Indian Institute of Technology Kanpur
CS698X: Topics in Probabilistic Modelling and Inference, 18–19

ASSIGNMENT 1

Student Name: Gurpreet Singh
Roll Number: 150259
Date: February 7, 2018

Please note that I have attempted all the questions, with each subsequent question starting from a new page.

Question 1

Below, I state the notations, terms given to us and some observations.

1. **Likelihood**

$$\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\mathbf{w},\beta\,\right] \quad = \quad \prod_{n=1}^{N}\mathcal{N}\left(y_n\,\middle|\,\mathbf{w}^{\mathrm{T}}\mathbf{x}_n,\beta^{-1}\right)$$

2. Let $\boldsymbol{\lambda}$ represent $[\,\lambda_1,\lambda_2\ldots\lambda_d\,]^{\mathrm{T}}$ and Λ represent the $D\times D$ diagonal matrix with element at (d,d) be equal to λ_d

3. **Prior**

$$\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\boldsymbol{\lambda}\,\right] \quad = \quad \prod_{d=1}^{D}\mathcal{N}\left(w_d\,\middle|\,0,\lambda_d^{-1}\right)$$

Part A

I will show that in both the cases, the marginal has a closed form, however will be showing the working for only the second case *i.e.* all λ_d's are different, and then simplify for the first case by putting $\forall\, d\in[\,D\,],\lambda_d=\lambda$.

In order to compute the marginal likelihood, we can compute the posterior of our parameter *i.e.* $\mathbf{w}$, and then use Bayes Rule to compute the marginal likelihood.

$$\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\beta,\boldsymbol{\lambda}\,\right] \quad = \quad \frac{\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\mathbf{w},\beta\,\right]\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\boldsymbol{\lambda}\,\right]}{\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\mathbf{y},X,\beta,\boldsymbol{\lambda}\,\right]} \tag{1}$$

We start by computing the posterior over $\mathbf{w}$.

$$\begin{aligned}
\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\mathbf{y},X,\beta,\boldsymbol{\lambda}\,\right] \quad &\propto \quad \mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\mathbf{w},\beta\,\right]\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\boldsymbol{\lambda}\,\right] \\
&\propto \quad \prod_{n=1}^{N}\mathcal{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n,\beta^{-1}\right)\prod_{d=1}^{D}\mathcal{N}\left(0,\lambda_d^{-1}\right) \\
&\propto \quad \prod_{n=1}^{N}\exp\left(\frac{-\beta}{2}\left(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2\right)\prod_{d=1}^{D}\exp\left(\frac{-\lambda_d}{2}\left(w_d\right)^2\right) \\
&\propto \quad \exp\left(\frac{-1}{2}\left(\beta\sum_{n=1}^{N}\left(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2+\sum_{d=1}^{D}\lambda_d w_d^2\right)\right) \\
&\propto \quad \exp\left(\frac{-1}{2}\left(\beta\mathbf{w}^{\mathrm{T}}X^{\mathrm{T}}X\mathbf{w}-2\mathbf{w}^{\mathrm{T}}X^{\mathrm{T}}\mathbf{y}+\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w}\right)\right)
\end{aligned}$$

Note. I have ignored all terms independent of $\mathbf{w}$ during the above computations

The last term looks like a Gaussian distribution, and hence, we can use the Completing the Squares trick to find out the exact distribution. Therefore, we get

$$\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\mathbf{y}, X, \beta, \boldsymbol{\lambda}\,\right] \quad = \quad \mathcal{N}\left(\mathbf{w}\,\middle|\,\boldsymbol{\mu}_w, \Sigma_w\right) \tag{2}$$

where

$$\begin{aligned}\Sigma_w &= \left(\beta X^{\mathrm{T}} X + \Lambda\right)^{-1} \\ \boldsymbol{\mu}_w &= \left(X^{\mathrm{T}} X + \frac{1}{\beta}\Lambda\right)^{-1} X^{\mathrm{T}}\mathbf{y}\end{aligned}$$

Therefore, we have the posterior. Now we can put this in Equation 1.

$$\begin{aligned}\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \beta, \boldsymbol{\lambda}\,\right] &= \frac{\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \mathbf{w}, \beta\,\right]\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\boldsymbol{\lambda}\,\right]}{\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\mathbf{y}, X, \beta, \boldsymbol{\lambda}\,\right]} \\ &\propto \exp\left(\frac{-1}{2}\left(\beta\sum_{n=1}^{N} y_n^2 + \beta\mathbf{w}^{\mathrm{T}}X^{\mathrm{T}}X\mathbf{w} - 2\beta\mathbf{w}^{\mathrm{T}}X^{\mathrm{T}}\mathbf{y} + \mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w}\right)\right. \\ &\quad \left. + \frac{1}{2}\left(\left(\mathbf{w}-\boldsymbol{\mu}_w\right)^{\mathrm{T}}\Sigma_w^{-1}\left(\mathbf{w}-\boldsymbol{\mu}\right)\right)\right) \\ &\propto \exp\left(\frac{-1}{2}\left(\beta\sum_{n=1}^{N} y_n^2 - {\boldsymbol{\mu}_w}^{\mathrm{T}}\Sigma_w^{-1}\boldsymbol{\mu}_w\right)\right)\end{aligned}$$

Putting $\boldsymbol{\mu}_w = \beta \cdot \Sigma_w X^{\mathrm{T}}\mathbf{y}$,

$$\begin{aligned}\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \beta, \boldsymbol{\lambda}\,\right] &\propto \exp\left(\frac{-1}{2}\left(\beta\sum_{n=1}^{N} y_n^2 - \left(\beta\cdot\Sigma_w X^{\mathrm{T}}\mathbf{y}\right)^{\mathrm{T}}\Sigma_w^{-1}\beta\cdot\Sigma_w X^{\mathrm{T}}\mathbf{y}\right)\right) \\ &\propto \exp\left(\frac{-1}{2}\left(\mathbf{y}^{\mathrm{T}}\left(\beta\mathbf{I} - \beta^2 X{\Sigma_w}^{\mathrm{T}}X^{\mathrm{T}}\right)\mathbf{y}\right)\right)\end{aligned}$$

This again is a gaussian distribution. Therefore, we can write the final form of this distribution as follows

$$\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \beta, \boldsymbol{\lambda}\,\right] \quad = \quad \mathcal{N}\left(\mathbf{y}\,\middle|\,\mathbf{0}, \Sigma\right) \tag{3}$$

where $\Sigma = \left(\beta\,\mathbf{I} - \beta^2\,X\,\Sigma_w\,X^{\mathrm{T}}\right)^{-1}$

We can further simplify the term of the gaussian using a property of matrix inverses. Consider the inverse property given in Equation 166 in [1], which states

$$\left(\mathbf{I} + AB\right)^{-1} \quad = \quad \mathbf{I} - A\left(\mathbf{I} + BA\right)^{-1}B \tag{4}$$

First let us rewrite Σ

$$\begin{aligned}\Sigma &= \frac{1}{\beta}\left(\mathbf{I} - \beta X\left(\beta X^{\mathrm{T}}X + \Lambda\right)^{-1}X^{\mathrm{T}}\right) \\ \implies \beta\,\Sigma &= \left(\mathbf{I} + \left(-X\left(X^{\mathrm{T}}X + \frac{1}{\beta}\Lambda\right)^{-1}\right)X^{\mathrm{T}}\right)\end{aligned}$$

We want to represent this in the form of the term given in the LHS of Equation 4.

Therefore, let $A = \left(-X\left(X^{\mathrm{T}}X + \frac{1}{\beta}\Lambda\right)^{-1}\right)$ and $B = X^{\mathrm{T}}$. Then

$$
\begin{aligned}
\beta\,\Sigma &= (\mathbf{I}+AB)^{-1} \\
&= \mathbf{I}-A\left(\mathbf{I}+BA\right)^{-1}B \\
&= \mathbf{I}-A\left(\mathbf{I}-X^{\mathrm{T}}X\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}\right)^{-1}\right)^{-1}B \\
&= \mathbf{I}-A\left(\mathbf{I}-\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}-\frac{\Lambda}{\beta}\right)\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}\right)^{-1}\right)^{-1}B \\
&= \mathbf{I}-A\left(\frac{\Lambda}{\beta}\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}\right)^{-1}\right)^{-1}B \\
&= \mathbf{I}-\left(-X\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}\right)^{-1}\left(X^{\mathrm{T}}X+\frac{\Lambda}{\beta}\right)\left(\frac{\Lambda}{\beta}\right)^{-1}\right)X^{\mathrm{T}} \\
&= \mathbf{I}+\beta X\Lambda^{-1}X^{\mathrm{T}} \\
\implies \Sigma &= \frac{1}{\beta}\mathbf{I}+X\Lambda^{-1}X^{\mathrm{T}}
\end{aligned}
$$

Hence, we can now rewrite our answer,

$$
\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\beta,\boldsymbol{\lambda}\,\right] = \mathcal{N}\left(\mathbf{y}\,\middle|\,,\mathbf{0},\Sigma\right) \tag{5}
$$

where $\Sigma = \beta^{-1}\mathbf{I}+X\Lambda^{-1}X^{\mathrm{T}}$

If for all $d = 1\ldots D$, $\lambda_d = \lambda$, we can simply put this in the result obtained in Equation 5

Part B

As discussed in class, we do MLE-II inference by performing MLE on the marginal likelihood $\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\beta,\boldsymbol{\lambda}\,\right]$, and then approximate the posterior using the conditional posterior of $\mathbf{w}$ $\left(\mathbb{P}\left[\,\mathbf{w}\,\middle|\,\mathbf{y},X,\widehat{\beta},\widehat{\boldsymbol{\lambda}}\,\right]\right)$.

Note. We define $\boldsymbol{\lambda}$ and Λ the same as in part A

Taking a cue from the MLE-II derivation given in PRML 3.5 [2], we represent the marginal in a different manner, and then try to maximize it for the purpose of MLE.

$$
\begin{aligned}
\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\beta,\boldsymbol{\lambda}\,\right] &= \int_{\mathbf{w}}\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X,\mathbf{w},\beta\,\right]\mathbb{P}\left[\,\mathbf{w}\,\middle|\,X,\beta,\boldsymbol{\lambda}\,\right]\,\mathrm{d}\mathbf{w} \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\sqrt{\frac{|\,\Lambda\,|}{(2\pi)^{D}}}\int_{\mathbf{w}}\exp\left(-E(\mathbf{w})\right)\,\mathrm{d}\mathbf{w}
\end{aligned}
$$

where $E:\mathbf{w}\mapsto\mathbb{R}$ is the evidence function [2] and is defined as

$$
E(\mathbf{w}) = \frac{\beta}{2}(\mathbf{y}-X\mathbf{w})^{\mathrm{T}}(\mathbf{y}-X\mathbf{w})+\frac{1}{2}\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w}
$$

We can also represent $E(\mathbf{w})$ using $\boldsymbol{\mu}_w$ which is the mean of the posterior as derived in part A.

$$
\begin{aligned}
2E(\mathbf{w}) &= \beta(\mathbf{y}-X\boldsymbol{\mu}_w-X(\mathbf{w}-\boldsymbol{\mu}_w))^{\mathrm{T}}(\mathbf{y}-X\boldsymbol{\mu}_w-X(\mathbf{w}-\boldsymbol{\mu}_w))+\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w} \\
&= \beta(\mathbf{y}-X\boldsymbol{\mu}_w)^{\mathrm{T}}(\mathbf{y}-X\boldsymbol{\mu}_w)+(\mathbf{w}-\boldsymbol{\mu}_w)^{\mathrm{T}}\beta X^{\mathrm{T}}X(\mathbf{w}-\boldsymbol{\mu}_w)-2(w-\boldsymbol{\mu}_w)^{\mathrm{T}}\beta X^{\mathrm{T}}(y-X\boldsymbol{\mu}_w)
\end{aligned}
$$

We can write $\beta X^{\mathrm{T}}\left(y - X\boldsymbol{\mu}_w\right)$ as $\Lambda\boldsymbol{\mu}_w$ by expanding $\boldsymbol{\mu}_w$ as shown below.

$$\begin{aligned}
\boldsymbol{\mu}_w &= \beta\left(\beta X^{\mathrm{T}}X + \Lambda\right)^{-1} X^{\mathrm{T}}\mathbf{y} \\
\implies \beta X^{\mathrm{T}}\left(\mathbf{y} - X\boldsymbol{\mu}_w\right) &= \beta\left(\mathbf{I} - \beta X^{\mathrm{T}}X\left(\beta X^{\mathrm{T}}X + \Lambda\right)^{-1}\right) X^{\mathrm{T}}\mathbf{y} \\
&= \beta\Lambda\left(\beta X^{\mathrm{T}}X + \Lambda\right)^{-1} X^{\mathrm{T}}\mathbf{y} \\
&= \Lambda\boldsymbol{\mu}_w
\end{aligned}$$

Therefore

$$\begin{aligned}
E(\mathbf{w}) &= \beta(\mathbf{y} - X\boldsymbol{\mu}_w)^{\mathrm{T}}\left(\mathbf{y} - X\boldsymbol{\mu}_w\right) + (\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\beta X^{\mathrm{T}}X\left(\mathbf{w} - \boldsymbol{\mu}_w\right) - 2(w - \boldsymbol{\mu}_w)^{\mathrm{T}}\Lambda\boldsymbol{\mu}_w + \mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w} \\
&= \beta(\mathbf{y} - X\boldsymbol{\mu}_w)^{\mathrm{T}}\left(\mathbf{y} - X\boldsymbol{\mu}_w\right) + (\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\beta X^{\mathrm{T}}X\left(\mathbf{w} - \boldsymbol{\mu}_w\right) - 2\mathbf{w}^{\mathrm{T}}\Lambda\boldsymbol{\mu}_w + 2\boldsymbol{\mu}_w{}^{\mathrm{T}}\Lambda\boldsymbol{\mu}_w + \mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w} \\
&= \beta(\mathbf{y} - X\boldsymbol{\mu}_w)^{\mathrm{T}}\left(\mathbf{y} - X\boldsymbol{\mu}_w\right) + (\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\left(\beta X^{\mathrm{T}}X + \Lambda\right)\left(\mathbf{w} - \boldsymbol{\mu}_w\right) + \boldsymbol{\mu}_w{}^{\mathrm{T}}\Lambda\boldsymbol{\mu}_w \\
&= 2E(\boldsymbol{\mu}_w) + (\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{w} - \boldsymbol{\mu}_w\right)
\end{aligned}$$

Hence, we can write

$$E(\mathbf{w}) = E(\boldsymbol{\mu}_w) + \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{w} - \boldsymbol{\mu}_w\right) \tag{6}$$

We can now write the Marginal Likelihood in this form

$$\int_{\mathbf{w}} \exp\left(-E(w)\right) = \exp\left(-E(\boldsymbol{\mu}_w)\right)\int_{\mathbf{w}} \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_w)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{w} - \boldsymbol{\mu}_w\right)\right)$$

Since the integral is just over the exponent part of the posterior distribution of $\mathbf{w}$, we can write

$$\int_{\mathbf{w}} \exp\left(-E(w)\right) = \exp\left(-E(\boldsymbol{\mu}_w)\right)\sqrt{(2\pi)^D\,|\,\Sigma\,|}$$

Hence, the negative log marginal (NLM) can be written as

$$-\log\left(\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \beta, \boldsymbol{\lambda}\,\right]\right) = E(\boldsymbol{\mu}_w) + \frac{1}{2}\log\left(\left|\,\Sigma_w^{-1}\,\right|\right) - \frac{1}{2}\log\left(|\,\Lambda\,|\right) - \frac{N}{2}\log\left(\beta\right) + \text{constants} \tag{7}$$

Now, we can use iterative MLE-II scheme to solve for the hyperparamters.

Suppose we have estimates of β and $\boldsymbol{\lambda}$ as β^t and $\boldsymbol{\lambda}^t$ from the t^{th} iteration of the MLE-II scheme, then we can compute the conditional posterior of $\mathbf{w}$ given these estimates. Note, we have already computed this posterior, as given in Equation 2. Therefore, we can write

$$\mathbb{P}\left[\,\mathbf{w}^{t+1}\,\middle|\,\mathbf{y}, X, \beta^t, \boldsymbol{\lambda}^t\,\right] = \mathcal{N}\left(\mathbf{w}^{t+1}\,\middle|\,\boldsymbol{\mu}_w^t, \Sigma_w^t\right)$$

where

$$\begin{aligned}
\Sigma_w^t &= \left(\beta^t X^{\mathrm{T}}X + \Lambda^t\right)^{-1} \\
\boldsymbol{\mu}_w^t &= \left(X^{\mathrm{T}}X + \frac{1}{\beta^t}\Lambda^t\right)^{-1} X^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

Now, we can continue with the point estimation step of the MLE-II method, *i.e.* estimating the update values β^{t+1} and $\boldsymbol{\lambda}^{t+1}$.

$$\beta^{t+1}, \boldsymbol{\lambda}^{t+1} = \underset{\beta, \boldsymbol{\lambda}}{\arg\min}\; -\log\left(\mathbb{P}\left[\,\mathbf{y}\,\middle|\,X, \beta, \boldsymbol{\lambda}\,\right]\right)$$

Since we wish to minimize the NLM (negative log marginal) given in the above equation, we can equate the partial derivatives of β and $\boldsymbol{\lambda}$ at $\beta = \beta^{t+1}$ and $\forall\, d \in [\,D\,]\; \lambda_d = \boldsymbol{\lambda}_d^{t+1}$ to 0. First, solving for λ_d^{t+1},

$$\left[\,\frac{\delta\left(\mathrm{NLM}\right)}{\delta\left(\lambda_d\right)}\,\right]_{\lambda_d = \lambda_d^{t+1}} = 0$$

Usign Equations 36, 42, 49 and 70 from the Matrix Cookbook [1], and replacing the NLM with the derived term in equation 7

$$0 \quad = \quad \left[\frac{1}{2} \frac{\delta \left(\boldsymbol{\mu}_w{}^{\mathrm{T}} \Lambda \boldsymbol{\mu}_w \right)}{\delta \left(\lambda_d \right)} + \frac{1}{2} \operatorname{Tr} \left(\Sigma_w \frac{\delta \left(\Sigma_w^{-1} \right)}{\delta \left(\lambda_d \right)} \right) - \frac{1}{2\lambda_d} \right]_{\lambda_d = \lambda_d^{t+1}}$$

Using Jacobi's method for solving non-linear equations, let us say that we know the value of $\boldsymbol{\mu}_w$ *i.e.* we estimate $\boldsymbol{\mu}_w = \boldsymbol{\mu}_w^t$ and $\Sigma = \Sigma^t$, therefore, we can write

$$\begin{aligned} \left[\frac{1}{2} \mu_{w,d}^2 + \frac{1}{2} \operatorname{Tr} \left(\Sigma_w \frac{\delta \left(\beta X^{\mathrm{T}} X + \Lambda \right)}{\delta \left(\lambda_d \right)} \right) - \frac{1}{2\lambda_d} \right]_{\lambda_d = \lambda_d^{t+1}} &= 0 \\ \implies \left\{ \lambda_d^{t+1} \right\}^{-1} &= \left\{ \mu_{w,d}^t \right\}^2 + \left\{ \Sigma_w^t \right\}_{(d,d)} \end{aligned}$$

If for all $d \in [\, D \,]$, $\lambda_d = \lambda$, we cannot write the above, since the derivatives in that case would be different. In that case, we follow the same procedure, but will obtain the following update equation

$$\left\{ \lambda^{t+1} \right\}^{-1} \quad = \quad \frac{1}{D} \left[\left\{ \boldsymbol{\mu}_w^t \right\}^{\mathrm{T}} \left\{ \boldsymbol{\mu}_w^t \right\} + \operatorname{Tr} \left(\Sigma_w^t \right) \right]$$

Now, we can also give an update equation for β, following the same steps.

$$\begin{aligned} 0 \quad &= \quad \left[\frac{\delta \left(NLM \right)}{\delta \left(\beta \right)} \right]_{\beta = \beta^{t+1}} \\ &= \quad \left[\frac{1}{2} \left(\mathbf{y} - X \boldsymbol{\mu}_w^t \right)^{\mathrm{T}} \left(\mathbf{y} - X \boldsymbol{\mu}_w^t \right) + \frac{1}{2} \operatorname{Tr} \left(\Sigma_w^t X^{\mathrm{T}} X \right) - \frac{N}{2\beta} \right]_{\beta = \beta^{t+1}} \\ \implies \left\{ \beta^{t+1} \right\}^{-1} \quad &= \quad \frac{1}{N} \left[\left(\mathbf{y} - X \boldsymbol{\mu}_w^t \right)^{\mathrm{T}} \left(\mathbf{y} - X \boldsymbol{\mu}_w^t \right) + \operatorname{Tr} \left(\Sigma_w^t X^{\mathrm{T}} X \right) \right] \end{aligned}$$

Since this is independent of the Λ, we can say that for both the cases, the updation of β will be the same.

Therefore, we give an iterative algorithm for this (as discussed in class) in Algorithm 1

Algorithm 1: MLE-II for Hyperparameter Estimation

1. Assume some initial estimates for the hyperparamters, β^0 and $\boldsymbol{\lambda}^0$
2. Repeat until convergence for t $= 0, 1 \ldots$
   (a) Estimate the posterior over $\mathbf{w}^{t+1}$ as

$$\mathbb{P}\left[\, \mathbf{w}^{t+1} \,\middle|\, \mathbf{y}, X, \boldsymbol{\beta}^t, \boldsymbol{\lambda}^t \,\right] \quad = \quad \mathcal{N}\left(\mathbf{w}^{t+1} \,\middle|\, \boldsymbol{\mu}_w^t, \Sigma_w^t\right)$$

   where

$$\begin{aligned} \Sigma_w^t &= \left(\beta^t X^{\mathrm{T}} X + \Lambda^t\right)^{-1} \\ \boldsymbol{\mu}_w^t &= \beta^t \Sigma^t X^{\mathrm{T}} \mathbf{y} \end{aligned}$$

   (b) Re-esimate the hyperparameters as

$$\left\{\beta^{t+1}\right\}^{-1} \quad = \quad \frac{1}{N}\left[\, \left(\mathbf{y} - X\boldsymbol{\mu}_w^t\right)^{\mathrm{T}} \left(\mathbf{y} - X\boldsymbol{\mu}_w^t\right) + \mathrm{Tr}\left(\Sigma_w^t X^{\mathrm{T}} X\right) \right]$$

   Case 1

$$\forall\, d \in [\, D\,],\ \left\{\lambda_d^{t+1}\right\}^{-1} \quad = \quad \left\{\mu_{w,d}^t\right\}^2 + \left\{\Sigma_w^t\right\}_{(d,d)}$$

   Case 2

$$\left\{\lambda^{t+1}\right\}^{-1} \quad = \quad \frac{1}{D}\left[\, \left\{\boldsymbol{\mu}_w^t\right\}^{\mathrm{T}} \left\{\boldsymbol{\mu}_w^t\right\} + \mathrm{Tr}\left(\Sigma_w^t\right) \right]$$

Question 2

We have the likelihood and the prior, and we want to find out the marginal likelihood. Given

$$\begin{aligned}\mathbb{P}\left[\, x \,\middle|\, \lambda \,\right] &= e^{-\lambda}\frac{\lambda^x}{(x)!} \\ \mathbb{P}\left[\, \lambda \,\middle|\, \alpha, \beta \,\right] &= \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\end{aligned}$$

Therefore, the marginal likelihood can be written as

$$\begin{aligned}\mathbb{P}\left[\, x \,\middle|\, \alpha, \beta \,\right] &= \int_\lambda e^{-\lambda}\frac{\lambda^x}{(x)!}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\ &= \frac{\beta^\alpha}{(x)!\,\Gamma(\alpha)}\int_\lambda \lambda^{x+\alpha-1}e^{-(1+\beta)\lambda}\,\mathrm{d}\lambda\end{aligned}$$

Since the term inside the integral is a form of the gamma distribution, we can try to normalize it.

$$\mathbb{P}\left[\, x \,\middle|\, \alpha, \beta \,\right] = \frac{\beta^\alpha\,\Gamma(\alpha+x)}{(x)!\,\Gamma(\alpha)\,(1+\beta)^{\alpha+x}}\int_\lambda \frac{(1+\beta)^{\alpha+x}}{\Gamma(\alpha+x)}\lambda^{x+\alpha-1}e^{-(1+\beta)\lambda}\,\mathrm{d}\lambda$$

As we know, the integral of a probability distribution is always 1, hence we can write

$$\mathbb{P}\left[\, x \,\middle|\, \alpha, \beta \,\right] = \frac{\beta^\alpha\,\Gamma(\alpha+x)}{(x)!\,\Gamma(\alpha)\,(1+\beta)^{\alpha+x}}$$

Putting in $\alpha = r$ and $\beta = \frac{1-p}{p}$, we get

$$\mathbb{P}\left[\, x \,\middle|\, r, p \,\right] = \frac{\Gamma(r+x)\,p^x(1-p)^r}{\Gamma(r)\;(x)!}$$

Given $r \in \mathbb{N}$, the above distribution would be *Negative Binomial Distribution.*

Question 3

We have the likelihood and the the prior.

$$\begin{aligned} \textbf{Likelihood:} \qquad & \mathbb{P}\left[\, x \,\middle|\, \sigma^2 \,\right] &=& \mathcal{N}\left(x \,\middle|\, 0, \sigma^2\right) \\ \textbf{Prior:} \qquad & \mathbb{P}\left[\, \sigma^2 \,\middle|\, \gamma \,\right] &=& \exp\left(\sigma^2 \,\middle|\, \gamma^2/2\right) \end{aligned}$$

We need to compute the marginal likelihood *i.e.* $\mathbb{P}\left[\, x \,\middle|\, \gamma \,\right]$.

$$\begin{aligned} \mathbb{P}\left[\, x \,\middle|\, \gamma \,\right] &= \int_{\sigma^2} \mathbb{P}\left[\, x \,\middle|\, \sigma^2 \,\right] \mathbb{P}\left[\, \sigma^2 \,\middle|\, \gamma \,\right] \, \mathrm{d}\sigma^2 \\ &= \frac{\gamma^2}{2\sqrt{2\pi}} \int_{\sigma^2} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{\gamma^2\sigma^2}{2}\right) \, \mathrm{d}\sigma^2 \end{aligned}$$

As hinted in the assignment, let us consider the MGF of the probability distribution $\mathbb{P}\left[\, x \,\middle|\, \gamma \,\right]$,

$$\begin{aligned} \mathop{\mathbb{E}}_{x\,|\,\gamma}\left[\, e^{sx} \,\right] &= \int_x e^{sx}\, \mathbb{P}\left[\, x \,\middle|\, \gamma \,\right] \, \mathrm{d}x \\ &= \int_x e^{sx} \left[\int_{\sigma^2} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{\gamma^2\sigma^2}{2}\right) \, \mathrm{d}\sigma^2 \right] \, \mathrm{d}x \end{aligned}$$

Since we can change the order of the integrals [3], we will try to integrate x first

$$\begin{aligned} \mathop{\mathbb{E}}_{x\,|\,\gamma}\left[\, e^{sx} \,\right] &= \frac{\gamma^2}{2\sqrt{2\pi}} \int_{\sigma^2} \int_x e^{sx} \frac{1}{\sigma} \exp\left(-\frac{x}{2\sigma^2}\right) \exp\left(-\frac{\gamma^2\sigma^2}{2}\right) \, \mathrm{d}x \cdot \, \mathrm{d}\sigma^2 \\ &= \frac{\gamma^2}{2\sqrt{2\pi}} \int_{\sigma^2} \frac{1}{\sigma} \exp\left(-\frac{\gamma^2\sigma^2}{2}\right) \int_x \exp\left(-\frac{1}{2\sigma^2}\left(x^2 - 2sx\sigma^2\right)\right) \, \mathrm{d}x \cdot \, \mathrm{d}\sigma^2 \end{aligned}$$

We can manipulate the exponent part in the inner integral to represent a gaussian. Therefore

$$\begin{aligned} \mathop{\mathbb{E}}_{x\,|\,\gamma}\left[\, e^{sx} \,\right] &= \frac{\gamma^2}{2} \int_{\sigma^2} \exp\left(-\frac{\gamma^2\sigma^2}{2} + \frac{s^2\sigma^2}{2}\right) \int_x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x^2 - 2sx\sigma^2 + s^2\sigma^4\right)\right) \, \mathrm{d}x \cdot \, \mathrm{d}\sigma^2 \\ &= \frac{\gamma^2}{2} \int_{\sigma^2} \exp\left(-\frac{\gamma^2\sigma^2}{2} + \frac{s^2\sigma^2}{2}\right) \int_x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - s\sigma^2\right)^2\right) \, \mathrm{d}x \cdot \, \mathrm{d}\sigma^2 \end{aligned}$$

Since we are integrating a gaussian distribution over x, we will get 1 as the answer of the inner integration, hence, we can write

$$\begin{aligned} \mathop{\mathbb{E}}_{x\,|\,\gamma}\left[\, e^{sx} \,\right] &= \frac{\gamma^2}{2} \int_{\sigma^2} \exp\left(\frac{\sigma^2}{2}\left(s^2 - \gamma^2\right)\right) \, \mathrm{d}\sigma^2 \\ &= \frac{\gamma^2}{2} \frac{2}{s^2 - \gamma^2} \left[\exp\left(\frac{\sigma^2}{2}\left(s^2 - \gamma^2\right)\right) \right]_0^\infty \end{aligned}$$

Clearly this will exist if and only if $s^2 \leq \gamma^2$. Suppose this is true, *i.e.* $|\, s \,| \leq |\, \gamma \,|$, then

$$\text{MGF} \quad = \quad \mathop{\mathbb{E}}_{x\,|\,\gamma}\left[\, e^{sx} \,\right] \quad = \quad \frac{1}{1 - \frac{s^2}{\gamma^2}}$$

Comparing with the various MGF in the list from [4], we can say that this is the MGF of a Laplace Distribution where $\mu = 0$ and $b = 1/\gamma$. Therefore

$$x \,|\, y \quad \sim \quad \text{Laplace}\left(0, \gamma^{-1}\right)$$

The plots of both the distributions $\mathbb{P}\left[\, x \,\middle|\, \sigma^2 \,\right]$ and $\mathbb{P}\left[\, x \,\middle|\, \gamma \,\right]$ is given in Figure 1

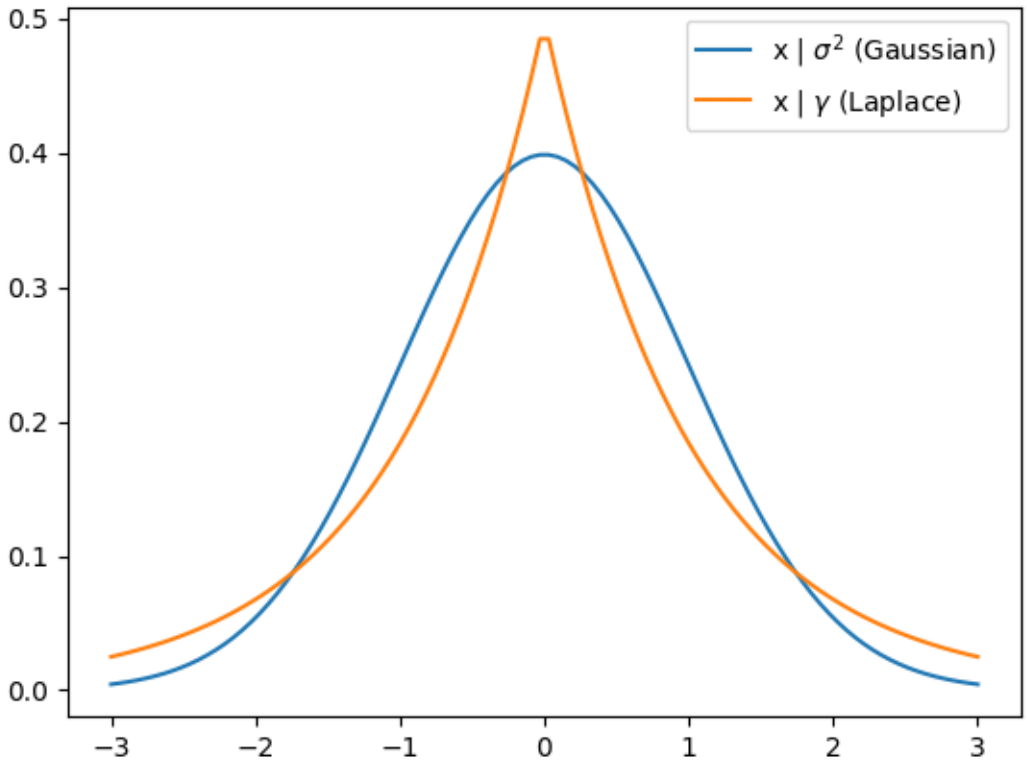

Figure 1: Comparasion of the Gaussian and Laplacian distributions for Question 3

Question 4

Means and Covariances Observed

1. **N = 1**

$$\mu_w = \begin{bmatrix} -0.429 \\ 0.0226 \end{bmatrix} \qquad \Sigma_w = \begin{bmatrix} 0.0382 & 0.0243 \\ 0.0243 & 0.4987 \end{bmatrix}$$

2. **N = 2**

$$\mu_w = \begin{bmatrix} -0.336 \\ 0.6969 \end{bmatrix} \qquad \Sigma_w = \begin{bmatrix} 0.0305 & -0.032 \\ -0.032 & 0.0899 \end{bmatrix}$$

3. **N = 5**

$$\mu_w = \begin{bmatrix} -0.403 \\ 0.6885 \end{bmatrix} \qquad \Sigma_w = \begin{bmatrix} 0.0239 & -0.027 \\ -0.027 & 0.0459 \end{bmatrix}$$

4. **N = 10**

$$\mu_w = \begin{bmatrix} -0.275 \\ 0.4244 \end{bmatrix} \qquad \Sigma_w = \begin{bmatrix} 0.0043 & -0.002 \\ 0 - .002 & 0.0102 \end{bmatrix}$$

5. **N = 20**

$$\mu_w = \begin{bmatrix} -0.239 \\ 0.3688 \end{bmatrix} \qquad \Sigma_w = \begin{bmatrix} 0.0020 & -0.000 \\ -0.000 & 0.0054 \end{bmatrix}$$

Plots for Values of $\mathbf{w}$ Sampled from Posterior

The plots are shown in Figures 2 to 6

Plots for Prediction Lines for Sampled Values of $\mathbf{w}$

The plots are shown in Figures 7 to 11

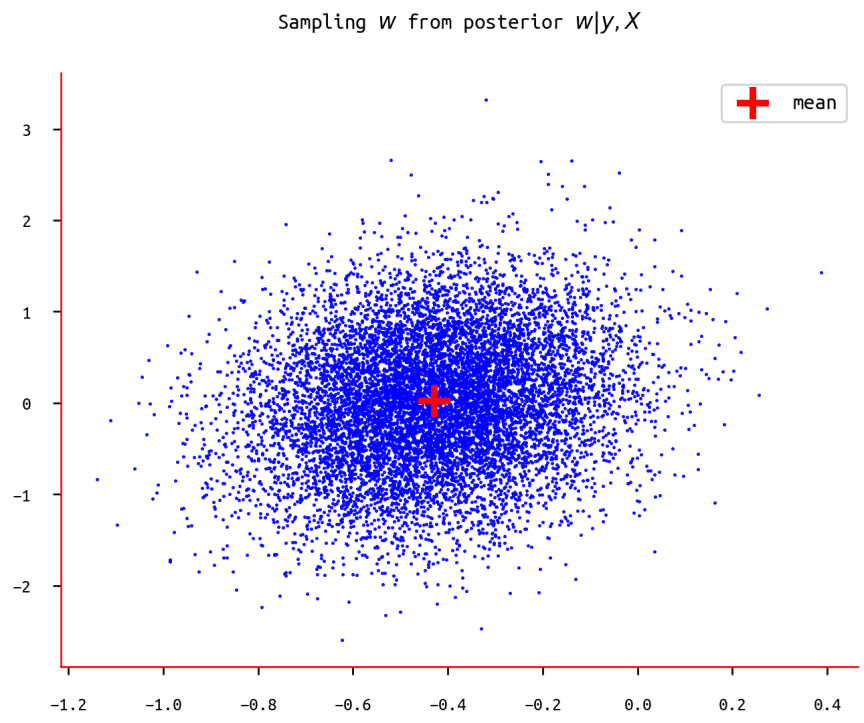

Figure 2: Sampling of w for N = 1

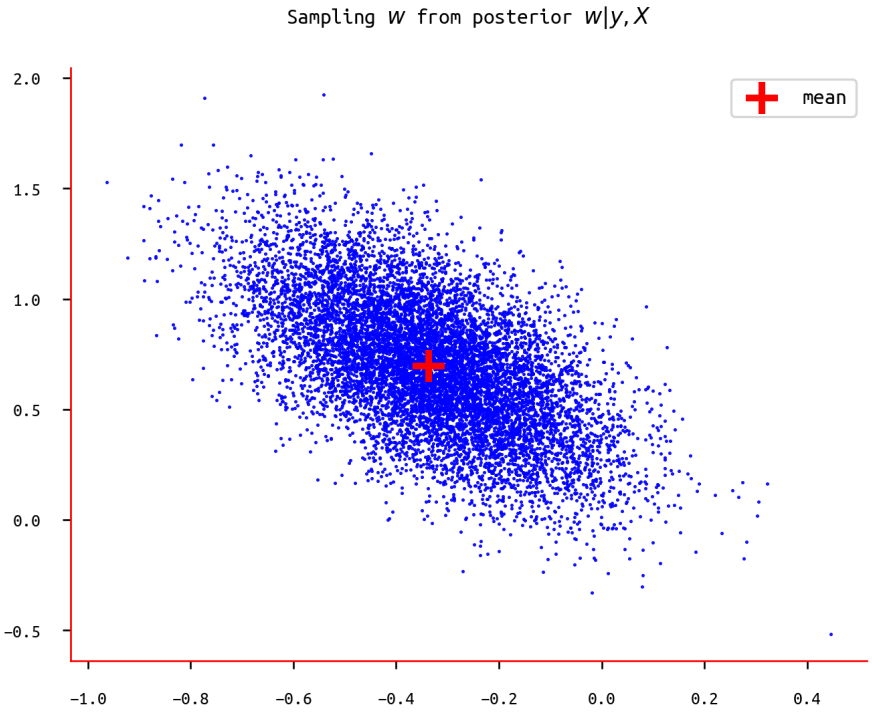

Figure 3: Sampling of w for N = 2

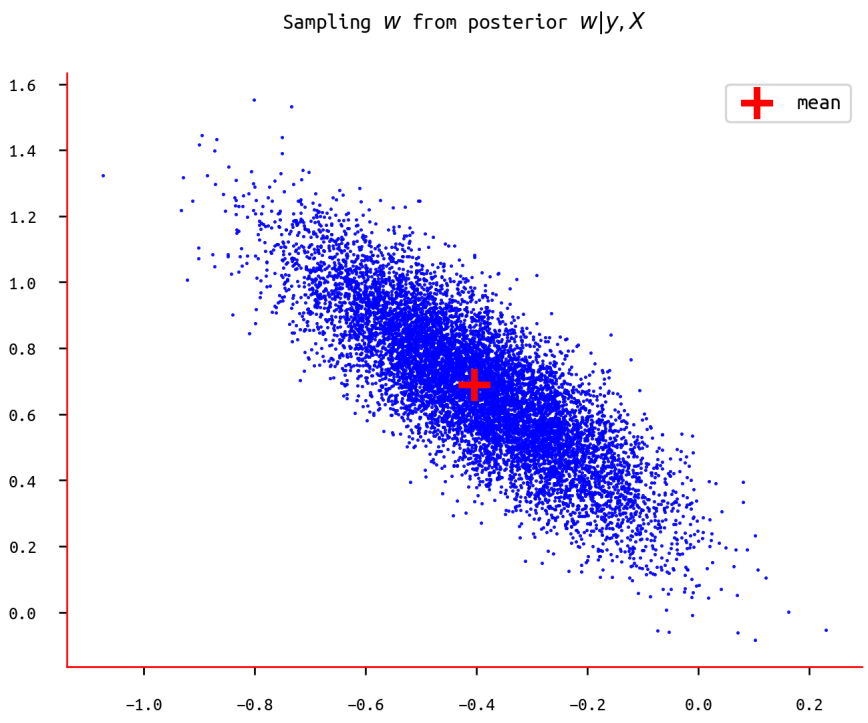

Figure 4: Sampling of w for N = 5

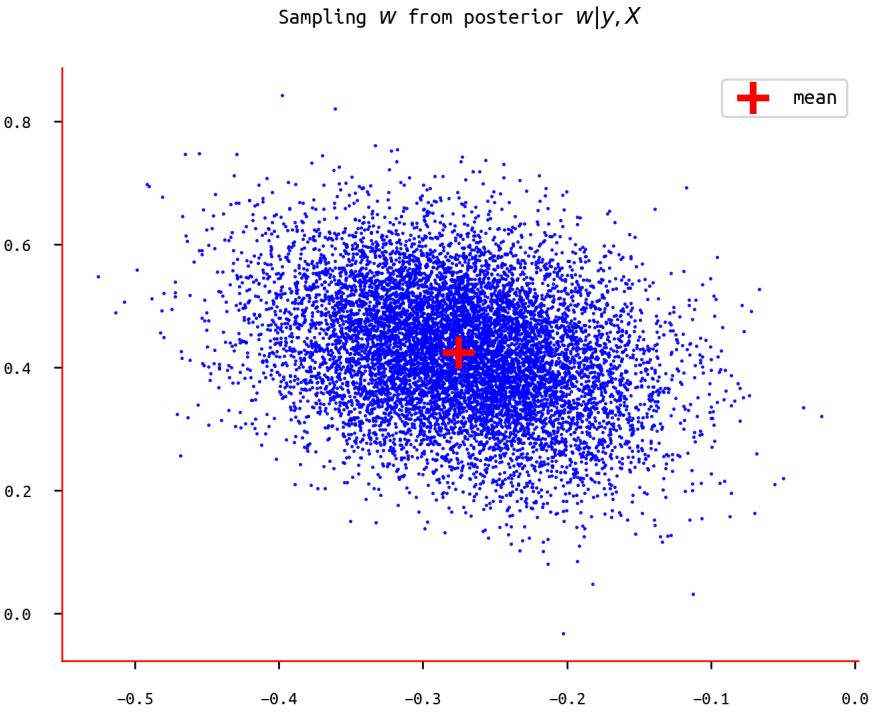

Figure 5: Sampling of w for N = 10

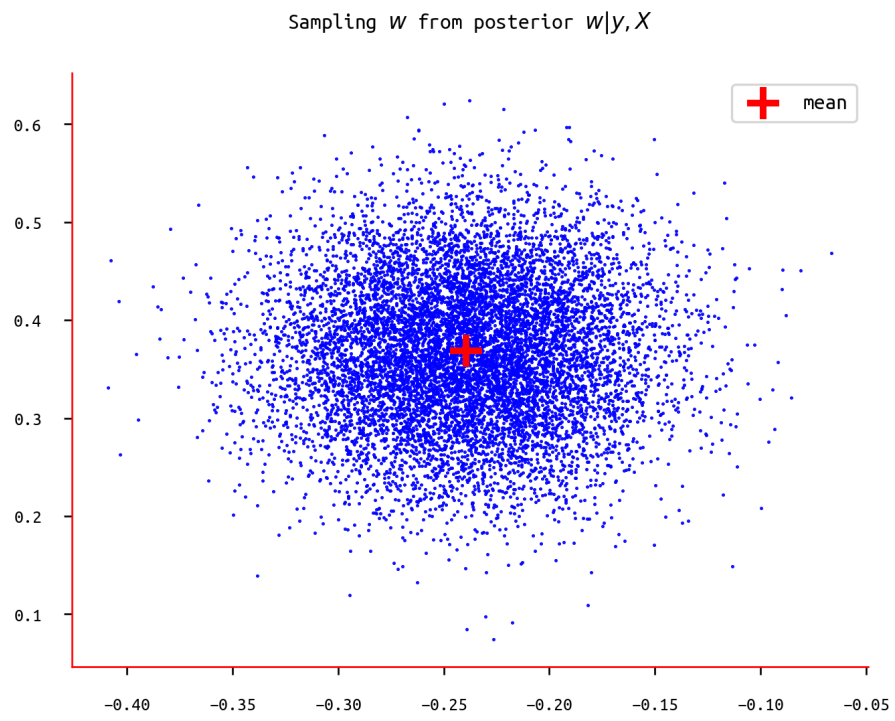

Figure 6: Sampling of w for N = 20

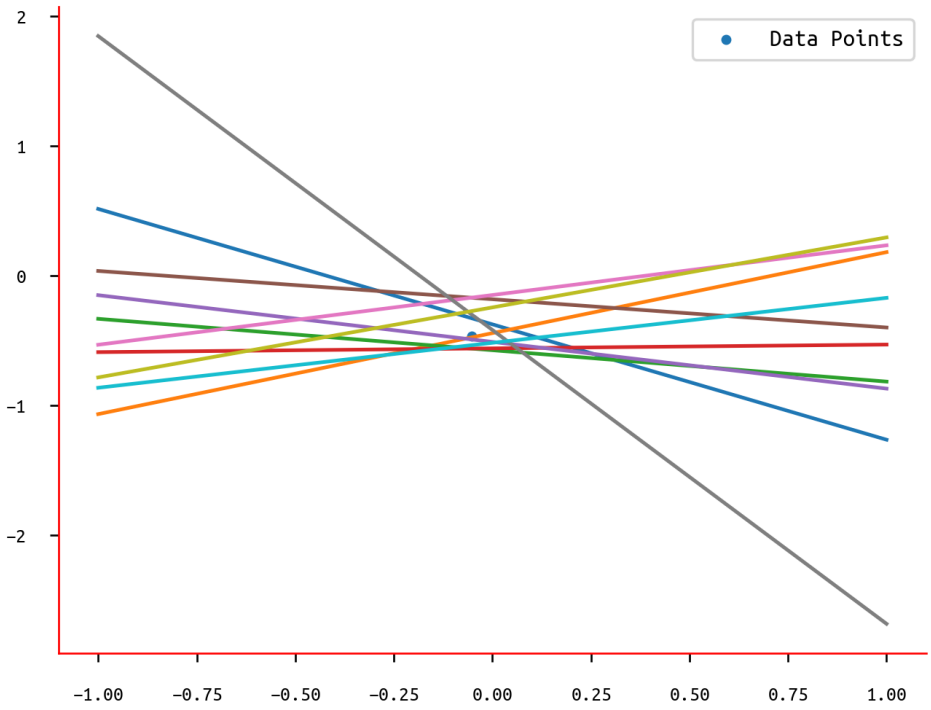

Figure 7: Sampling of w for N = 1

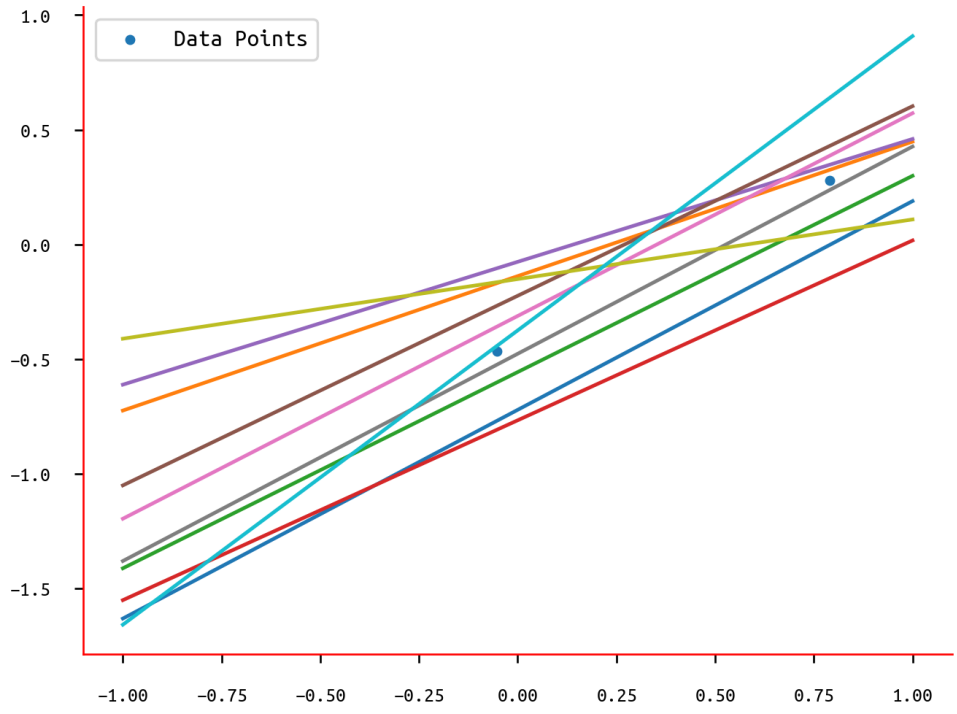

Figure 8: Sampling of w for N = 2

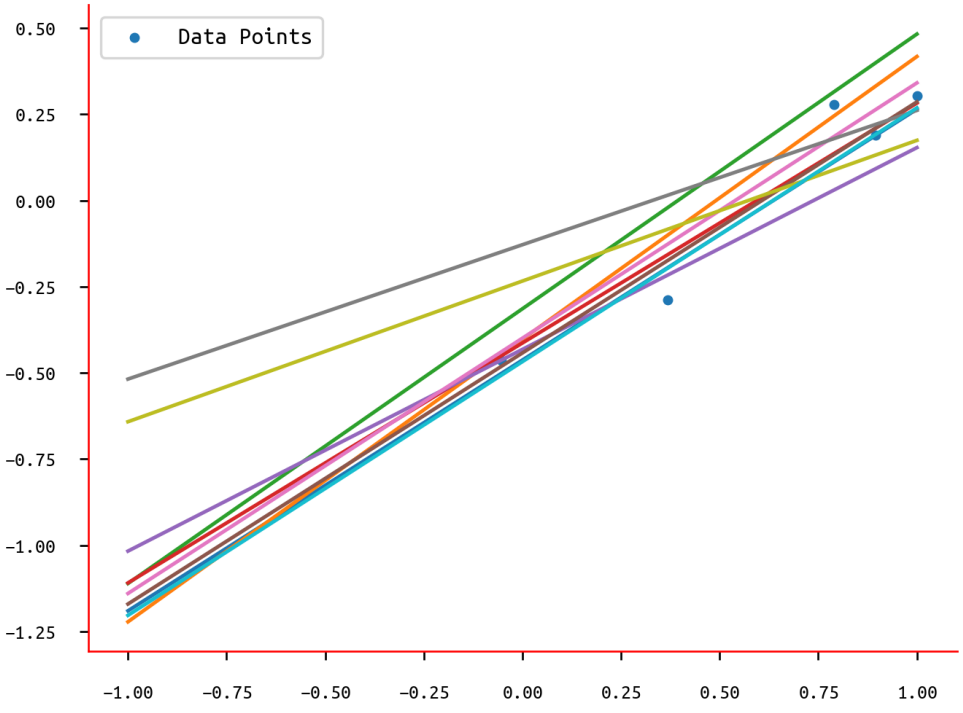

Figure 9: Sampling of w for N = 5

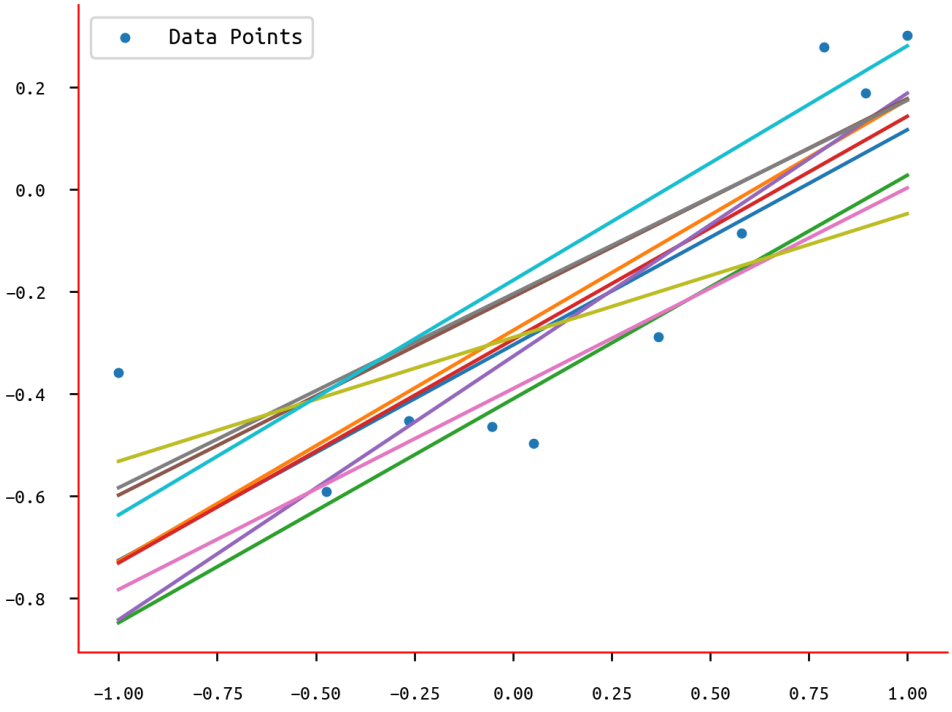

Figure 10: Sampling of w for N = 10

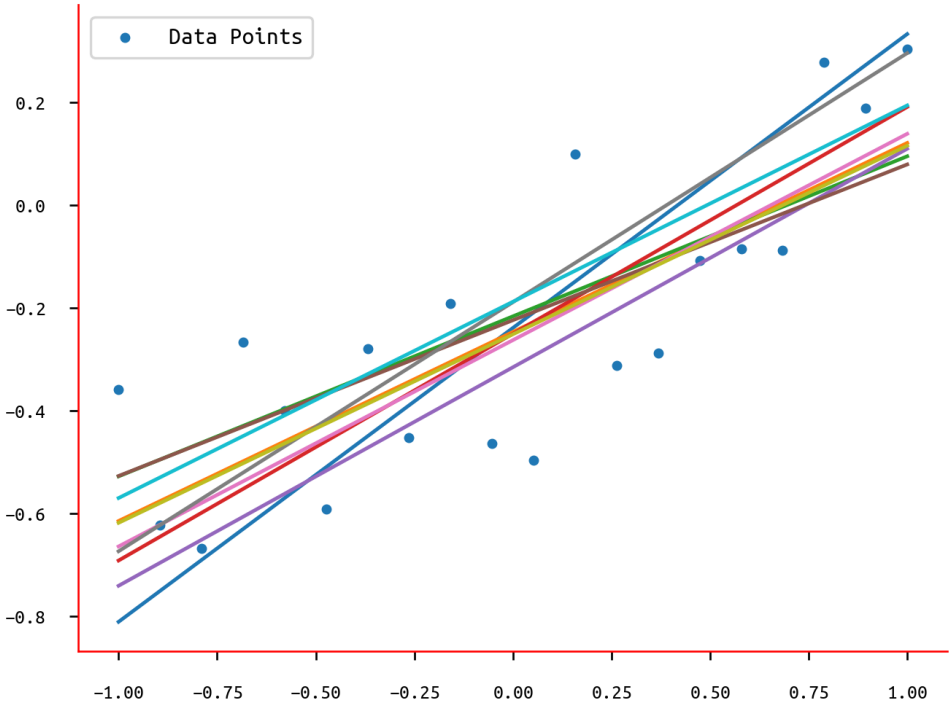

Figure 11: Sampling of w for N = 20

Observations

It can be very clearly seen from both the data in the first section, as well as from the figures that the absolute values of the covariance matrix are going down as the number of points are increasing. This is what is expected, as as the number of points increase, there is more surety about the fit (assuming the assumed model is a good fit, and there is no overfitting for small number of points), and therefore the variance in the inferred parameters will reduce.

Another observation that can be made is that the mean of the posterior is converging, which makes sense from the computed value of the mean, that converges as the covariance matrix converges.

Question 5

Edward is a Turing-complete probabilistic programming language, integrated into TensorFlow (in python), providing significant speedups over existing probabilistic systems.

Edward is built around an iterative process for probabilistic modeling. The process is as follows: formulating a model; inferring the model's hidden structure; criticize how well the model captures the data's generative process. As the tool criticizes the model's fit to the data, it revises components of the model and repeats to form an iterative loop.

1. For *modeling*, Edward provides a language of random variables to construct a broad class of models: directed graphical models, stochastic neural networks, and programs with stochastic control flow
2. For *inference*, Edward provides algorithms such as stochastic and black box variational inference, Hamiltonian Monte Carlo, and Stochastic Gradient Langevin Dynamics.
3. For model *criticism*, Edward provides methods from scoring rules and predictive checks

Models

In Edward, all the models are represented as composotional graphs using random variables. They are essentially class objects with inbuilt methods such as computing the log density or sampling from the random variable.

Furthermore, each random variable X is associated to a tensor (multi-dimensional array) X^*, which represents a single sample X^*X. This association embeds the random variable into a computational graph, where nodes represent operations on tensors and edges represent tensors communicated between them.

This design facilitates developing probabilistic programs in a computational graph framework. This makes it easy to compose random variables with complex deterministic structure such as deep neural networks, a diverse set of math operations, and third party libraries that build on the same framework. The design also enables compositions of random variables to capture complex stochastic structure.

Edward allows all distributions supported by Tensorflow [1]. Using compositional graphical implementation, Edward supports a wide set of applications, such as Graphical Models, Neural Networks as well as Bayesian Nonparametrics using Collapsed Infinite Space (Gaussian Processes and Poisson Processes) or Lazily Defined Infinite Space (Dirichlet Processes).

Inference Tasks

Edward supports many different ways to perform inference for different sets of problems. Inference is an abstract class which takes two inputs: a collection of latent variables, with model variables bound to posterior variables; and a collection of observed variables, with model variables bound to data. I is also possible to fine control the inference process.

Edward also supports passing Model Parameters to compute point-estimates, or even doing conditional inference, by binding random variables to other random variables in the data itself. Latent variables can also be defined in the model without any posterior inference, and will be implicitly marginalized out using a single sample.

The various classes of inference are discussed below.

[1] https://www.tensorflow.org/versions/master/api_guides/python/contrib.distributions

Variational Inference

By specifying the model parameters and an approximation (distribution) class, Edward allows us to perform Variational Inference to approximate the posterior over latent variables.

It is also possible to compute MAP Estimates over the posterior usign Point Mass Random Variables as the approximation class.

Monte Carlo Inference

Monte Carlo approximation simply approximates the posterior with an emperical distribution over the samples generated by updating one sample at a time. It is also possible to leverage the gradients and the graph structure, weherever applicable.

Exact Inference

For some conjugate classes, it is possible to integrate out the paramters, thus computing the exact posterior, for example, exponential class of distributions. This avoids the approximation of posterior and automatically derives exact expressions for classical Gibbs and mean-field updates.

Hybrid Algorithms

Hybrid Algorithms use different class of inference for different variables in the model, for example the EM algorithm with E-step over local (latent) variables and M-step over global variables (model parameters).

Message Passing Algorithms

Edward allows inference using Message Passing, which operates on the posterior distribution using a collection of local inferences. With TensorFlow's distributed training, compositionality enables distributed message passing over a cluster with many workers.

This extends to many algorithms, such as classical message passing, Gibbs sampling, expectation propogation, etc. In all of these, Edward is built to perform local inferences split over infiviual random variables.

Data Subsampling

Edward supports inference steps after looking at only partial data (sample). Only certain algorithms, such as MAP Estimation, KL Divergence and Stochastic Gradient Langevin Dynamics, support data subsampling.

Model Criticism

Model Criticism helps justify the model as an approximation of the true data distribution. Edward explores model criticism using point-based evaluations, such as mean squared error, or posterior predictive checks, for making probabilistic assessments of the model fit using discrepancy functions.

References

[1] Matrix Cookbook. *Kaare Brandt Petersen, Michael Syskind Pedersen*
`http://www.matrixcookbook.com`

[2] Pattern Recongnition and Machine Learning. 2006. *Christopher M. Bishop*

[3] Order of Integration (n.d.) *Wikipedia*
`https://en.wikipedia.org/wiki/Order_of_integration_(calculus)`

[4] Moment Generating Function *Wikipedia*
`https://en.wikipedia.org/wiki/Moment-generating_function`