Machine Learning	Prof. Dr. Mario Fritz
Sheet 3 (Solution)	***Tutorial 2022.05.05***

Exercise 1: Orthogonal

i) Show that the matrix in Equation 1 takes any vector $\boldsymbol{v}$ and projects it onto the space spanned by the columns of $\boldsymbol{\Phi}$.

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} \tag{1}$$

Solution: To prove: (1) $\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}$ is a projection matrix; (2) the projection $\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{v}$ is in the column space of $\boldsymbol{\Phi}$.

(1).
Need to prove that $P = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}$ is idempotent, i.e., $P^2 = P$.

Simply applying the associativity of matrix multiplication:

$$P^2 = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} = P$$

(2).
W.l.o.g. let $\boldsymbol{\Phi} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{v}^* = \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{v}$, then we know that $\boldsymbol{v}^*$ should be a M-dimensional vector (i.e., $\boldsymbol{v}^* \in \mathbb{R}^{M\times 1}$).

By definition of matrix multiplication,

$$\begin{aligned}\boldsymbol{\Phi}\boldsymbol{v}^* &= \left(\boldsymbol{\varphi}_0, ..., \boldsymbol{\varphi}_{M-1}\right)\begin{pmatrix} v_0^* \\ \vdots \\ v_{M-1}^* \end{pmatrix} \\ &= v_0^*\boldsymbol{\varphi}_0 + ... + v_{M-1}^*\boldsymbol{\varphi}_{M-1}\end{aligned}$$

where $\boldsymbol{\varphi}_j \in \mathbb{R}^N$ denotes the j^{th} column of $\boldsymbol{\Phi}$, and $v_j^* \in \mathbb{R}$ is the j^{th} element of $\boldsymbol{v}^*$.

Thus, the projection lies in the space spanned by the columns of $\boldsymbol{\Phi}$.

ii) Use this result to show that the least-squares solution given in Equation 2 corresponds to an orthogonal projection of the vector $\boldsymbol{t}$ onto the manifold $\mathcal{S}$ as shown in Figure 1.

$$\boldsymbol{w}_{\text{ML}} = \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{t} \tag{2}$$

Solution: To prove: the residual of the projection is orthogonal to the column space of $\boldsymbol{\Phi}$.

By definition, $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w}_{\text{ML}} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{t}$.

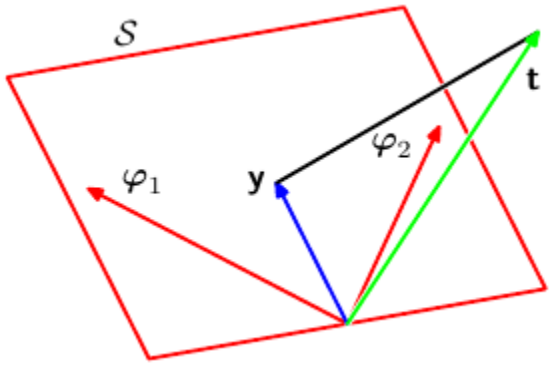

Figure 1: Geometrical interpretation of the least-squares solution: the least-squares regression function is obtained by finding the orthogonal projection of the data vector $\boldsymbol{t}$ onto the column space of $\boldsymbol{\Phi}$ ($\boldsymbol{\varphi}_j$ denotes the j^{th} column of $\boldsymbol{\Phi}$).

Thus, we have

$$\begin{aligned}
(\boldsymbol{y}-\boldsymbol{t})^\top \boldsymbol{\Phi} &= \left(\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{t}-\boldsymbol{t}\right)^\top\boldsymbol{\Phi} \\
&= \left(\left(\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{t}\right)^\top-\boldsymbol{t}^\top\right)\boldsymbol{\Phi} \\
&= \left(\boldsymbol{t}^\top\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top-\boldsymbol{t}^\top\right)\boldsymbol{\Phi} \\
&= \boldsymbol{t}^\top\left(\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top-I\right)\boldsymbol{\Phi} \\
&= \boldsymbol{t}^\top\left(\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}-\boldsymbol{\Phi}\right) \\
&= 0
\end{aligned}$$

Hence, we prove the orthogonality.

Exercise 2: Sample weights

Consider a data set in which each data point $(\boldsymbol{x}_n, y_n)$ is associated with a weighting factor $r_n > 0$, so that the sum-of-squares loss function is as in Equation 3.

i) Find an expression for the solution $\boldsymbol{w}^*$ that minimizes this error function.

ii) Give an interpretation of the weighted sum-of-squares error function in terms of data dependent noise variance.

iii) Give an interpretation of the weighted sum-of-squares error function in terms of replicated data points.

$$L(\boldsymbol{X},\boldsymbol{y},\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\left(y_n-\boldsymbol{w}^\top\varphi(\boldsymbol{x}_n)\right)^2 \tag{3}$$

Solution:

i) To find the minimum, we need to take the derivative w.r.t. $\boldsymbol{w}$:

$$\begin{aligned}\nabla_{\boldsymbol{w}} L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) &= \nabla_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} r_n \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2 \\ &= \frac{1}{2} \sum_{n=1}^{N} r_n \nabla_{\boldsymbol{w}} \left(-2 y_n \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n) + \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n) \varphi(\boldsymbol{x}_n)^\top \boldsymbol{w}\right) \\ &= \sum_{n=1}^{N} r_n \left(-y_n \varphi(\boldsymbol{x}_n) + \varphi(\boldsymbol{x}_n) \varphi(\boldsymbol{x}_n)^\top \boldsymbol{w}\right)\end{aligned}$$

The solution $\boldsymbol{w}^*$ makes the derivative $= 0$ (Checking the second-order derivative we see that the Hessian $\varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top$ is indeed positive semi-definite):

$$\boldsymbol{w}^* = \left(\sum_{n=1}^{N} r_n \varphi(\boldsymbol{x}_n) \varphi(\boldsymbol{x}_n)^\top\right)^{-1} \left(\sum_{n=1}^{N} r_n y_n \varphi(\boldsymbol{x}_n)\right)$$

Let $y'_n = \sqrt{r_n} y_n$, $\boldsymbol{y}' = \begin{pmatrix} y'_1 \\ \vdots \\ y'_N \end{pmatrix}$, $\boldsymbol{\Phi} = \begin{pmatrix} \varphi(\boldsymbol{x}_1) \\ \vdots \\ \varphi(\boldsymbol{x}_N) \end{pmatrix}$, $\boldsymbol{\Phi}' = \begin{pmatrix} \sqrt{r_1}\varphi(\boldsymbol{x}_1) \\ \vdots \\ \sqrt{r_N}\varphi(\boldsymbol{x}_N) \end{pmatrix}$, we can rewrite $\boldsymbol{w}^*$ as:

$$\boldsymbol{w}^* = \left(\boldsymbol{\Phi}'^\top \boldsymbol{\Phi}'\right)^{-1} \boldsymbol{\Phi}'^\top \boldsymbol{y}'$$

ii) Recall the likelihood, log-likelihood, and the error functions of the vanilla least-squares regression:

$$\begin{aligned}p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}(y_n | \boldsymbol{w}^T \varphi(\boldsymbol{x}_n), \beta^{-1}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n))^2}{2\beta^{-1}}\right) \\ \ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(y_n | \boldsymbol{w}^T \varphi(\boldsymbol{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \frac{\beta}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2 \\ L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) &= \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2\end{aligned}$$

Compare with the weighted sum-of-squares error function:

$$L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2$$

We see that the weighted sum-of-squares error function corresponds to log-likelihood of the form:

$$\ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta) \sim \frac{\beta}{2} \sum_{n=1}^{N} r_n \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2 \sim \sum_{n=1}^{N} \ln \mathcal{N}(y_n | \boldsymbol{w}^T \varphi(\boldsymbol{x}_n), r_n^{-1} \beta^{-1})$$

which is Gaussian with data-dependent noise variance.

iii) It can be seen from the log-likelihood function that r_n can be viewed as the effective number of observation of $(\boldsymbol{x}_n, y_n)$, i.e., you can treat $(\boldsymbol{x}_n, y_n)$ as repeatedly occurring r_n times in the data set.

Exercise 3: Independent noise

Consider the linear model given in Equation 4 together with the sum-of-squares loss function given in 5. Now suppose that Gaussian noise ε_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i \varepsilon_j] = \delta_{ij}\sigma^2$ show that minimizing $L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w})$ averaged over the noise distribution is equivalent to minimizing the sum of squares loss for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

$$f(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{4}$$

$$L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(f(\boldsymbol{x}_n, \boldsymbol{w}) - y_n\right)^2 \tag{5}$$

Solution:

Let

$$\begin{aligned}
\widetilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i (x_{ni} + \varepsilon_{ni}) \\
&= w_0 + \sum_{i=1}^{D} w_i x_{ni} + \sum_{i=1}^{D} w_i \varepsilon_{ni} \\
&= f(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{i=1}^{D} w_i \varepsilon_{ni}
\end{aligned}$$

From Equation 5, we then have:

$$\begin{aligned}
\widetilde{L}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) &= \frac{1}{2} \sum_{n=1}^{N} \left(\widetilde{y}_n - y_n\right)^2 \\
&= \frac{1}{2} \sum_{n=1}^{N} \left(\widetilde{y}_n^2 - 2\widetilde{y}_n y_n + y_n^2\right) \\
&= \frac{1}{2} \sum_{n=1}^{N} \left(\left(f(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{i=1}^{D} w_i \varepsilon_{ni} \right)^2 - 2 \left(f(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{i=1}^{D} w_i \varepsilon_{ni} \right) y_n + y_n^2 \right) \\
&= \frac{1}{2} \sum_{n=1}^{N} \left(f(\boldsymbol{x}_n, \boldsymbol{w})^2 + \left(\sum_{i=1}^{D} w_i \varepsilon_{ni} \right)^2 + 2 f(\boldsymbol{x}_n, \boldsymbol{w}) \sum_{i=1}^{D} w_i \varepsilon_{ni} - 2 f(\boldsymbol{x}_n, \boldsymbol{w}) y_n - 2 \sum_{i=1}^{D} w_i \varepsilon_{ni} y_n + y_n^2 \right)
\end{aligned}$$

Taking the expectation over ε_{ni}, we have:

$\mathbb{E}\left[2 f(\boldsymbol{x}_n, \boldsymbol{w}) \sum_{i=1}^{D} w_i \varepsilon_{ni}\right] = 0$ and $\mathbb{E}\left[2 \sum_{i=1}^{D} w_i \varepsilon_{ni} y_n\right] = 0$, since $\mathbb{E}[\varepsilon_{ni}] = 0$

$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \varepsilon_{ni}\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j \varepsilon_{ni} \varepsilon_{nj}\right] = \sum_{i=1}^{D} w_i^2 \sigma^2$, since $\mathbb{E}[\varepsilon_{ni} \varepsilon_{nj}] = \delta_{ij}\sigma^2$

Thus,

$$\begin{aligned}\mathbb{E}\left[\widetilde{L}(\boldsymbol{X},\boldsymbol{y},\boldsymbol{w})\right] &= \mathbb{E}\left[\frac{1}{2}\sum_{n=1}^{N}\left(f(\boldsymbol{x}_n,\boldsymbol{w})^2 - 2f(\boldsymbol{x}_n,\boldsymbol{w})y_n + y_n^2\right)\right] + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i\varepsilon_{ni}\right)^2\right] \\ &= \mathbb{E}\left[L(\boldsymbol{X},\boldsymbol{y},\boldsymbol{w})\right] + \frac{N\sigma^2}{2}\sum_{i=1}^{D} w_i \\ &= L(\boldsymbol{X},\boldsymbol{y},\boldsymbol{w}) + \frac{N\sigma^2}{2}\sum_{i=1}^{D} w_i\end{aligned}$$

where

- Left-hand side: the sum of squares loss (for *noisy* input) averaged over the noise distribution
- Right-hand side: the sum of squares loss (for *noise-free* input) + a weight-decay regularization term

Exercise 4: Linear basis functions

Consider a linear basis function regression model for a multivariate target variable $\boldsymbol{y}$ having a Gaussian distribution as given in Equation 6 where $f(\boldsymbol{x},\boldsymbol{W}) = \boldsymbol{W}^\top\varphi(\boldsymbol{x})$ together with a training data set comprising input basis vectors $\varphi(\boldsymbol{x}_n)$ and corresponding target vectors $\boldsymbol{y}_n$, with $n = 1,...,N$.

$$p(\boldsymbol{y}\,|\,\boldsymbol{W},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{y}\,|\,f(\boldsymbol{x},\boldsymbol{W}),\boldsymbol{\Sigma}) \tag{6}$$

i) Show that the maximum likelihood solution $\boldsymbol{W}_{\mathrm{ML}}$ for the parameter matrix $\boldsymbol{W}$ has the property that each column is given by an expression shown in Equation 2, which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix $\boldsymbol{\Sigma}$.

ii) Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{y}_n - \boldsymbol{W}_{\mathrm{ML}}^\top\varphi(\boldsymbol{x}_n)\right)\left(\boldsymbol{y}_n - \boldsymbol{W}_{\mathrm{ML}}^\top\varphi(\boldsymbol{x}_n)\right)^\top$$

Solution: The likelihood and log-likelihood functions are:

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{W},\boldsymbol{\Sigma}) = \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{y}_n|\boldsymbol{W}^T\varphi(\boldsymbol{x}_n),\boldsymbol{\Sigma}) = \prod_{n=1}^{N}\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))\right)$$

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{W},\boldsymbol{\Sigma}) &= \sum_{n=1}^{N}\mathcal{N}(\boldsymbol{y}_n|\boldsymbol{W}^T\varphi(\boldsymbol{x}_n),\boldsymbol{\Sigma}) \\ &= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{ND}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))\end{aligned}$$

i) Take the derivative of log-likelihood w.r.t. $\boldsymbol{W}$:

$$\begin{aligned}\nabla_{\boldsymbol{W}}\ln p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{W},\boldsymbol{\Sigma}) &= -\frac{1}{2}\nabla_{\boldsymbol{W}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)) \\ &= -\frac{1}{2}\nabla_{\boldsymbol{W}}\sum_{n=1}^{N}\left(-\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n) - \varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_n + \varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\right)\end{aligned}$$

Let's investigate each term separately:

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(-\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\right) &= \nabla_{\boldsymbol{W}}\operatorname{tr}(-\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)) \\ &= -\nabla_{\boldsymbol{W}}\operatorname{tr}\left(\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\right) \\ &= -\varphi(\boldsymbol{x}_n)\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(-\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_n\right) &= \nabla_{\boldsymbol{W}}\operatorname{tr}(-\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_n) \\ &= -\operatorname{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_n\varphi(\boldsymbol{x}_n)^\top) \\ &= -(\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_n\varphi(\boldsymbol{x}_n)^\top)^\top \\ &= -\varphi(\boldsymbol{x}_n)\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\right) &= \nabla_{\boldsymbol{W}}\operatorname{tr}\left(\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\right) \\ &= \nabla_{\boldsymbol{W}}\operatorname{tr}\left(\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top\right) \\ &= \varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1} + \varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}(\boldsymbol{\Sigma}^{-1})^\top \\ &= 2\varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\end{aligned}$$

As the derivative vanishes at $\boldsymbol{W}_{\mathrm{ML}}$, we have:

$$\sum_{n=1}^{N}\varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^\top\boldsymbol{W}_{\mathrm{ML}}\boldsymbol{\Sigma}^{-1} = \sum_{n=1}^{N}\varphi(\boldsymbol{x}_n)\boldsymbol{y}_n^\top\boldsymbol{\Sigma}^{-1}$$

Right multiply both side by $\boldsymbol{\Sigma}$ and rewrite the equation using the design matrix $\boldsymbol{\Phi}$ and target matrix $\boldsymbol{Y}$:

$$\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{W}_{\mathrm{ML}} = \boldsymbol{\Phi}^\top\boldsymbol{Y}$$

i.e.,

$$\boldsymbol{W}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{Y}$$

ii) Take the derivative of log-likelihood function w.r.t. $\boldsymbol{\Sigma}^{-1}$:

$$\begin{aligned}\nabla_{\boldsymbol{\Sigma}^{-1}}\ln p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{W},\boldsymbol{\Sigma}) &= \nabla_{\boldsymbol{\Sigma}^{-1}}\frac{N}{2}\ln|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\nabla_{\boldsymbol{\Sigma}^{-1}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n)) \\ &= \nabla_{\boldsymbol{\Sigma}^{-1}}\frac{N}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\nabla_{\boldsymbol{\Sigma}^{-1}}\sum_{n=1}^{N}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\right) \\ &= \frac{N}{2}\boldsymbol{\Sigma}^\top - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top \\ &= \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top\end{aligned}$$

Hence, we have

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))(\boldsymbol{y}_n - \boldsymbol{W}^\top\varphi(\boldsymbol{x}_n))^\top$$

Reference: Chapter 15 of [3]

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[3] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.