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

Student Name: Atreya Goswami
Roll Number: 190201
Date: March 30, 2023

QUESTION

1

Gaussian-ify the Gamma!

(a) Given the gamma distribution, whose p.d.f. is given by

$$p\left(x \mid a, b\right) = \mathsf{Gamma}\left(x \mid a, b\right) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

where, a and b are the shape and rate parameters respectively, we first need to approximate this distribution using the Laplace approximation.
Laplace approximation approximates any distribution $p\left(\theta\right)$ using the Gaussian distribution $\mathcal{N}(\theta \mid \hat{\theta}, \mathbf{H}^{-1})$, where $\hat{\theta}$ is the mode of the distribution $p\left(\theta\right)$ to be approximated, and $\mathbf{H}$ is the Hessian matrix of the negative log p.d.f. evaluated at $\hat{\theta}$.

Mode of the gamma distribution[1] $x_{\text{mode}} = \frac{a-1}{b}$ (assuming $a \geq 1$).

Since x is a scalar R.V., $\mathbf{H}$ will be the second derivative of the negative log p.d.f. evaluated at x_{mode}.

$$\begin{aligned}
\mathbf{H} &= \frac{\partial^2}{\partial x^2}\left[-\log p\left(x \mid a, b\right)\right]\Bigg|_{x=x_{\text{mode}}} \\
&= -\frac{\partial^2}{\partial x^2}\left[a \log b - \log \Gamma(a) + (a-1)\log x - bx\right]\Bigg|_{x=x_{\text{mode}}} \\
&= \frac{a-1}{x^2}\Bigg|_{x=x_{\text{mode}}} \\
&= \frac{b^2}{a-1}
\end{aligned}$$

Hence, using the Laplace approximation, the gamma distribution can be approximated using a gaussian distribution as

$$\mathsf{Gamma}\left(x \mid a, b\right) \approx \mathcal{N}\left(x \mid x_{\text{mode}}, \mathbf{H}^{-1}\right) = \mathcal{N}\left(x \;\middle|\; \frac{a-1}{b}, \frac{a-1}{b^2}\right) = q\left(x \mid a, b\right)$$

Mean of gamma distribution[1] is given by $\mu = \frac{a}{b}$.
Variance of gamma distribution[1] is given by $\sigma^2 = \frac{a}{b^2}$

Now, we need to approximate the gamma distribution using a gaussian distribution with the

[1] Expressions of mean, mode and variance of the gamma distribution were taken from the following Wikipedia entry: https://en.wikipedia.org/wiki/Gamma_distribution

same mean and variance.
Thus the appropriate gaussian distribution is $\mathcal{N}\left(x \mid \frac{a}{b}, \frac{a}{b^2}\right)$

The two approximations will be roughly same for $b \gg a$.

(b) In order to approximate the gamma function $\Gamma(a)$, let us evaluate the p.d.f of the true distribution and the gaussian p.d.f. derived using Laplace approximation at the mode $x_{\text{mode}} = \frac{a-1}{b}$.

$$p\left(x = x_{\text{mode}} \mid a, b\right) = \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{b}\right)^{a-1} e^{-(a-1)} = \frac{b}{\Gamma(a)} \left(a-1\right)^{a-1} e^{-(a-1)}$$

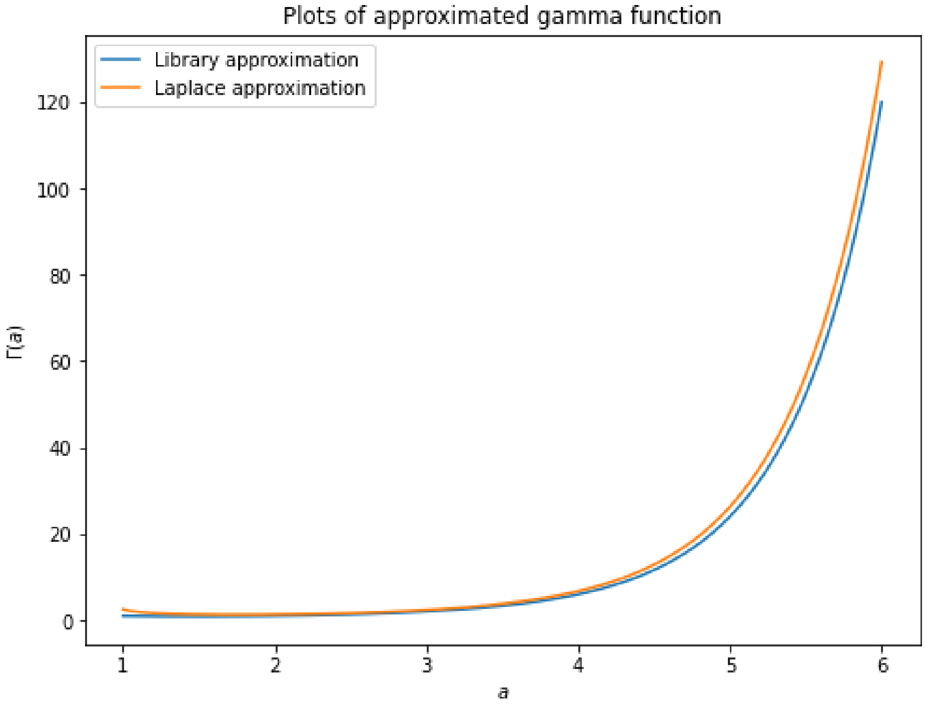
$$q\left(x = x_{\text{mode}} \mid a, b\right) = \frac{b}{\sqrt{2\pi(a-1)}}$$

Now, since $p\left(x = x_{\text{mode}} \mid a, b\right) \approx q\left(x = x_{\text{mode}} \mid a, b\right)$,

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

Hence for $a \geq 1$, the gamma function $\Gamma(a)$ can be approximated as

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



The gamma function was plotted using the Laplace approximation derived above and using the Scipy library function for the same. Our approximation is fairly accurate, and most closely resembles the library approximation at $a = 2.3$. The two curves differ slightly at $a \approx 1$ and for large values of a.

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

Student Name: Atreya Goswami
Roll Number: 190201
Date: March 30, 2023

QUESTION
2

Local Conjugacy and Gibbs Sampling

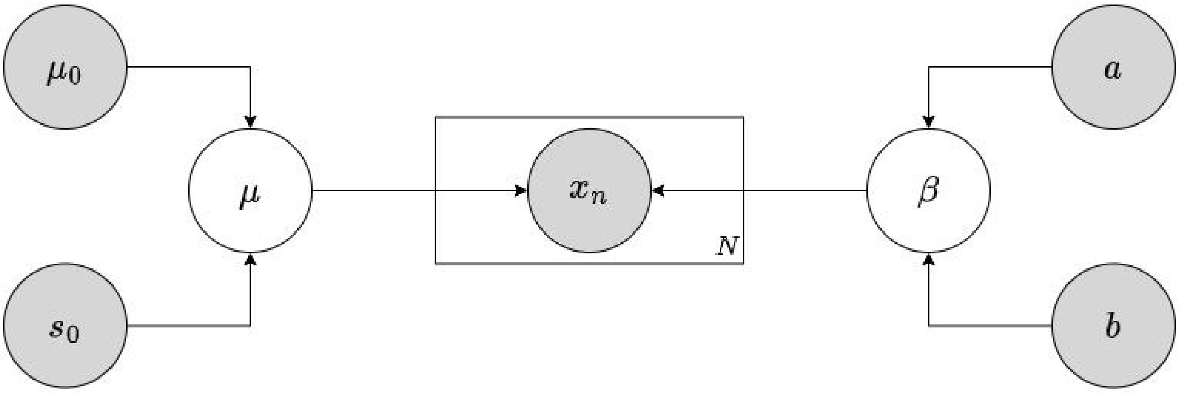
Given N scalar observations

$$x_1, x_2, \cdots, x_N \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(x \,|\, \mu, \beta^{-1}\right)$$

Let us define $\mathbf{X}$ as x_n's stacked vertically for all $n \in [1, 2, \cdots, N]$. We can write the overall likelihood function as

$$p\left(\mathbf{X} \,|\, \mu, \beta\right) = \prod_{n=1}^{N} p\left(x_n \,|\, \mu, \beta\right) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \,|\, \mu, \beta^{-1}\right)$$

We assume a gaussian prior $p\left(\mu \,|\, \mu_0, s_0\right) = \mathcal{N}\left(\mu \,|\, \mu_0, s_0\right)$ on the mean μ and a gamma prior $p\left(\beta \,|\, a, b\right) = \text{Gamma}\left(\beta \,|\, a, b\right)$ on the precision β. The corresponding plate notation can be diagrammatically represented as follows.



Now, considering the Markov blanket property, we can write the conditional posteriors of μ and β as $p\left(\mu \,|\, \mathbf{X}, \beta, \mu_0, s_0\right)$ and $p\left(\beta \,|\, \mathbf{X}, \mu, a, b\right)$ respectively. To derive the required CPs, we use the technique introduced on Slide 13 of Lecture 18, i.e., we first write down the joint posterior of μ and β and then use terms containing μ and β only to get their corresponding CPs.

$$\begin{aligned} p\left(\mu, \beta \,|\, \mathbf{X}, \mu_0, s_0, a, b\right) &= p\left(\mathbf{X} \,|\, \mu, \beta\right)\, p\left(\mu \,|\, \mu_0, s_0\right)\, p\left(\beta \,|\, a, b\right) \\ &= \prod_{n=1}^{N} \mathcal{N}\left(x_n \,|\, \mu, \beta^{-1}\right)\, \mathcal{N}\left(\mu \,|\, \mu_0, s_0\right)\, \text{Gamma}\left(\beta \,|\, a, b\right) \end{aligned}$$

Hence,

$$p\left(\mu \,|\, \mathbf{X}, \beta, \mu_0, s_0\right) \propto \exp\left[-\sum_{n=1}^{N} \frac{\beta}{2}\left(x_n - \mu\right)^2\right] \times \exp\left[-\frac{\left(\mu - \mu_0\right)^2}{2 s_0}\right]$$

Using local conjugacy and the results presented on Slide 13 of Lecture 4, we can write

$$\boxed{p\left(\mu \,|\, \mathbf{X}, \beta, \mu_0, s_0\right) = \mathcal{N}\left(\mu \,|\, \mu_N, s_N\right)}$$

where,

$$\begin{aligned} s_N &= \left(\frac{1}{s_0} + N\beta\right)^{-1} \\ \mu_N &= \left(\frac{1}{1 + N\beta s_0}\right)\mu_0 + \left(\frac{\beta s_0}{1 + N\beta s_0}\right)\sum_{n=1}^{N} x_n \end{aligned}$$

Similarly,

$$p\left(\beta \,|\, \mathbf{X}, \mu, a, b\right) \propto \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\beta \sum_{n=1}^{N} \frac{\left(x_n - \mu\right)^2}{2}\right] \times \beta^{a-1} \exp[-\beta b]$$

Using local conjugacy and the results presented on Slide 16 of Lecture 4, we can write

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

Now, we use the Gibbs sampling technique to obtain a sampling-based approximation of the joint posterior of μ and β, i.e., $p\left(\mu, \beta \,|\, \mathbf{X}, \mu_0, s_0, a, b\right)$.

Gibbs Sampling algorithm to approximate the joint posterior $p\left(\mu, \beta \,|\, \mathbf{X}, \mu_0, s_0, a, b\right)$

Gibbs sampler iteratively draws random samples from CPs of μ and β. When run long enough, the sampler produces samples from the joint posterior.

1. Initialise $\beta^{(0)}$
2. For $s = 1, 2, 3, \cdots, S$, repeat
   (a) Draw a random sample for μ as
   $\mu^{(s)} \sim p\left(\mu \,|\, \mathbf{X}, \beta^{(s-1)}, \mu_0, s_0\right) = \mathcal{N}\left(\mu \,|\, \mu_N^{(s)}, s_N^{(s)}\right)$ where,

$$\begin{aligned} s_N^{(s)} &= \left(\frac{1}{s_0} + N\beta^{(s-1)}\right)^{-1} \\ \mu_N^{(s)} &= \left(\frac{1}{1 + N\beta^{(s-1)} s_0}\right)\mu_0 + \left(\frac{\beta^{(s-1)} s_0}{1 + N\beta^{(s-1)} s_0}\right)\sum_{n=1}^{N} x_n \end{aligned}$$

   (b) Draw a random sample for β as
   $\beta^{(s)} \sim p\left(\beta \,|\, \mathbf{X}, \mu^{(s)}, a, b\right) = \text{Gamma}\left(\beta \,\middle|\, a + \frac{N}{2}, b + \sum_{n=1}^{N} \frac{\left(x_n - \mu^{(s)}\right)^2}{2}\right)$

These S random samples $\left\{\left(\mu^{(s)}, \beta^{(s)}\right)\right\}_{s=1}^{S}$ represent a sampling-based approximation of the joint posterior.

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

QUESTION 3

Student Name: Atreya Goswami
Roll Number: 190201
Date: March 30, 2023

Given the probabilistic linear regression model $p\left(y_n \mid \boldsymbol{w}, \boldsymbol{x}_n, \beta\right) = \mathcal{N}\left(y_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}\right)$, where $\boldsymbol{x}_n, \boldsymbol{w} \in \mathbb{R}^D$, and a gaussian prior $p\left(\boldsymbol{w} \mid \lambda\right) = \mathcal{N}\left(\boldsymbol{w} \mid 0, \lambda^{-1}\mathbf{I}\right)$ on $\boldsymbol{w}$, we will use the EM algorithm to estimate the hyperparameters λ and β. Let us define $\mathbf{X}$ and $\boldsymbol{y}$ as $\boldsymbol{x}_n$ and y_n stacked vertically for all $n \in [1, 2, 3, \cdots, N]$. Hence assuming i.i.d observations, we can write the overall likelihood as

$$p\left(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{w}, \beta\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}\right) = \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}\right)$$

In order to apply the EM algorithm, we consider λ and β as parameters for which point estimates will suffice, and $\boldsymbol{w}$ as the latent variable for which we will compute the conditional posterior.

E step

In the E step, we need to compute the CP of the latent variable i.e., $\boldsymbol{w}$, and expectation of the complete-data log likelihood w.r.t. the CP of $\boldsymbol{w}$ given current values of the parameters $\lambda^{(\text{old})}$ and $\beta^{(\text{old})}$.

CP of $\boldsymbol{w}$ is given by

$$\begin{aligned} p\left(\boldsymbol{w} \mid \mathbf{X}, \boldsymbol{y}, \beta^{(\text{old})}, \lambda^{(\text{old})}\right) &= \frac{p\left(\boldsymbol{w} \mid \lambda^{(\text{old})}\right)\, p\left(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{w}, \beta^{(\text{old})}\right)}{p\left(\boldsymbol{y} \mid \mathbf{X}, \beta^{(\text{old})}\right)} \\ &\propto \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{X}\boldsymbol{w}, \left(\beta^{(\text{old})}\right)^{-1}\mathbf{I}\right) \mathcal{N}\left(\boldsymbol{w} \mid 0, \left(\lambda^{(\text{old})}\right)^{-1}\mathbf{I}\right) \end{aligned}$$

Using the results of probabilistic linear regression presented on Slide 15 of Lecture 5, we can write the CP of $\boldsymbol{w}$ as

$$p\left(\boldsymbol{w} \mid \mathbf{X}, \boldsymbol{y}, \beta^{(\text{old})}, \lambda^{(\text{old})}\right) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right)$$

where,

$$\text{cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_N = \left(\beta^{(\text{old})}\mathbf{X}^\top\mathbf{X} + \lambda^{(\text{old})}\mathbf{I}\right)^{-1} \tag{3.1}$$

$$\mathbb{E}[\boldsymbol{w}] = \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left[\beta^{(\text{old})}\mathbf{X}^\top\boldsymbol{y}\right] = \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda^{(\text{old})}}{\beta^{(\text{old})}}\mathbf{I}\right)^{-1}\mathbf{X}^\top\boldsymbol{y} \tag{3.2}$$

The complete-data log likelihood (CLL) is given by

$$
\begin{aligned}
\log p\left(\boldsymbol{w}, \boldsymbol{y} \,|\, \mathbf{X}, \beta, \lambda\right) = & \quad \log\left[p\left(\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}, \beta, \lambda\right) \times p\left(\boldsymbol{w} \,|\, \lambda\right)\right] \\
&= \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}\right) + \log \mathcal{N}\left(\boldsymbol{w} \,|\, 0, \lambda^{-1}\mathbf{I}\right) \\
&= \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2 - \frac{1}{2}\log|\lambda^{-1}\mathbf{I}| - \frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} + C \\
&= \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}\left[y_n^2 - 2y_n\boldsymbol{w}^\top\boldsymbol{x}_n + \boldsymbol{x}_n^\top\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{x}_n\right] - \frac{1}{2}\log|\lambda^{-1}\mathbf{I}| \\
&\qquad - \frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} + C
\end{aligned}
$$

Hence by linearity of expectation, the expected CLL $\mathcal{L}\left(\lambda, \beta, \lambda^{(\text{old})}, \beta^{(\text{old})}\right)$ is given by

$$
\begin{aligned}
\mathcal{L}\left(\lambda, \beta, \lambda^{(\text{old})}, \beta^{(\text{old})}\right) &= \mathbb{E}_{p\left(\boldsymbol{w} \,|\, \mathbf{X}, \boldsymbol{y}, \beta^{(\text{old})}, \lambda^{(\text{old})}\right)}\left[\log p\left(\boldsymbol{w}, \boldsymbol{y} \,|\, \mathbf{X}, \beta, \lambda\right)\right] \\
= \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}\left[y_n^2 - 2y_n\,\mathbb{E}\left[\boldsymbol{w}^\top\right]\boldsymbol{x}_n + \boldsymbol{x}_n^\top\,\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\top\right]\boldsymbol{x}_n\right] &- \frac{1}{2}\log|\lambda^{-1}\mathbf{I}| - \frac{\lambda}{2}\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] + C
\end{aligned}
\tag{3.3}
$$

where $\mathbb{E}[\cdot]$ is computed w.r.t. $p\left(\boldsymbol{w} \,|\, \mathbf{X}, \boldsymbol{y}, \beta^{(\text{old})}, \lambda^{(\text{old})}\right)$, the CP of $\boldsymbol{w}$ given the current values of the parameters $\lambda^{(\text{old})}$ and $\beta^{(\text{old})}$. Hence using results 3.1 and 3.2, we can derive the required expectations as

$$
\mathbb{E}\left[\boldsymbol{w}^\top\right] = \mathbb{E}\left[\boldsymbol{w}\right]^\top = \boldsymbol{\mu}_N^\top \tag{3.4}
$$

$$
\mathbb{E}\left[\boldsymbol{w}^\top\right] = \mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\top\right] = \mathbb{E}\left[\boldsymbol{w}\right]\mathbb{E}\left[\boldsymbol{w}\right]^\top + \text{cov}(\boldsymbol{w}) = \boldsymbol{\mu}_N\boldsymbol{\mu}_N^\top + \boldsymbol{\Sigma}_N \tag{3.5}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \mathbb{E}\left[\text{trace}\left(\boldsymbol{w}^\top\boldsymbol{w}\right)\right] && \left(\because\ \boldsymbol{w}^\top\boldsymbol{w} \text{ is a scalar}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \mathbb{E}\left[\text{trace}\left(\boldsymbol{w}\boldsymbol{w}^\top\right)\right] && \left(\text{using trace property [Mur22, 7.33]}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \text{trace}\left(\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\top\right]\right) && \left(\text{using linearity of expectation}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \text{trace}\left(\mathbb{E}\left[\boldsymbol{w}\right]\mathbb{E}\left[\boldsymbol{w}\right]^\top + \text{cov}(\boldsymbol{w})\right) && \left(\text{using property [Mur22, 3.4]}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \text{trace}\left(\mathbb{E}\left[\boldsymbol{w}\right]\mathbb{E}\left[\boldsymbol{w}\right]^\top\right) + \text{trace}\left(\text{cov}(\boldsymbol{w})\right) && \left(\text{using trace property [Mur22, 7.31]}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \text{trace}\left(\mathbb{E}\left[\boldsymbol{w}\right]^\top\mathbb{E}\left[\boldsymbol{w}\right]\right) + \text{trace}\left(\text{cov}(\boldsymbol{w})\right) && \left(\text{using trace property [Mur22, 7.33]}\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \mathbb{E}\left[\boldsymbol{w}\right]^\top\mathbb{E}\left[\boldsymbol{w}\right] + \text{trace}\left(\text{cov}(\boldsymbol{w})\right) \\
\text{or,}\quad &\mathbb{E}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] = \boldsymbol{\mu}_N^\top\boldsymbol{\mu}_N + \text{trace}\left(\boldsymbol{\Sigma}_N\right)
\end{aligned}
\tag{3.6}
$$

Hence, using results 3.4, 3.5 and 3.6, we can rewrite equation 3.3 as

$$
\begin{aligned}
\mathcal{L}\left(\lambda, \beta, \lambda^{(\text{old})}, \beta^{(\text{old})}\right) = \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^{N}&\left[y_n^2 - 2y_n\,\boldsymbol{\mu}_N^\top\boldsymbol{x}_n + \boldsymbol{x}_n^\top\,\left(\boldsymbol{\mu}_N\boldsymbol{\mu}_N^\top + \boldsymbol{\Sigma}_N\right)\boldsymbol{x}_n\right] \\
&+\frac{D}{2}\log\lambda - \frac{\lambda}{2}\left(\boldsymbol{\mu}_N^\top\boldsymbol{\mu}_N + \text{trace}\left(\boldsymbol{\Sigma}_N\right)\right) + C
\end{aligned}
\tag{3.7}
$$

M step

In the M step, the parameters λ and β are updated to their new values $\lambda^{(\text{new})}$ and $\beta^{(\text{new})}$ respectively, by maximising the expected CLL $\mathcal{L}\left(\lambda, \beta, \lambda^{(\text{old})}, \beta^{(\text{old})}\right)$ w.r.t. the parameters. Hence

$$\left\{\lambda^{(\text{new})}, \beta^{(\text{new})}\right\} = \underset{\lambda,\beta}{\arg\max}\, \mathcal{L}\left(\lambda, \beta, \lambda^{(\text{old})}, \beta^{(\text{old})}\right)$$

To update λ,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda} &= 0 \\ \text{or,} \qquad \frac{D}{2\lambda} - \frac{1}{2}\left(\boldsymbol{\mu}_N^\top \boldsymbol{\mu}_N + \text{trace}\left(\boldsymbol{\Sigma}_N\right)\right) = 0 \\ \text{or,} \qquad \lambda^{(\text{new})} = \frac{D}{\boldsymbol{\mu}_N^\top \boldsymbol{\mu}_N + \text{trace}\left(\boldsymbol{\Sigma}_N\right)} \end{aligned}$$

To update β,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta} &= 0 \\ \text{or,} \qquad \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\left[y_n^2 - 2y_n\, \boldsymbol{\mu}_N^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top\, \left(\boldsymbol{\mu}_N \boldsymbol{\mu}_N^\top + \boldsymbol{\Sigma}_N\right) \boldsymbol{x}_n\right] = 0 \\ \text{or,} \qquad \beta^{(\text{new})} = \frac{N}{\sum_{n=1}^{N}\left[y_n^2 - 2y_n\, \boldsymbol{\mu}_N^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top\, \left(\boldsymbol{\mu}_N \boldsymbol{\mu}_N^\top + \boldsymbol{\Sigma}_N\right) \boldsymbol{x}_n\right]} \end{aligned}$$

where, $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ have the expressions derived in equations 3.1 and 3.2 respectively.

Now, let us sketch the entire EM algorithm to estimate the hyperparameters λ and β.

EM algorithm to estimate hyperparameters λ and β in probabilistic linear regression

1. Initialise $\lambda^{(0)}, \beta^{(0)}$, set $t = 1$.
2. Repeat till convergence

   (a) **E step** : Compute CP of latent variable $\boldsymbol{w}$ given current parameters $\lambda^{(t-1)}, \beta^{(t-1)}$.

   $$\text{CP of } \boldsymbol{w} : \quad p\left(\boldsymbol{w} \mid \mathbf{X}, \boldsymbol{y}, \beta^{(t-1)}, \lambda^{(t-1)}\right) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{\mu}_N^{(t)}, \boldsymbol{\Sigma}_N^{(t)}\right)$$

   where,

   $$\begin{aligned}
   \boldsymbol{\Sigma}_N^{(t)} &= \left(\beta^{(t-1)}\mathbf{X}^\top\mathbf{X} + \lambda^{(t-1)}\mathbf{I}\right)^{-1} \\
   \boldsymbol{\mu}_N^{(t)} &= \boldsymbol{\Sigma}_N^{(t)}\left[\beta^{(t-1)}\mathbf{X}^\top\boldsymbol{y}\right] = \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda^{(t-1)}}{\beta^{(t-1)}}\mathbf{I}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}
   \end{aligned}$$

   (b) **M step** : Maximise the expected CLL (eqn. 3.7) w.r.t. λ, β to update parameters.

   $$\begin{aligned}
   &\left\{\lambda^{(t)}, \beta^{(t)}\right\} = \underset{\lambda,\beta}{\arg\max}\, \mathcal{L}\left(\lambda, \beta, \lambda^{(t-1)}, \beta^{(t-1)}\right) \\
   &\lambda^{(t)} = \frac{D}{\left(\boldsymbol{\mu}_N^{(t)}\right)^\top \boldsymbol{\mu}_N^{(t)} + \text{trace}\left(\boldsymbol{\Sigma}_N^{(t)}\right)} \\
   &\beta^{(t)} = \frac{N}{\sum_{n=1}^N \left[y_n^2 - 2y_n\, \boldsymbol{\mu}_N^{(t)\top}\boldsymbol{x}_n + \boldsymbol{x}_n^\top\left(\boldsymbol{\mu}_N^{(t)}\boldsymbol{\mu}_N^{(t)\top} + \boldsymbol{\Sigma}_N\right)\boldsymbol{x}_n\right]}
   \end{aligned}$$

   where $\boldsymbol{\mu}_N^{(t)}$ and $\boldsymbol{\Sigma}_N^{(t)}$ have been evaluated in the E step.

   (c) Set $t = t + 1$.

   On convergence at let's say $t = T$ steps, $\left\{\lambda^{(T)}, \beta^{(T)}\right\}$ will give us a point estimate of the hyperparameters λ, β.

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

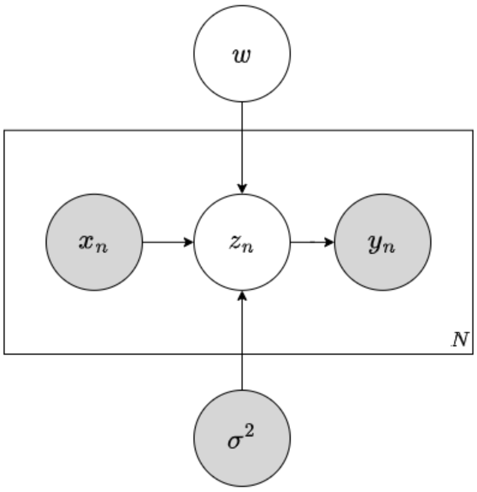
QUESTION 4

Student Name: Atreya Goswami
Roll Number: 190201
Date: March 30, 2023

Given training data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0,1\}$, we assume y_n to be generated from a latent variable $z_n \in \mathbb{R}$ as $y_n = \mathbb{I}[z_n > 0]$, where z_n is a gaussian latent variable with

$$p\left(z_n \mid \boldsymbol{w}, \boldsymbol{x}_n\right) = \mathcal{N}\left(z_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2\right) \qquad (4.1)$$

The plate notation diagram of the model has been shown on the right. $\sigma^2 = 1$ has been assumed. $\mathbb{I}[\cdot]$ is assumed to be the standard indicator function.



Given the model description, we want to design an EM algorithm to get a point estimate of the parameter $\boldsymbol{w}$ and consequently conditional posterior (CP) will also be computed for the latent variables $\{z_n\}_{n=1}^N$.

E step

Using the Markov blanket property, we can write the model likelihood function as

$$p\left(y_n = 1 \mid z_n\right) = \mathbb{I}[z_n > 0] \qquad \text{and} \qquad p\left(y_n = 0 \mid z_n\right) = \mathbb{I}[z_n \leq 0]$$

Let us define $\mathbf{X}, \boldsymbol{y}$ and $\mathbf{z}$ as $\{\boldsymbol{x}_n\}_{n=1}^N$, $\{y_n\}_{n=1}^N$ and $\{z_n\}_{n=1}^N$ respectively stacked vertically. Using the new notations and taking $\sigma^2 = 1$, we can rewrite 4.1 as

$$p\left(\mathbf{z} \mid \boldsymbol{w}, \mathbf{X}\right) = \mathcal{N}\left(\mathbf{z} \mid \mathbf{X}\boldsymbol{w}, \mathbf{I}\right)$$

We now define the Complete-data Log Likelihood (CLL) as

$$\begin{aligned}\log p\left(\mathbf{z}, \boldsymbol{y} \mid \mathbf{X}, \boldsymbol{w}\right) &= \log p\left(\boldsymbol{y} \mid \mathbf{z}\right) + \log p\left(\mathbf{z} \mid \boldsymbol{w}, \mathbf{X}\right) \\ &= \sum_{n=1}^N \log p\left(y_n \mid z_n\right) - \frac{1}{2}\left(\mathbf{z} - \mathbf{X}\boldsymbol{w}\right)^\top \left(\mathbf{z} - \mathbf{X}\boldsymbol{w}\right) + C \qquad (4.2)\end{aligned}$$

Now, we need to take expectation of the CLL w.r.t. the CP of z_n. Hence, the CP of z_n is given by

$$\begin{aligned} p\left(z_n \mid y_n, \boldsymbol{w}, \boldsymbol{x}_n\right) &\propto p\left(y_n \mid z_n\right)\ p\left(z_n \mid \boldsymbol{w}, \boldsymbol{x}_n\right) \\ &= \begin{cases} \mathcal{N}\left(z_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right) \mathbb{I}[z_n > 0] & \text{if } y_n = 1 \\ \mathcal{N}\left(z_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right) \mathbb{I}[z_n \leq 0] & \text{if } y_n = 0 \end{cases} \\ &= \begin{cases} \mathcal{TN}\left(z_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right) [0 < z_n < \infty] & \text{if } y_n = 1 \\ \mathcal{TN}\left(z_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right) [-\infty < z_n \leq 0] & \text{if } y_n = 0 \end{cases} \end{aligned}$$

Hence the CP of z_n is a truncated normal distribution[2] for both cases $y_n = 1$ and $y_n = 0$ with their corresponding supports as shown above. Using the standard notations for truncated normal distribution, for $y_n = 1$, let us define $a = 0$, $b = \infty$, $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$, $\sigma = 1$, ϕ as the standard normal distribution, and Φ as the C.D.F of ϕ.

Hence the expectation[2] of z_n w.r.t $\mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[0 < z_n < \infty]$ can be written as

$$\begin{aligned}
\mathbb{E}_{\mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[0<z_n<\infty]}\left[z_n\right] &= \mu_n + \frac{\phi\left(\frac{a-\mu_n}{\sigma}\right) - \phi\left(\frac{b-\mu_n}{\sigma}\right)}{\Phi\left(\frac{b-\mu_n}{\sigma}\right) - \Phi\left(\frac{a-\mu_n}{\sigma}\right)} \\
&= \boldsymbol{w}^\top \boldsymbol{x}_n + \frac{\phi(-\boldsymbol{w}^\top \boldsymbol{x}_n) - \phi(\infty)}{\Phi(\infty) - \Phi(-\boldsymbol{w}^\top \boldsymbol{x}_n)} \\
&= \boldsymbol{w}^\top \boldsymbol{x}_n + \frac{\phi(\boldsymbol{w}^\top \boldsymbol{x}_n)}{\Phi(\boldsymbol{w}^\top \boldsymbol{x}_n)} \qquad (4.3)
\end{aligned}$$

Similarly, for $y_n = 0$, let us define $a = -\infty$, $b = 0$, $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$, $\sigma = 1$, ϕ as the standard normal distribution, and Φ as the C.D.F of ϕ. Then the expectation[2] of z_n w.r.t $\mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[-\infty < z_n \leq 0]$ can be written as

$$\begin{aligned}
\mathbb{E}_{\mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[-\infty<z_n\leq 0]}\left[z_n\right] &= \mu_n + \frac{\phi\left(\frac{a-\mu_n}{\sigma}\right) - \phi\left(\frac{b-\mu_n}{\sigma}\right)}{\Phi\left(\frac{b-\mu_n}{\sigma}\right) - \Phi\left(\frac{a-\mu_n}{\sigma}\right)} \\
&= \boldsymbol{w}^\top \boldsymbol{x}_n + \frac{\phi(-\infty) - \phi(-\boldsymbol{w}^\top \boldsymbol{x}_n)}{\Phi(-\boldsymbol{w}^\top \boldsymbol{x}_n) - \Phi(-\infty)} \\
&= \boldsymbol{w}^\top \boldsymbol{x}_n - \frac{\phi(\boldsymbol{w}^\top \boldsymbol{x}_n)}{\Phi(-\boldsymbol{w}^\top \boldsymbol{x}_n)} \\
&= \boldsymbol{w}^\top \boldsymbol{x}_n - \frac{\phi(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1 - \Phi(\boldsymbol{w}^\top \boldsymbol{x}_n)} \qquad (4.4)
\end{aligned}$$

Combining results 4.3 and 4.4, we can write

$$\xi_n = \mathbb{E}_{p(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n)}\left[z_n\right] = \begin{cases} \boldsymbol{w}^\top \boldsymbol{x}_n + \frac{\phi(\boldsymbol{w}^\top \boldsymbol{x}_n)}{\Phi(\boldsymbol{w}^\top \boldsymbol{x}_n)} & \text{if } y_n = 1 \\ \boldsymbol{w}^\top \boldsymbol{x}_n - \frac{\phi(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1-\Phi(\boldsymbol{w}^\top \boldsymbol{x}_n)} & \text{if } y_n = 0 \end{cases} \qquad (4.5)$$

We now define $\boldsymbol{\xi} = \mathbb{E}_{p(\mathbf{z} \,|\, \boldsymbol{y}, \boldsymbol{w}, \mathbf{X})}[\mathbf{z}] = \left\{\mathbb{E}_{p(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n)}\left[z_n\right]\right\}_{n=1}^N = \{\xi_n\}_{n=1}^N$ stacked vertically.

Now, in order to derive the expectation of the CLL w.r.t. the CP of $\mathbf{z}$, we calculate the expectations of the terms on the RHS of equation 4.2 individually.

$$\mathbb{E}_{p(\mathbf{z} \,|\, \boldsymbol{y}, \boldsymbol{w}, \mathbf{X})}\left[\sum_{n=1}^N \log p\left(y_n \,|\, z_n\right)\right] = \sum_{n=1}^N \mathbb{E}_{p(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n)}\left[\log p\left(y_n \,|\, z_n\right)\right]$$

For $n : y_n = 1$, the corresponding part of the posterior $p\left(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n\right) = \mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[0 < z_n < \infty]$ has support $z_n > 0$. Hence, in this case $\log p\left(y_n \,|\, z_n\right) = \log p\left(y_n = 1 \,|\, z_n > 0\right) = \log \mathbb{I}[z_n > 0] = 0$. Similarly for $n : y_n = 0$, the corresponding part of the posterior $p\left(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n\right) =$

[2] Expressions for p.d.f and expectations of the truncated normal distribution have been taken from the following Wikipedia entry: https://en.wikipedia.org/wiki/Truncated_normal_distribution

$\mathcal{TN}\left(z_n \,|\, \boldsymbol{w}^\top \boldsymbol{x}_n, 1\right)[-\infty < z_n \leq 0]$ has support $z_n \leq 0$. Hence, $\log p\left(y_n \,|\, z_n\right) = \log p\left(y_n = 0 \,|\, z_n \leq 0\right)$ $= \log \mathbb{I}[z_n \leq 0] = 0$. Thus,

$$\mathbb{E}_{p(\mathbf{z} \,|\, \boldsymbol{y}, \boldsymbol{w}, \mathbf{X})}\left[\sum_{n=1}^{N} \log p\left(y_n \,|\, z_n\right)\right] = \sum_{n=1}^{N} \mathbb{E}_{p(z_n \,|\, y_n, \boldsymbol{w}, \boldsymbol{x}_n)}\left[\log p\left(y_n \,|\, z_n\right)\right] = 0 \tag{4.6}$$

Also,

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z} \,|\, \boldsymbol{y}, \boldsymbol{w}, \mathbf{X})}\left[-\frac{1}{2}\left(\mathbf{z} - \mathbf{X}\boldsymbol{w}\right)^\top \left(\mathbf{z} - \mathbf{X}\boldsymbol{w}\right)\right] &= \mathbb{E}\left[\boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{z} - \frac{1}{2}\boldsymbol{w}^\top \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w}\right] \\ &= \boldsymbol{w}^\top \mathbf{X}^\top \mathbb{E}[\boldsymbol{z}] - \frac{1}{2}\boldsymbol{w}^\top \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w} \\ &= \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{\xi} - \frac{1}{2}\boldsymbol{w}^\top \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w} \qquad (4.7)\end{aligned}$$

Hence from results 4.6 and 4.7 the expected CLL w.r.t the posterior of **z** is given by

$$\mathcal{L} = \mathbb{E}\left[\log p\left(\mathbf{z}, \boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}\right)\right] = \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{\xi} - \frac{1}{2}\boldsymbol{w}^\top \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w} + C' \tag{4.8}$$

where C' is a constant w.r.t. w.

M step

In the M step, we need to maximise the expected CLL obtained (equation 4.8) w.r.t. the model parameter $\boldsymbol{w}$ to update $\boldsymbol{w}$. Thus to get $\boldsymbol{w}^{(\text{new})}$, we set

$$\begin{aligned} & \nabla_{\boldsymbol{w}} \mathcal{L}\,\big|_{\boldsymbol{w} = \boldsymbol{w}^{(\text{new})}} = 0 \\ \text{or,} \qquad & \nabla_{\boldsymbol{w}}\left[\boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{\xi} - \frac{1}{2}\boldsymbol{w}^\top \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w} + C'\right] = 0 \\ \text{or,} \qquad & \mathbf{X}^\top \boldsymbol{\xi} - \left(\mathbf{X}^\top \mathbf{X}\right) \boldsymbol{w} = 0 \\ \text{or,} \qquad & \boldsymbol{w}^{(\text{new})} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\xi}\end{aligned}$$

Now, we sketch the overall EM algorithm as follows.

EM algorithm for a binary classification model

1. Initialise $\boldsymbol{w}^{(0)}$, set $t = 1$.
2. Repeat till convergence

    (a) **E step** : Compute CP of latent variable $\left\{z_n^{(t)}\right\}_{n=1}^{N}$ given current parameter $\boldsymbol{w}^{(t-1)}$

    $$p\left(z_n^{(t)} \mid y_n, \boldsymbol{w}^{(t-1)}, \boldsymbol{x}_n\right) = \begin{cases} \mathcal{TN}\left(z_n^{(t)} \mid \boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n, 1\right)\left[0 < z_n^{(t)} < \infty\right] & \text{if } y_n = 1 \\ \mathcal{TN}\left(z_n^{(t)} \mid \boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n, 1\right)\left[-\infty < z_n^{(t)} \leq 0\right] & \text{if } y_n = 0 \end{cases}$$

    Compute expectation $\boldsymbol{\xi}^{(t)} = \mathbb{E}_{p\left(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{w}^{(t-1)}, \mathbf{X}\right)}[\mathbf{z}] = \left\{\xi_n^{(t)}\right\}_{n=1}^{N}$ where

    $$\xi_n^{(t)} = \begin{cases} \boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n + \frac{\phi(\boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n)}{\Phi(\boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n)} & \text{if } y_n = 1 \\ \boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n - \frac{\phi(\boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n)}{1-\Phi(\boldsymbol{w}^{(t-1)\top}\boldsymbol{x}_n)} & \text{if } y_n = 0 \end{cases}$$

    (b) **M step** : Maximise the expected CLL (eqn. 4.8) w.r.t. $\boldsymbol{w}$ to update parameters.

    $$\boldsymbol{w}^{(t)} = \arg\max_{\boldsymbol{w}} \mathcal{L}$$
    $$\boldsymbol{w}^{(t)} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\xi}^{(t)}$$

    where $\boldsymbol{\xi}^{(t)}$ has been evaluated in the E step.

    (c) Set $t = t + 1$.

    On convergence at let's say $t = T$ steps, $\boldsymbol{w}^{(T)}$ will give us a point estimate of the model parameter $\boldsymbol{w}$.

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

Student Name: Atreya Goswami
Roll Number: 190201
Date: March 30, 2023

QUESTION

5

Gaussian Processes

Part 1: GP Posterior

Given training data $(\mathbf{X}, \boldsymbol{y}) = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, we assume a likelihood $p\left(y_n \,|\, \boldsymbol{x}_n, f\right) = \mathcal{N}\left(y_n \,|\, f(\boldsymbol{x}_n), \sigma^2\right)$ and a GP prior $p\left(f\right) = \mathcal{GP}\left(0, \kappa\right)$, equivalently $p\left(\boldsymbol{f}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$

where, $\boldsymbol{f} = \left[f(x_1), f(x_2), \cdots, f(x_N)\right]^\top$ and $\mathbf{K}$ is a $N \times N$ kernel matrix with $\mathbf{K}_{nm} = \kappa\left(\boldsymbol{x}_n, \boldsymbol{x}_m\right)$.

By vertically stacking $\boldsymbol{x}_n$'s and y_n's in $\mathbf{X}$ and $\boldsymbol{y}$ respectively, we can write the overall likelihood as

$$p\left(\boldsymbol{y} \,|\, \boldsymbol{f}\right) = \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{f}, \sigma^2 \mathbf{I}\right)$$

Therefore using Bayes's rule, we can write the GP posterior as

$$\begin{aligned} p\left(\boldsymbol{f} \,|\, \boldsymbol{y}\right) &= \frac{p\left(\boldsymbol{y} \,|\, \boldsymbol{f}\right)\, p\left(\boldsymbol{f}\right)}{p\left(\boldsymbol{y}\right)} \\ &\propto \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{f}, \sigma^2 \mathbf{I}\right)\, \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right) \end{aligned}$$

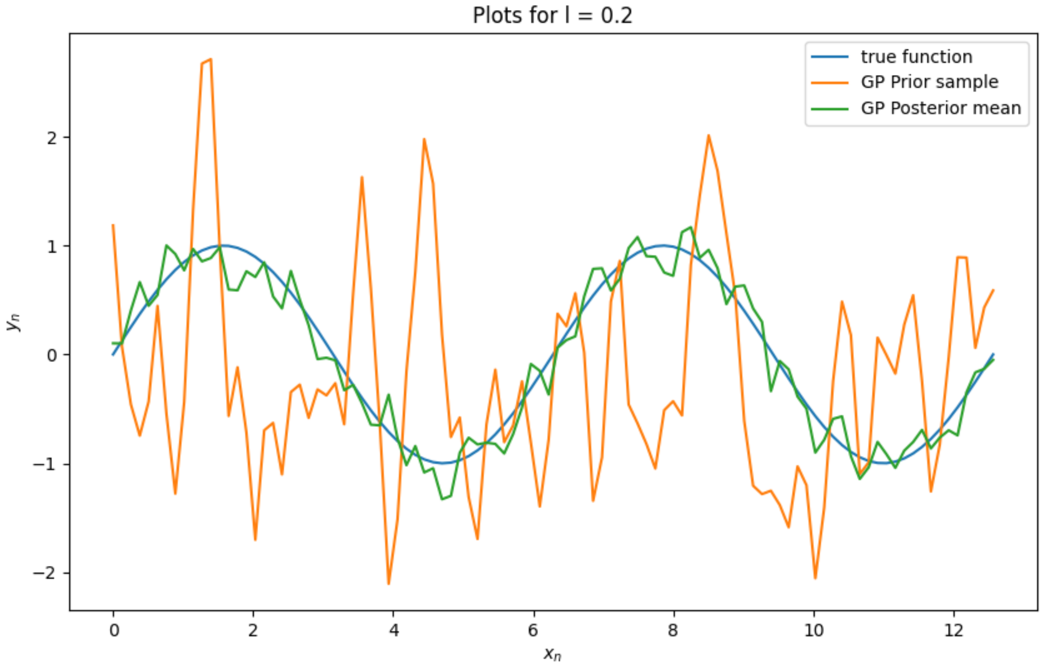
Using standard gaussian results [Bis06, 2.116] and [Bis06, 2.117], we can write the GP posterior

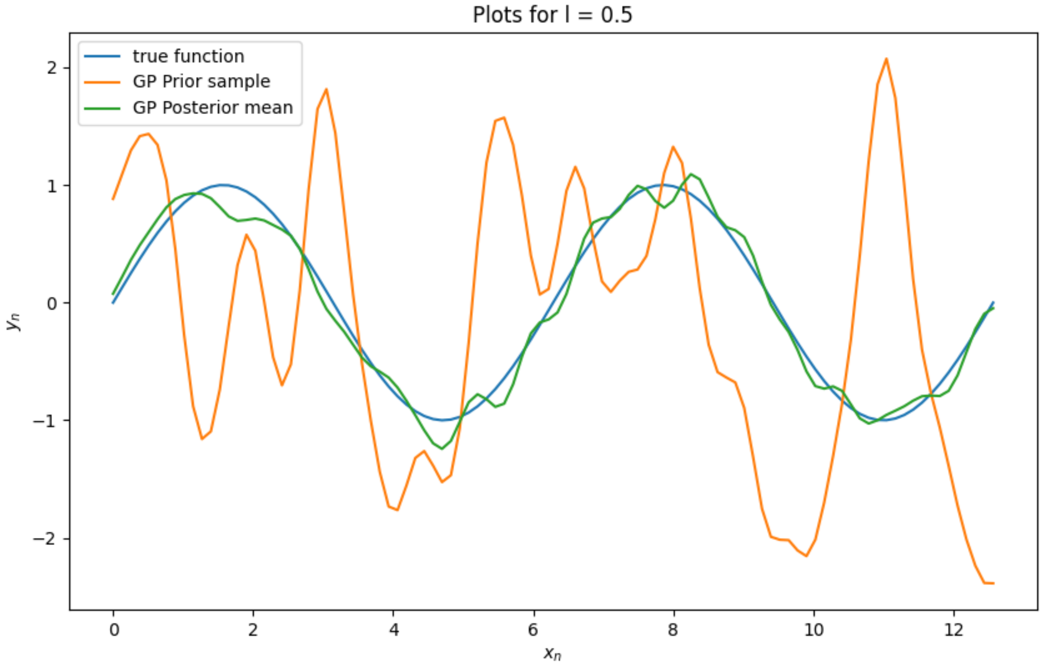
$$p\left(\boldsymbol{f} \,|\, \boldsymbol{y}\right) = \mathcal{N}\left(\boldsymbol{f} \,|\, \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right)$$

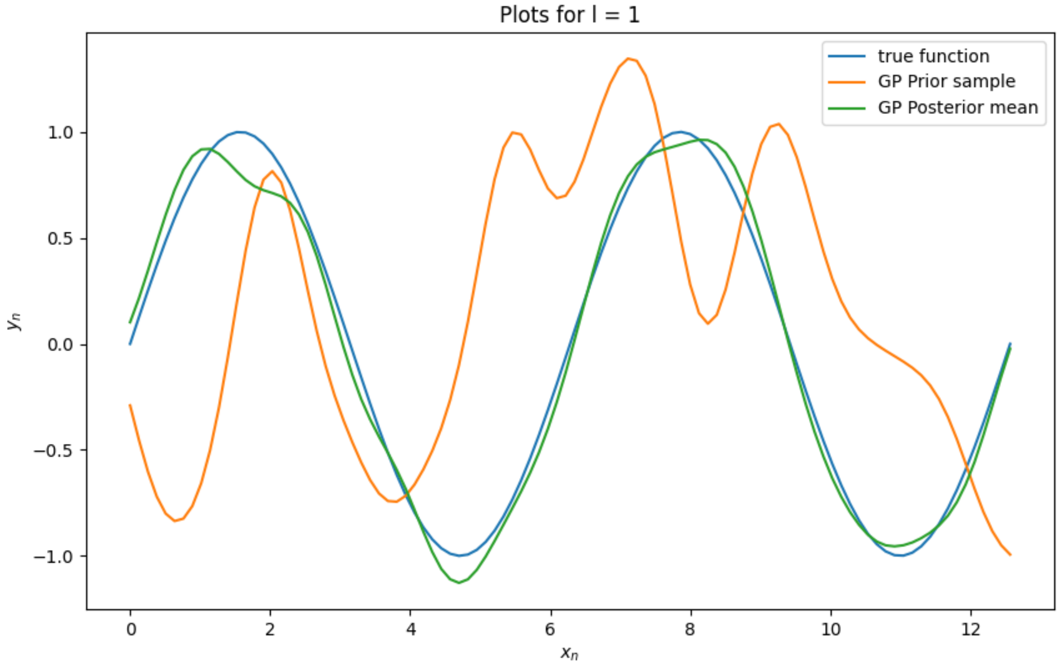
where,

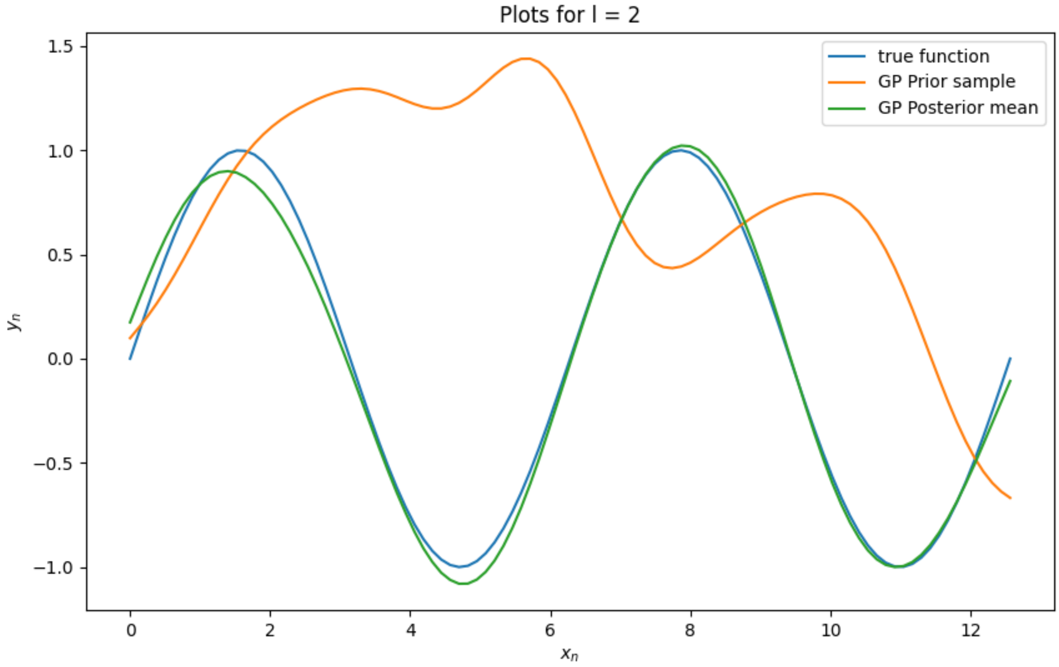
$$\begin{aligned} \boldsymbol{\Sigma}_N &= \left(\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I}\right)^{-1} \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N \left[\sigma^{-2}\mathbf{I} \cdot \boldsymbol{y}\right] = \sigma^{-2}\boldsymbol{\Sigma}_N \cdot \boldsymbol{y} = \left(\sigma^2\mathbf{K} + \mathbf{I}\right)^{-1} \cdot \boldsymbol{y} \end{aligned}$$

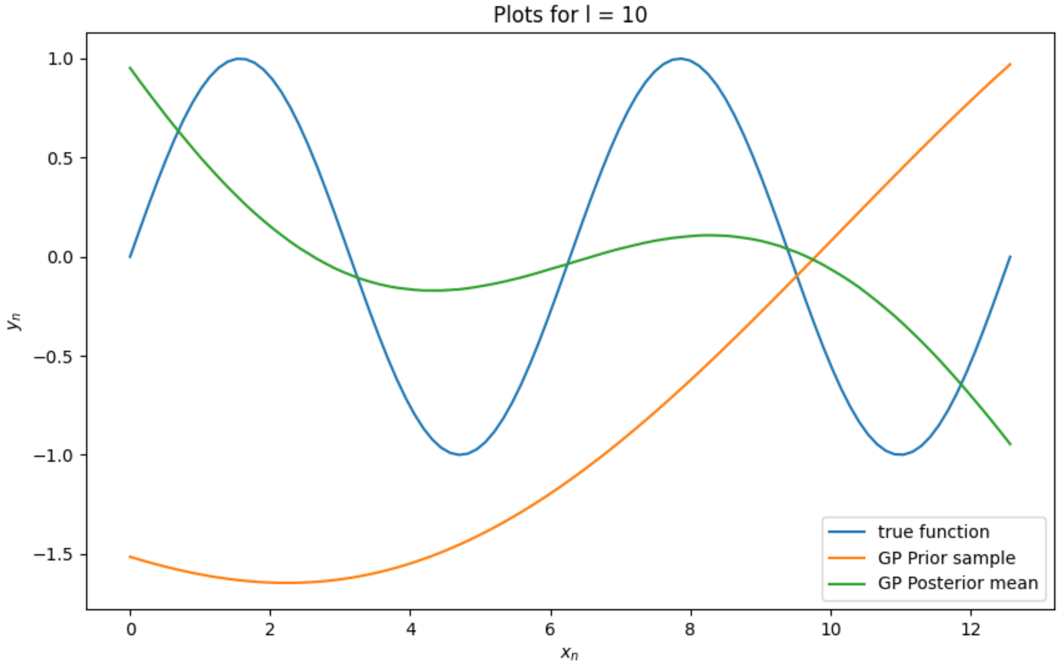
Part 2: Visualizing GP Priors and Posteriors for Regression






Plots for l = 1
true function
GP Prior sample
GP Posterior mean
1.0
0.5
0.0
−0.5
−1.0
y_n
0
2
4
6
8
10
12
x_n


Plots for l = 2
true function
GP Prior sample
GP Posterior mean
1.5
1.0
0.5
0.0
−0.5
−1.0
y_n
0
2
4
6
8
10
12
x_n



Comments on the generated plots

GP Prior sample: The random sample from GP prior doesn't depend on the data, but it depends on the value of l. For small value of l, the curve has more jitters since the off-diagonal terms in the kernel matrix K increase. It is smooth for larger values of l.

GP posterior mean: For small values of l, the posterior mean curve overfits the training data as it tries to model the noise as well. Curve with $l = 2$ best fits the true function as shown by the above RMSE calculations.

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Mur22] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022. URL: http://probml.github.io/book1.