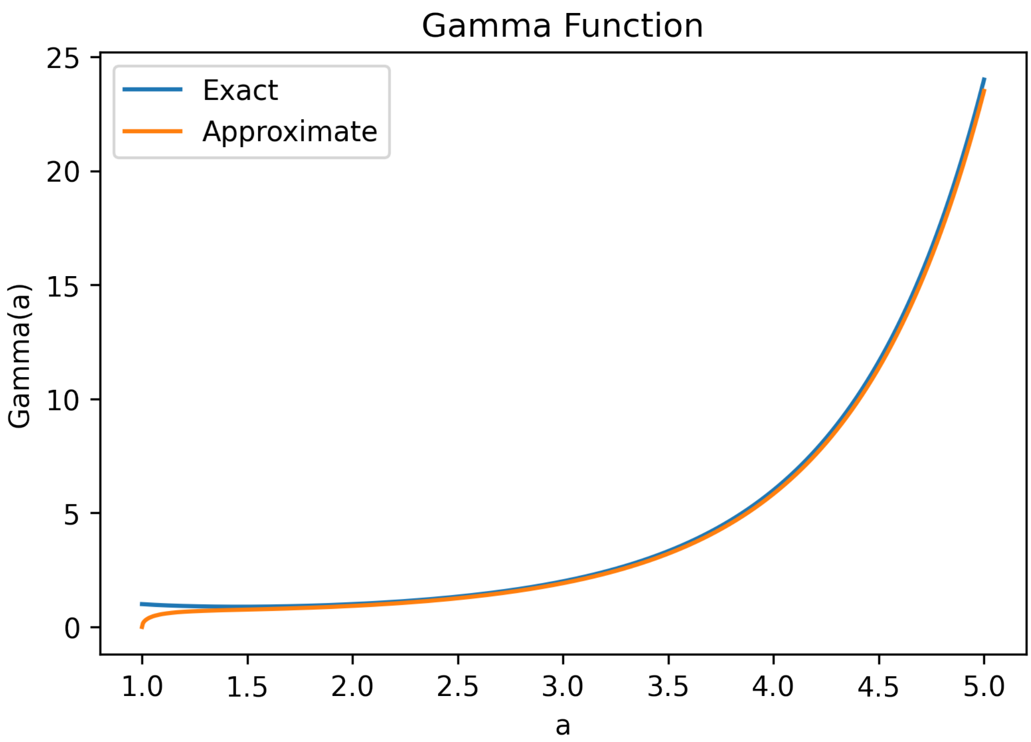

Figure 1: Approximation plots of $\Gamma(a)$

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: March 30, 2023

QUESTION 2

Let $\mathbf{X} = [x_1, x_2, \ldots, x_N]^T$
$\Rightarrow p(\mu|\mathbf{X}, \beta) = \frac{p(\mathbf{X}|\mu,\beta)p(\mu)}{\int p(\mathbf{X}|\mu,\beta)p(\mu)d\mu} \propto \Pi_{n=1}^{N} exp\left[-\frac{\beta(x_n-\mu)^2}{2}\right] \times exp\left[-\frac{(\mu-\mu_0)^2}{2s_0}\right]$

Now using the results from [1]. $p(\mu|\mathbf{X}, \beta) \propto exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$
where μ_N and σ_N are as folows:

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

$$\Rightarrow \mu_N = \frac{1}{N\beta s_0 + 1}\mu_0 + \frac{N\beta s_0}{N\beta s_0 + 1}\bar{x} \qquad \left(\text{where } \bar{x} = \frac{\Sigma_{n=1}^{N} x_n}{N}\right)$$

Therfore,

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

$\Rightarrow p(\beta|\mathbf{X}, \mu) = \frac{p(\mathbf{X}|\mu,\beta)p(\beta)}{\int p(\mathbf{X}|\mu,\beta)p(\beta)d\beta} \propto Gamma(\beta|a, b)\,\mathcal{N}(\mathbf{X}|\mu\mathbf{I}, \beta^{-1}\mathbf{I})$
Using the results from [2], we get:

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

Gibbs Sampling algorithm

- Initialize $\beta^{(0)}$
- For $s = 1, 2, \ldots, S$
  - $\mu^{(s)} \sim p(\mu|\mathbf{X}, \beta^{(s-1)})$
  - $\beta^{(s)} \sim p(\beta|\mathbf{X}, \mu^{(s)})$

After running the above algorithm for a large number of iterations, $(\mu^{(s)}, \beta^{(s)})_{s=1}^{S}$ will give us the joint posterior of μ and β.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: March 30, 2023

QUESTION 3

EM algorithm

- **E step:**

$$p(\mathbf{w}^{(t)}|\mathbf{X},\mathbf{y},\lambda^{(t-1)},\beta^{(t-1)}) = \frac{p(\mathbf{w}^{(t-1)}|\lambda^{(t-1)})p(\mathbf{y}|\mathbf{X},\mathbf{w}^{(t)},\beta^{(t-1)})}{p(\mathbf{y},\mathbf{X},\lambda^{(t-1)},\beta^{(t-1)})}$$

- **M step:**

$$\{\lambda^{(t)},\beta^{(t)}\} = arg\,max_{\lambda,\beta}\mathbb{E}[log\,p(\mathbf{y},\mathbf{w}^{(t)}|\mathbf{X},\beta,\lambda)]$$

Let's use the result from [3] to find the posterior of $\mathbf{w}$ as follows:

$$p(\mathbf{w}^{(t)}|\mathbf{X},\mathbf{y},\lambda^{(t-1)},\beta^{(t-1)}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D)\times\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\beta^{-1}\mathbf{I}_N)$$

Therefore, the estimated E step will be:

$$p(\mathbf{w}^{(t)}|\mathbf{X},\mathbf{y},\lambda^{(t-1)},\beta^{(t-1)}) = \mathcal{N}(\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)$$

where $\boldsymbol{\Sigma}_N = (\beta^{(t-1)}\mathbf{X}^T\mathbf{X}+\lambda^{(t-1)}\mathbf{I}_D)^{-1}$ and $\boldsymbol{\mu}_N = (\mathbf{X}^T\mathbf{X}+\frac{\lambda^{(t-1)}}{\beta^{(t-1)}}\mathbf{I}_D)^{-1}\mathbf{X}^T\mathbf{y}$

The complete data log likelihood is:
$CLL = log(p(\mathbf{y},\mathbf{w}|\mathbf{X},\beta,\lambda)) = log(p(\mathbf{y}|\mathbf{X},\mathbf{w},\beta)) + log(p(\mathbf{w}|\lambda))$
$\Rightarrow \frac{1}{2}(Nlog\beta - \beta(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) - Nlog(2\pi) + Dlog\lambda - \lambda\mathbf{w}^T\mathbf{w} - Dlog(2\pi))$
Expected value of this CLL is:
$\mathbb{E}[CLL] = \frac{1}{2}(Nlog\beta - \beta(\mathbf{y}^T\mathbf{y} - \mathbb{E}[\mathbf{w}^T]\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbb{E}[\mathbf{w}] + \mathbb{E}[\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}]) + Dlog\lambda - \lambda\mathbb{E}[\mathbf{w}^T\mathbf{w}] - (N+D)log(2\pi))$
Let's first find the expectations that are need to compute the above expression:
$\Rightarrow \mathbb{E}[\mathbf{w}] = E[\mathbf{w}^T] = \boldsymbol{\mu}_N$
$\Rightarrow \mathbb{E}[\mathbf{w}\mathbf{w}^T] = Cov(\mathbf{w}) + \mathbb{E}[\mathbf{w}][\mathbf{w}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T$
$\Rightarrow \mathbb{E}[\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}] = Tr(\mathbf{X}^T\mathbf{X}\,\mathbb{E}[\mathbf{w}\mathbf{w}^T]) = Tr(\mathbf{X}^T\mathbf{X}(\boldsymbol{\Sigma}+\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T))$
Therefore, $\mathbb{E}[CLL] = \frac{1}{2}(Nlog\,\beta - \beta(\mathbf{y}^T\mathbf{y} - \boldsymbol{\mu}_N\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\mu}_N + Tr(\mathbf{X}^T\mathbf{X}(\boldsymbol{\Sigma}+\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T)) + Dlog\,\lambda - (N+D)log(2\pi)))$
Now, $\lambda = arg\,max_\lambda \quad \mathbb{E}[CLL]$
After taking derivative w.r.t λ, we get:

$$\lambda^{(t)} = \frac{D}{Tr(\mathbf{X}^T\mathbf{X}(\boldsymbol{\Sigma}+\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T))}$$

and similarly we get:

$$\beta^{(t)} = \frac{N}{\mathbf{y}^T\mathbf{y} - \boldsymbol{\mu}_N\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\mu}_N + Tr(\mathbf{X}^T\mathbf{X}(\boldsymbol{\Sigma}+\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T))}$$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: March 30, 2023

QUESTION 4

Let's first find the CP of latent variables:
$p(z_n^{(t)}|y_n, \mathbf{w}^{(t-1)}, \mathbf{x}_n) \propto p(z_n^{(t)}|\mathbf{w}^{(t-1)}, \mathbf{x}_n) \times p(y_n|z_n^{(t)}, \mathbf{w}^{(t-1)}, \mathbf{x}_n)$
We know that $p(z_n^{(t)}|\mathbf{w}^{(t-1)} = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, 1)$ and $p(y_n|z_n^{(t)}, \mathbf{w}^{(t-1)}, \mathbf{x}_n) = \mathbb{I}(z_n^{(t)} > 0)$ hence

Conditional Posterior

$$\begin{aligned} p(z_n^{(t)}|y_n, \mathbf{w}^{(t-1)}, \mathbf{x}_n) = (\mathcal{N}(z_n^t|\mathbf{w}^{(t-1)T}\mathbf{x}_n, 1)\mathbb{I}(z_n^t > 0))^{y_n} \\ \times(\mathcal{N}(z_n^t|\mathbf{w}^{(t-1)T}\mathbf{x}_n, 1)\mathbb{I}(z_n^t < 0))^{(1-y_n)} \end{aligned}$$

The terms inside $(.)^{y_n}$ and $(.)^{(1-y_n)}$ are truncated normal distribution.

Let's now calculate the CLL:
$log(p(\mathbf{z}^t, \mathbf{y}|\mathbf{X}, \mathbf{w}^{(t-1)})) = log(p(\mathbf{y}|\mathbf{z}^{(t)})) + log(p(\mathbf{z}^{(t)}|\mathbf{X}, \mathbf{w}^{(t-1)}))$
$\Rightarrow CLL = \Sigma_{n=1}^N log(p(y_n|z_n^{(t)})) - \frac{1}{2}(\mathbf{z}^{(t)} - \mathbf{X}\mathbf{w}^{(t-1)})^T(\mathbf{z}^{(t)} - \mathbf{X}\mathbf{w}^{(t-1)}) + const$
After taking expectation and replacing z_n with $\mathbb{E}[z_n]$, we get:
$\mathbb{E}[CLL] = const - \frac{Det.(\mathbb{E}[\mathbf{z}^{(t)}] - \mathbf{X}\mathbf{w}^{(t-1)})^2}{2}$ where $Det(A)$ is the determinant of the matrix A.
Now let's find the maximum value of this expected value of CLL w.r.t w:
$\mathbf{w}^{(t)} = arg\,max_{\mathbf{w}} \quad \Sigma_{n=1}^N \mathbb{E}_{p(z_n^{(t)}|y_n, \mathbf{x}_n, \mathbf{w}^{(t-1)})}[log(p(y_n, z_n^{(t)}|\mathbf{w}))]$
$\Rightarrow arg\,max_{\mathbf{w}} \quad -\frac{1}{2}Det.(\mathbb{E}[\mathbf{z}^{(t)}] - \mathbf{X}\mathbf{w}^{(t-1)})^2$, therefore

Maximizing CLL

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

where $\mathbb{E}[\mathbf{z}^{(t)}] = \left[\mathbb{E}[z_1^{(t)}], \mathbb{E}[z_2^{(t)}], \ldots, \mathbb{E}[z_N^{(t)}]\right]^T$

EM algorithm

- Initialize $\mathbf{w}^{(0)}$
- For $t = 1, 2, \ldots, T$, until convergence do:
  - **E step:** Compute N CPs
    $p(z_n^{(t)}|y_n, \mathbf{x}_n, \mathbf{w}^{(t-1)}) = \left(\frac{\mathbb{I}[z_n > 0]}{1 - \mathbf{\Phi}_n}\right)^{y_n} \left(\frac{\mathbb{I}[z_n > 0]}{\mathbf{\Phi}_n}\right)^{(1-y_n)} \mathcal{N}(z_n^{(t)}|\mathbf{w}^{(t-1)T}\mathbf{x}_n, 1)$
    Compute $E[z_n]$ as well
  - **M step:** Compute $\mathbf{w}$
    $\mathbf{w}^{(t)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{z}^{(t)}]$
    where $\mathbb{E}[\mathbf{z}^{(t)}] = \left[\mathbb{E}[z_1^{(t)}], \mathbb{E}[z_2^{(t)}], \ldots, \mathbb{E}[z_N^{(t)}]\right]^T$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Abhishek Pardhi
Roll Number: 200026
Date: March 30, 2023

QUESTION 5

Part 1:
$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{f})p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}) \times \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}_N)$
Using the results of [4], we get the posterior:

$$p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\mathbf{\Sigma}\frac{\mathbf{y}}{\sigma^2}, \mathbf{\Sigma})$$

where $\mathbf{\Sigma} = (\mathbf{K}^{-1} + \frac{\mathbf{I}_N}{\sigma^2})^{-1}$
Part 2:
It could be seen that smaller value of l (such as $l = 0.2$) leads to over-fitting and larger values of l (such as $l = 10$) leads to under-fitting. However, choosing the value that is in between them ($l = 2$) gives us the best estimate of the true function.

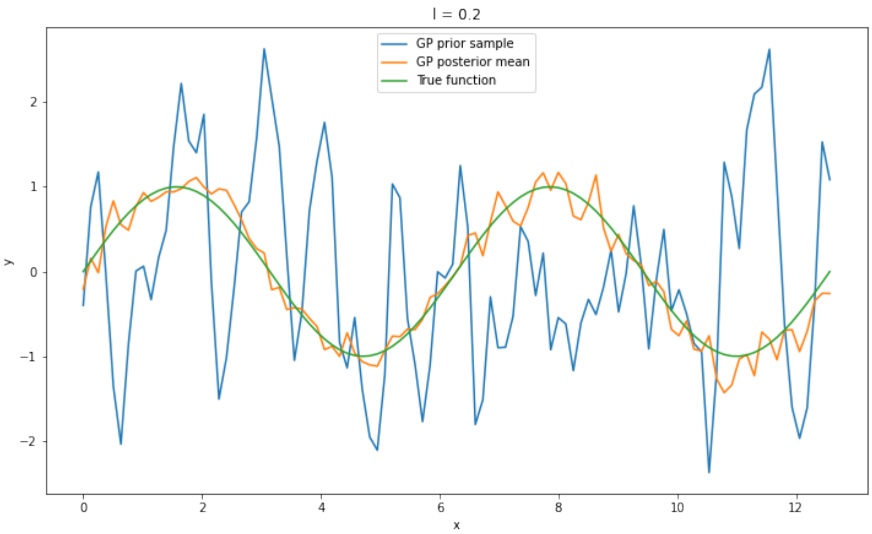

Figure 2: Distributions for $l = 0.2$

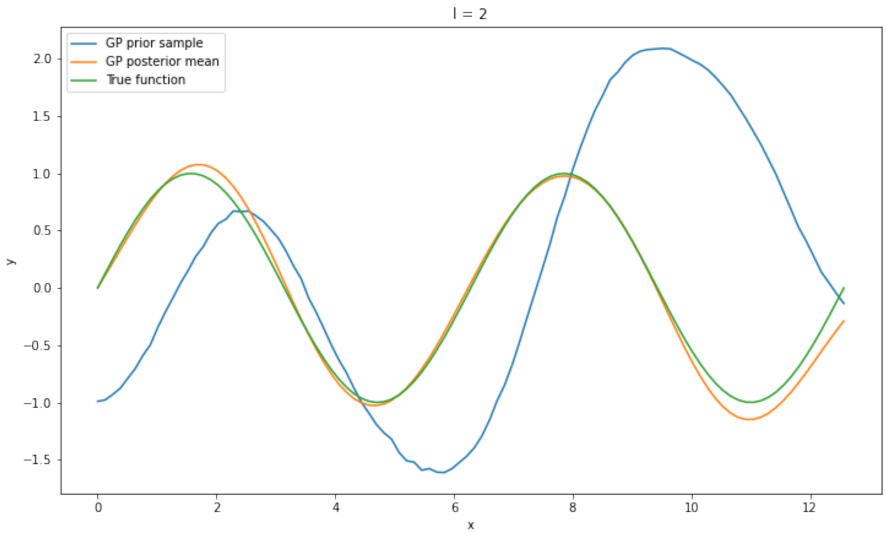

Figure 3: Distributions for $l = 2$

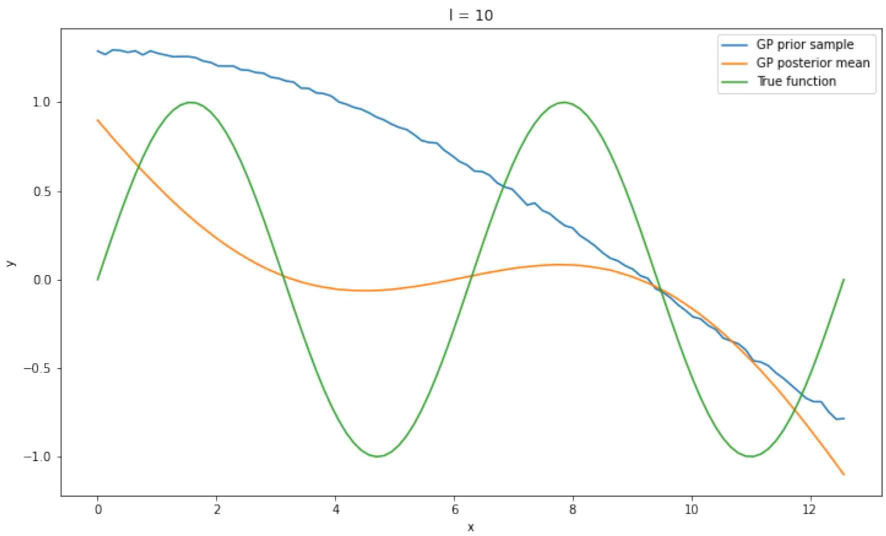

Figure 4: Distributions for $l = 10$

References

[1] Piyush Rai. *CS772 Lecture*. Slide 4, Page 13.

[2] Piyush Rai. *CS772 Lecture*. Slide 4, Page 16.

[3] Piyush Rai. *CS772 Lecture*. Slide 5, Page 15.

[4] Piyush Rai. *CS772 Lecture*. Slide 5, Page 11.