**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 2, 2018

# Question 1

Here, 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 $\Lambda$ represent the $D\times D$ diagonal matrix with element at $d,d$ be equal to $\lambda_d$

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 have shown that the marginal likelihood is in closed form for both the cases, however am showing the working for only the second case, 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.

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

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

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

First let us rewrite $\Sigma$

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

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 $\Lambda$ 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 $\beta X^{\mathrm{T}}(y - X\boldsymbol{\mu})$ 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

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

Therefore,

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

Therefore, we can now write the Marginal Likelihood in this form

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

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 can be written as

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

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

Suppose we have estimates of $\beta$ and $\boldsymbol{\lambda}$ as $\beta^t$ and $\boldsymbol{\lambda}^t$ from the $\mathrm{t}^{\mathrm{th}}$ 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 of $\beta$ and $\boldsymbol{\lambda}$.

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

Note, we do not directly estimate the MLE updates, instead use an iterative method (like Jacobi method). Therefore, instead of solving for the equations (partial derivates equated to zero), we will try to find an equation which allows for easy updates of the hyperparamters.

Again, taking a cue from PRML [2], we will consider the estimation of $\boldsymbol{\lambda}$ (particulary $\lambda_d$ for some $d$) first.

From the properties of Determinants and Eigenvalues, we know $|A| = \prod_{k=1}^{K} \alpha_k$ where $\alpha_k$ is the $k^{\text{th}}$ eigenvalue of $A$.

Let the eigenvalues of $\left(\beta X^{\mathrm{T}} X + \Lambda^t - \lambda_d^{t+1}\right)$ be represented by the set $\{\gamma_{d'}\}_{d'=1}^{D}$, then we can say that the eigen values of $\left(\beta X^{\mathrm{T}} X + \Lambda^t\right)$ will be $\left\{\gamma_{d'} + \lambda_d^{t+1}\right\}_{d'=1}^{D}$. Therefore, we can now put this in Equation 7

**Note.** $|\Lambda| = \prod_{d'=1}^{D} \lambda_{d'}$ as $\Lambda$ is just a diagonal matrix

$$-\log\left(\mathbb{P}\left[\mathbf{y} \,\middle|\, X, \beta, \boldsymbol{\lambda}\right]\right) = \frac{1}{2}E(\boldsymbol{\mu}_w^t) - \frac{1}{2}\sum_{d'=1}^{D}\log\left(\gamma_{d'} + \lambda_d^{t+1}\right) - \frac{1}{2}\sum_{d'=1}^{D}\log\left(\lambda_{d'}^{t+1}\right) - \frac{N}{2}\log(\beta)$$

Although we cannot compute $\gamma_{d'}$ as we do not yet know the value of $\lambda_d^{t+1}$, however, using the idea of the Jacobi method, we can estimate it by putting $\lambda_d^t$ instead of $\lambda_d^{t+1}$. Now, differentiating with respect to $\lambda_d^{t+1}$, and equating to 0, we get

$$0 = \frac{1}{2\lambda_d^{t+1}} - \frac{1}{2}\left\{\mu_{d,w}^t\right\}^2 - \frac{1}{2}\sum_{d'=1}^{D}\frac{1}{\gamma_{d'} + \lambda_d^{t+1}}$$

**Note.** We are putting in the value of $\lambda_d^{t+1}$ instead of $\lambda_d^t$ in the term for $\Lambda$ and hence we get the derivative of $E(\boldsymbol{\mu}_w)$ as what is shown.

$$\lambda_d^{t+1} = \frac{1}{\left\{\mu_{d,w}^t\right\}^2}\left(1 - \sum_{d'=1}^{D}\frac{\lambda_d^{t+1}}{\gamma_{d'} + \lambda_d^{t+1}}\right)$$

In the fraction part, again using the idea from Jacobi method, we can estimate $\lambda_d^{t+1}$ by $\lambda_d^t$, and therefore we can write

$$\lambda_d^{t+1} = \frac{1}{\left\{\mu_{d,w}^t\right\}^2}\left(1 - \sum_{d'=1}^{D}\frac{\lambda_d^t}{\gamma_{d'} + \lambda_d^t}\right) \triangleq \frac{\alpha_d}{\left\{\mu_{d,w}^t\right\}^2}$$

If for all $d = 1 \ldots D$, $\lambda_d = \lambda$, we cannot directly write this, since all $\lambda_{d'}$ are not independent. The only change in that case will be the derivatives with respect to $\lambda$. Now,

$$\begin{aligned} 0 &= \frac{D}{2}\lambda^{t+1} - \frac{1}{2}\{\boldsymbol{\mu}_w^t\}^{\mathrm{T}}\boldsymbol{\mu}_w^t - \frac{1}{2}\sum_{d'=1}^{D}\frac{1}{\gamma_{d'} + \lambda^{t+1}} \\ \implies \lambda^{t+1} &= \frac{\alpha}{\{\boldsymbol{\mu}_w^t\}^{\mathrm{T}}\boldsymbol{\mu}_w^t} \end{aligned}$$

We now give the MLE update for $\beta^{t+1}$. Since in this case we can use all the estimates $\{\lambda_d^t\}_{d=1}^{D}$, we can simply define $\gamma_d'$ to be the eigenvalues of $\beta X^{\mathrm{T}} X$.

Since for some eigenvector $\mathbf{u}_d$, $\beta X^{\mathrm{T}} X \mathbf{u}_d = \gamma_d'$, we can say

$$\begin{aligned} \frac{d(\gamma_d')}{d(\beta)}\mathbf{u}_d &= X^{\mathrm{T}} X \mathbf{u}_d \\ \implies \frac{d(\gamma_d')}{d(\beta)}\beta\mathbf{u}_d &= \beta X^{\mathrm{T}} X \mathbf{u}_d = \gamma_d'\mathbf{u}_d \\ \implies \frac{d(\gamma_d')}{d(\beta)} &= \frac{\gamma_d'}{\beta} \end{aligned}$$

Therefore, we can write

$$0 \quad = \quad \frac{N}{2\beta} - \frac{1}{2}\left(\mathbf{y} - X\boldsymbol{\mu}_w^t\right)^{\mathrm{T}}\left(\mathbf{y} - X\boldsymbol{\mu}_w^t\right) - \sum_{d=1}^{D} \frac{\beta}{-} < ++ >$$

<++>

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

**Algorithm 1**: MLE-II for Hyperparameter Estimation

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

$$\mathbb{P}\left[\,\mathbf{w}^t \,\middle|\, \mathbf{y}, X, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^{t-1}\,\right] \quad = \quad \mathcal{N}\left(\mathbf{w}^t \,\middle|\, \boldsymbol{\mu}_w^{t-1}, \Sigma_2^{t-1}\right)$$

   where

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

   (b) Re-esimate the hyperparameters as

$$\begin{aligned} \boldsymbol{\lambda}^t &= \\ \beta^t &= \end{aligned}$$

# 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`