1 Bayesian Linear Regression approach

1.1 Rewrite the problem in vector form

First we go to the vector form:

$$\mathbf{\Phi} := \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_m)^T \end{bmatrix} \in \mathbb{R}^{m \times d}$$

In vector form, we can write:

$$y_{1:n} = \mathbf{\Phi w} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}_n \sigma_\varepsilon^2)$$

1.2 Look for the posterior

$$\begin{aligned}
p(\mathbf{w} \mid y_{1:m}) &= \frac{p(\mathbf{w}, y_{1:m})}{p(y_{1:m} \mid)} = \frac{1}{Z} p(y_{1:m} \mid \mathbf{w}) p(\mathbf{w}) \\
&= \frac{1}{Z'} \exp\left(-\frac{1}{2} (y_{1:m} - \mathbf{\Phi w})^T \overbrace{(\sigma_n^2 \mathbf{I}_n)^{-1}}^{\frac{1}{\sigma_n^2}} (y_{1:m} - \mathbf{\Phi w}) \right) \exp\left(-\frac{1}{2} \mathbf{w}^T \overbrace{(\sigma_p^2 \mathbf{I}_d)^{-1}}^{\frac{1}{\sigma_p^2}} \mathbf{w} \right) \\
&= \frac{1}{Z'} \exp\left(-\frac{1}{2} \left(\frac{1}{\sigma_n^2} \|y_{1:m}\|^2 + \frac{1}{\sigma_n^2} \mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi w} + \frac{1}{\sigma_p^2} \mathbf{w}^T \mathbf{w} - \frac{1}{\sigma_n^2} y_{1:m}^T \mathbf{\Phi w} - \frac{1}{\sigma_n^2} \mathbf{w}^T \mathbf{\Phi}^T y_{1:m} \right) \right) \\
&= \frac{1}{Z''} \exp\left(-\frac{1}{2} \left(\mathbf{w}^T \left(\frac{1}{\sigma_n^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_p^2} \mathbf{I}_d \right) \mathbf{w} - \frac{1}{\sigma_n^2} y_{1:m}^T \mathbf{\Phi w} - \frac{1}{\sigma_n^2} \mathbf{w}^T \mathbf{\Phi}^T y_{1:m} \right) \right) \\
&= \frac{1}{Z'''} \exp\left(-\frac{1}{2} \left((\mathbf{w} - \mu)^T \left(\frac{1}{\sigma_n^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_p^2} \mathbf{I}_d \right) (\mathbf{w} - \mu) \right) \right) \\
\bar{\mu} &= \frac{1}{\sigma_n^2} \overbrace{\left(\frac{1}{\sigma_n^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_p^2} \mathbf{I}_d \right)^{-1}}^{\bar{\Sigma}} \mathbf{\Phi}^T y_{1:m}
\end{aligned}$$

Therefore:

$$p(\mathbf{w} \mid y_{1:m}) = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$$

Notice that $\bar{\Sigma} \in \mathbb{R}^{d \times d}$

1.3 Prediction

$$y^* = \phi(x^*)^T \mathbf{w} + \epsilon^*$$

Therefore:

$$y^* \sim \mathcal{N}(\phi(x^*)^T \mu, \phi(x^*)^T \bar{\Sigma} \phi(x^*) + \sigma_n^2)$$

2 GP approach

We now take a 0 mean prior and as kernel we take $k(x,y) = \sigma_p^2 \phi(x)^t \phi(y)$

$$\begin{bmatrix} y^* \\ y_{1:m} \end{bmatrix} = \begin{bmatrix} \phi(x^*) \\ \mathbf{\Phi} \end{bmatrix} \mathbf{w} + \begin{bmatrix} 0 \\ \mathbf{I}_m \end{bmatrix} \epsilon$$

Therefore, since it's all gaussian:

$$\begin{aligned} \begin{bmatrix} y^* \\ y_{1:m} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi(x^*)^T \\ \mathbf{\Phi} \end{bmatrix} \sigma_p^2 \mathbf{I}_d \begin{bmatrix} \phi(x^*) & \mathbf{\Phi}^T \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{I}_m \end{bmatrix} \sigma_\varepsilon^2 \mathbf{I}_m \begin{bmatrix} 0 & \mathbf{I}_m \end{bmatrix} \right) \\ &= \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi(x^*)^T \phi(x^*) \sigma_p^2 & \sigma_p^2 \phi(x^*)^T \mathbf{\Phi}^T \\ \sigma_p^2 \mathbf{\Phi} \phi(x^*) & \sigma_p^2 \mathbf{\Phi}\mathbf{\Phi}^T + \sigma_\varepsilon^2 \mathbf{I}_m \end{bmatrix} \right) \end{aligned}$$

Now, remembering the definition of the kernel we notice:

- $$\sigma_p^2 \phi(x)^T \phi(x') = k(x, x')$$

- $$\sigma_p^2 \mathbf{\Phi} \phi(x^*) = \sigma_p^2 \begin{bmatrix} \phi(x_1) \\ \vdots \\ \phi(x_n) \end{bmatrix} \phi(x^*) = \begin{bmatrix} \sigma_p^2 \phi(x_1)^T \phi(x^*) \\ \vdots \\ \sigma_p^2 \phi(x_v)^T \phi(x^*) \end{bmatrix} = \begin{bmatrix} k(x_1, x^*) \\ \vdots \\ k(x_n, x^*) \end{bmatrix} := \mathbf{k}_{Ax^*}$$

- $$\begin{aligned} \sigma_p^2 \mathbf{\Phi}^T \mathbf{\Phi} &= \sigma_p^2 \begin{bmatrix} \phi(x_1)^T \phi(x) \\ \vdots \\ \phi(x_v)^T \phi(x) \end{bmatrix} \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_n) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_p^2 \phi(x_1)^T \phi(x_1) & \cdots & \sigma_p^2 \phi(x_1)^T \phi(x_n) \\ \vdots & \ddots & \vdots \\ \sigma_p^2 \phi(x_n)^T \phi(x_1) & \cdots & \sigma_p^2 \phi(x_n)^T \phi(x_n) \end{bmatrix} \\ &= \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} := \mathbf{K}_{AA} \end{aligned}$$

Then we can rewrite:

$$\begin{bmatrix} y^* \\ y_{1:m} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) & \mathbf{k}_{Ax^*}^T \\ \mathbf{k}_{Ax^*} & \mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m \end{bmatrix} \right)$$

Now we condition, using the formula before, and we get directly the prediction:

$$\begin{aligned} y^* \mid y_{1:n} &\sim \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right) \\ \tilde{\mu} &= 0 + \mathbf{k}_{Ax^*}^T (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} y_{1:m} \\ \tilde{\Sigma} &= \sigma_p^2 \mathbf{I}_d - \mathbf{k}_{Ax^*}^t (\mathbf{K}_{AA} + \sigma_\varepsilon^2 \mathbf{I}_m)^{-1} \mathbf{k}_{Ax^*} \end{aligned}$$