

Machine Learning with Python

MTH786U/P 2023/24

Detailed solutions Coursework 2

Nicola Perra, Queen Mary University of London (QMUL)

Problem 1

Problem 1. In the lecture we discussed different types of error measures. In this task you are asked to compare the stability/robustness of both measures to an outlier. The data samples given are $(x^{(1)}, y^{(1)}) = (-2, 1)$, $(x^{(2)}, y^{(2)}) = (-1, 2)$, $(x^{(3)}, y^{(3)}) = (0, 3)$, $(x^{(4)}, y^{(4)}) = (1, 4)$.

1. Compute the MSE for the 1-parameter model by hand:

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

for $w^{(0)} \in \{1, 2, 3, 4, 5, 6, 7\}$. Between the above values find $w^{(0)}$ that minimises the MSE. A new data sample $(x^{(5)}, y^{(5)}) = (2, 20)$ is added. Evaluate new error measure and corresponding minimiser.

You may find it useful to fill in the missing entries of the following table:

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$							
$y^{(2)} = 2$							
$y^{(3)} = 3$							
$y^{(4)} = 4$							
$\text{MSE}(w) \cdot 2s$							
$y^{(5)} = 20$							
$\text{MSE}(w) \cdot 2s$							

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$
$y^{(1)} = 1$	0
$y^{(2)} = 2$	1
$y^{(3)} = 3$	4
$y^{(4)} = 4$	9
$\text{MSE}(w) \cdot 2s$	14

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$
$y^{(1)} = 1$	0	1
$y^{(2)} = 2$	1	0
$y^{(3)} = 3$	4	1
$y^{(4)} = 4$	9	4
$\text{MSE}(w) \cdot 2s$	14	6

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$
$y^{(1)} = 1$	0	1	4
$y^{(2)} = 2$	1	0	1
$y^{(3)} = 3$	4	1	0
$y^{(4)} = 4$	9	4	1
$\text{MSE}(w) \cdot 2s$	14	6	6

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$
$y^{(1)} = 1$	0	1	4	9
$y^{(2)} = 2$	1	0	1	4
$y^{(3)} = 3$	4	1	0	1
$y^{(4)} = 4$	9	4	1	0
$\text{MSE}(w) \cdot 2s$	14	6	6	14

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$
$y^{(1)} = 1$	0	1	4	9	16
$y^{(2)} = 2$	1	0	1	4	9
$y^{(3)} = 3$	4	1	0	1	4
$y^{(4)} = 4$	9	4	1	0	1
$\text{MSE}(w) \cdot 2s$	14	6	6	14	30

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$
$y^{(1)} = 1$	0	1	4	9	16	25
$y^{(2)} = 2$	1	0	1	4	9	16
$y^{(3)} = 3$	4	1	0	1	4	9
$y^{(4)} = 4$	9	4	1	0	1	4
$\text{MSE}(w) \cdot 2s$	14	6	6	14	30	54

Problem 1

$$\text{MSE}(w^{(0)}) = \frac{1}{2s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|^2,$$

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$
$y^{(1)} = 1$	0	1	4	9	16	25
$y^{(2)} = 2$	1	0	1	4	9	16
$y^{(3)} = 3$	4	1	0	1	4	9
$y^{(4)} = 4$	9	4	1	0	1	4
$\text{MSE}(w) \cdot 2s$	14	6	6	14	30	54

What are the minimum values?

- After another data sample is added one gets an updated table as follows

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$	0	1	4	9	16	25	36
$y^{(2)} = 2$	1	0	1	4	9	16	25
$y^{(3)} = 3$	4	1	0	1	4	9	16
$y^{(4)} = 4$	9	4	1	0	1	4	9
$y^{(5)} = 20$	361	324	289	256	225	196	169
$\text{MSE}(w) \cdot 2s$	375	330	295	270	255	250	255

Minimal value of the MSE is now achieved at $w^{(0)} = 6$.

- After another data sample is added one gets an updated table as follows

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$	0	1	4	9	16	25	36
$y^{(2)} = 2$	1	0	1	4	9	16	25
$y^{(3)} = 3$	4	1	0	1	4	9	16
$y^{(4)} = 4$	9	4	1	0	1	4	9
$y^{(5)} = 20$	361	324	289	256	225	196	169
$\text{MSE}(w) \cdot 2s$	375	330	295	270	255	250	255

Minimal value of the MSE is now achieved at $w^{(0)} = 6$.

Note the values of the MSE!

2. Repeat the same exercise for what is known as the Mean Absolute Error (MAE), i.e.

$$\text{MAE}(w^{(0)}) = \frac{1}{s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|.$$

What do you observe, in particular with regards to the outlier $y^{(5)}$?



2. Repeat the same exercise for what is known as the Mean Absolute Error (MAE), i.e.

$$\text{MAE}(w^{(0)}) = \frac{1}{s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|.$$

What do you observe, in particular with regards to the outlier $y^{(5)}$?

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$	0	1	2	3	4	5	6
$y^{(2)} = 2$	1	0	1	2	3	4	5
$y^{(3)} = 3$	2	1	0	1	2	3	4
$y^{(4)} = 4$	3	2	1	0	1	2	3
$\text{MAE}(w) \cdot s$	6	4	4	6	10	14	18

Minimal value of the MAE is achieved at $w^{(0)} = 2$ and $w^{(0)} = 3$.

2. Repeat the same exercise for what is known as the Mean Absolute Error (MAE), i.e.

$$\text{MAE}(w^{(0)}) = \frac{1}{s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|.$$

What do you observe, in particular with regards to the outlier $y^{(5)}$?

- After another data sample is added one gets an updated table as follows

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$	0	1	2	3	4	5	6
$y^{(2)} = 2$	1	0	1	2	3	4	5
$y^{(3)} = 3$	2	1	0	1	2	3	4
$y^{(4)} = 4$	3	2	1	0	1	2	3
$y^{(5)} = 20$	19	18	17	16	15	14	13
$\text{MAE}(w) \cdot s$	25	22	21	22	25	28	31

2. Repeat the same exercise for what is known as the Mean Absolute Error (MAE), i.e.

$$\text{MAE}(w^{(0)}) = \frac{1}{s} \sum_{i=1}^s |y^{(i)} - w^{(0)}|.$$

What do you observe, in particular with regards to the outlier $y^{(5)}$?

- After another data sample is added one gets an updated table as follows

	$w^{(0)} = 1$	$w^{(0)} = 2$	$w^{(0)} = 3$	$w^{(0)} = 4$	$w^{(0)} = 5$	$w^{(0)} = 6$	$w^{(0)} = 7$
$y^{(1)} = 1$	0	1	2	3	4	5	6
$y^{(2)} = 2$	1	0	1	2	3	4	5
$y^{(3)} = 3$	2	1	0	1	2	3	4
$y^{(4)} = 4$	3	2	1	0	1	2	3
$y^{(5)} = 20$	19	18	17	16	15	14	13
$\text{MAE}(w) \cdot s$	25	22	21	22	25	28	31

Minimal value of the MAE is now achieved at $w^{(0)} = 3$. Compared to the MSE, the additional outlier does not affect the location of the minimum dramatically.

Problem 2

Problem 2. Assume we are given s i.i.d. samples x_1, \dots, x_s , and we know that they are drawn from a normal distribution with mean μ and variance σ^2 . We do not know these two parameters and want to estimate them from the data using the maximum likelihood principle.

1. Write down the likelihood for this data, i.e., the joint probability distribution function $\rho(x_1, \dots, x_s | \mu, \sigma^2)$, where the notation reminds us that this PDF depends on the two parameters μ and σ^2 .
2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.
3. Use the maximum likelihood principle to estimate the parameter σ^2 . Proceed in the same manner as in section 2, but this time with the parameter σ^2 instead of μ .
4. Verify that $-\nabla \log(\rho(w)) = 0$ automatically implies $\nabla \rho(w) = 0$, regardless of the choice of probability density function ρ .

1. Write down the likelihood for this data, i.e., the joint probability distribution function $\rho(x_1, \dots, x_s | \mu, \sigma^2)$, where the notation reminds us that this PDF depends on the two parameters μ and σ^2 .

$$\rho(x_1, \dots, x_s | \mu, \sigma^2) = \prod_{n=1}^s \rho(x_n | \mu, \sigma^2)$$



1. Write down the likelihood for this data, i.e., the joint probability distribution function $\rho(x_1, \dots, x_s | \mu, \sigma^2)$, where the notation reminds us that this PDF depends on the two parameters μ and σ^2 .

$$\begin{aligned}\rho(x_1, \dots, x_s | \mu, \sigma^2) &= \prod_{n=1}^s \rho(x_n | \mu, \sigma^2) \\ &= \prod_{n=1}^s \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right)\end{aligned}$$

1. Write down the likelihood for this data, i.e., the joint probability distribution function $\rho(x_1, \dots, x_s | \mu, \sigma^2)$, where the notation reminds us that this PDF depends on the two parameters μ and σ^2 .

$$\begin{aligned}\rho(x_1, \dots, x_s | \mu, \sigma^2) &= \prod_{n=1}^s \rho(x_n | \mu, \sigma^2) \\ &= \prod_{n=1}^s \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^s \exp\left(-\frac{\sum_{n=1}^s (x_n - \mu)^2}{2\sigma^2}\right).\end{aligned}$$

2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$-\log(\rho(x_1, \dots, x_s | \mu, \sigma^2)) = -\log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^s \exp\left(-\frac{\sum_{n=1}^s (x_n - \mu)^2}{2\sigma^2}\right)\right)$$



2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$\begin{aligned} -\log(\rho(x_1, \dots, x_s | \mu, \sigma^2)) &= -\log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^s \exp\left(-\frac{\sum_{n=1}^s (x_n - \mu)^2}{2\sigma^2}\right)\right) \\ &= \frac{s}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{n=1}^s (x_n - \mu)^2 \end{aligned}$$

2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$\begin{aligned} -\log(\rho(x_1, \dots, x_s | \mu, \sigma^2)) &= -\log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^s \exp\left(-\frac{\sum_{n=1}^s (x_n - \mu)^2}{2\sigma^2}\right)\right) \\ &= \frac{s}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{n=1}^s (x_n - \mu)^2 \\ &= \frac{s}{2} \log(2\pi) + \frac{s}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{n=1}^s (x_n - \mu)^2. \end{aligned}$$

2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$-\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \mu} = \frac{1}{2\sigma^2} \frac{\partial \left(\sum_{n=1}^s (x_n^2 - 2x_n\mu + \mu^2) \right)}{\partial \mu}$$

- Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$\begin{aligned} -\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \mu} &= \frac{1}{2\sigma^2} \frac{\partial \left(\sum_{n=1}^s (x_n^2 - 2x_n\mu + \mu^2) \right)}{\partial \mu} \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^s (-2x_n + 2\mu) \end{aligned}$$

2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$\begin{aligned} -\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \mu} &= \frac{1}{2\sigma^2} \frac{\partial \left(\sum_{n=1}^s (x_n^2 - 2x_n\mu + \mu^2) \right)}{\partial \mu} \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^s (-2x_n + 2\mu) \\ &= \frac{1}{\sigma^2} \sum_{n=1}^s (-x_n + \mu). \end{aligned}$$

2. Use the maximum likelihood principle to estimate the parameter μ . More precisely, compute the gradient of the negative log-likelihood with respect to μ , set it to zero and solve for μ . This gives us an estimator $\hat{\mu}$ of μ that depends on the data.

$$\begin{aligned} -\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \mu} &= \frac{1}{2\sigma^2} \frac{\partial \left(\sum_{n=1}^s (x_n^2 - 2x_n\mu + \mu^2) \right)}{\partial \mu} \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^s (-2x_n + 2\mu) \\ &= \frac{1}{\sigma^2} \sum_{n=1}^s (-x_n + \mu). \end{aligned}$$

Setting this expression to zero, we obtain $\hat{\mu} = \frac{1}{s} \sum_{n=1}^s x_n$.

3. Use the maximum likelihood principle to estimate the parameter σ^2 . Proceed in the same manner as in section 2, but this time with the parameter σ^2 instead of μ .

$$\begin{aligned} -\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \sigma^2} &= \frac{s}{2} \frac{\partial \log(\sigma^2)}{\partial \sigma^2} + \frac{\partial \frac{1}{\sigma^2}}{\partial \sigma^2} \frac{1}{2} \sum_{n=1}^s (x_n - \mu)^2 \\ &= \frac{s}{2} \frac{1}{\sigma^2} - \frac{1}{\sigma^4} \frac{1}{2} \sum_{n=1}^s (x_n - \mu)^2. \end{aligned}$$

3. Use the maximum likelihood principle to estimate the parameter σ^2 . Proceed in the same manner as in section 2, but this time with the parameter σ^2 instead of μ .

$$\begin{aligned} -\frac{\partial \log(\rho(x_1, \dots, x_s | \mu, \sigma^2))}{\partial \sigma^2} &= \frac{s}{2} \frac{\partial \log(\sigma^2)}{\partial \sigma^2} + \frac{\partial \frac{1}{\sigma^2}}{\partial \sigma^2} \frac{1}{2} \sum_{n=1}^s (x_n - \mu)^2 \\ &= \frac{s}{2} \frac{1}{\sigma^2} - \frac{1}{\sigma^4} \frac{1}{2} \sum_{n=1}^s (x_n - \mu)^2. \end{aligned}$$

$$\hat{\sigma}^2 = \frac{1}{s} \sum_{n=1}^s (x_n - \hat{\mu})^2.$$

4. Verify that $-\nabla \log(\rho(w)) = 0$ automatically implies $\nabla \rho(w) = 0$, regardless of the choice of probability density function ρ .



4. Verify that $-\nabla \log(\rho(w)) = 0$ automatically implies $\nabla \rho(w) = 0$, regardless of the choice of probability density function ρ .

$$h(x) = f(g(x))$$



4. Verify that $-\nabla \log(\rho(w)) = 0$ automatically implies $\nabla \rho(w) = 0$, regardless of the choice of probability density function ρ .

$$h(x) = f(g(x))$$

$$h'(x) = f'(g(x)) g'(x)$$



4. Verify that $-\nabla \log(\rho(w)) = 0$ automatically implies $\nabla \rho(w) = 0$, regardless of the choice of probability density function ρ .

$$h(x) = f(g(x))$$

$$h'(x) = f'(g(x)) g'(x)$$

$$\nabla \log(\rho(w)) = \frac{1}{\rho(w)} \nabla \rho(w)$$



Problem 3

Problem 3. In the lecture we have seen that the the general MSE cost function for the linear regression is of the form

$$\text{MSE}(\mathbf{W}) = \frac{1}{2s} \|\mathbf{XW} - \mathbf{Y}\|^2,$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_1^{(s)} & x_2^{(s)} & \dots & x_d^{(s)} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} w_1^{(0)} & w_2^{(0)} & w_3^{(0)} & \dots & w_n^{(0)} \\ w_1^{(1)} & w_2^{(1)} & w_3^{(1)} & \dots & w_n^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_1^{(d)} & w_2^{(d)} & w_3^{(d)} & \dots & w_n^{(d)} \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \dots & y_n^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \dots & y_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(s)} & y_2^{(s)} & \dots & y_n^{(s)} \end{pmatrix},$$

and the norm $\|\cdot\|$ is a Frobenius norm defined by

$$\|\mathbf{M}\|^2 = \sum_{i,j} m_{i,j}^2.$$

Prove that the gradient of MSE is given by

$$\nabla \text{MSE}(w) = \frac{1}{s} \mathbf{X}^\top (\mathbf{XW} - \mathbf{Y}).$$

Problem 3

Problem 3. In the lecture we have seen that the the general MSE cost function for the linear regression is of the form

$$\text{MSE}(\mathbf{W}) = \frac{1}{2s} \|\mathbf{XW} - \mathbf{Y}\|^2,$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_1^{(s)} & x_2^{(s)} & \dots & x_d^{(s)} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} w_1^{(0)} & w_2^{(0)} & w_3^{(0)} & \dots & w_n^{(0)} \\ w_1^{(1)} & w_2^{(1)} & w_3^{(1)} & \dots & w_n^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_1^{(d)} & w_2^{(d)} & w_3^{(d)} & \dots & w_n^{(d)} \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \dots & y_n^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \dots & y_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(s)} & y_2^{(s)} & \dots & y_n^{(s)} \end{pmatrix},$$

Note how this is a more general form with more than 1 label!

and the norm $\|\cdot\|$ is a Frobenius norm defined by

$$\|\mathbf{M}\|^2 = \sum_{i,j} m_{i,j}^2.$$

Prove that the gradient of MSE is given by

$$\nabla \text{MSE}(w) = \frac{1}{s} \mathbf{X}^\top (\mathbf{XW} - \mathbf{Y}).$$

Problem 3

By the definition of Frobenius norm and MSE one has

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n (\mathbf{XW} - \mathbf{Y})_{i,j}^2.$$



Problem 3

By the definition of Frobenius norm and MSE one has

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n (\mathbf{XW} - \mathbf{Y})_{i,j}^2.$$

Using the definition of a matrix product one can write the above as follows

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} \mathbf{X}_{i,k} \mathbf{W}_{k,j} - \mathbf{Y}_{i,j} \right)^2.$$



Problem 3

By the definition of Frobenius norm and MSE one has

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n (\mathbf{XW} - \mathbf{Y})_{i,j}^2.$$

Using the definition of a matrix product one can write the above as follows

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} \mathbf{X}_{i,k} \mathbf{W}_{k,j} - \mathbf{Y}_{i,j} \right)^2.$$

Finally, using the definition of matrices $\mathbf{X}, \mathbf{Y}, \mathbf{W}$ one can write

$$MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right)^2.$$

Problem 3

$$\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) = \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \frac{\partial}{\partial w_q^{(p)}} \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right)^2$$



Problem 3

$$\begin{aligned}\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) &= \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \frac{\partial}{\partial w_q^{(p)}} \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right)^2 \\ &= \frac{1}{s} \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \frac{\partial}{\partial w_q^{(p)}} \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right)\end{aligned}$$



Problem 3

$$\begin{aligned}\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) &= \frac{1}{2s} \sum_{i=1}^s \sum_{j=1}^n \frac{\partial}{\partial w_q^{(p)}} \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right)^2 \\ &= \frac{1}{s} \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \frac{\partial}{\partial w_q^{(p)}} \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \\ &= \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \left(\sum_{k=1}^{d+1} x_k^{(i)} \frac{\partial}{\partial w_q^{(p)}} w_j^{(k)} \right).\end{aligned}$$

Problem 3

$$\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) = \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \left(\sum_{k=1}^{d+1} x_k^{(i)} \frac{\partial}{\partial w_q^{(p)}} w_j^{(k)} \right).$$



Problem 3

Only terms with $j=q$ and $k=p$ are 1!

$$\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) = \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \left(\sum_{k=1}^{d+1} x_k^{(i)} \frac{\partial}{\partial w_q^{(p)}} w_j^{(k)} \right).$$



Problem 3

Only terms with $j=q$ and $k=p$ are 1!

$$\begin{aligned}\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) &= \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \left(\sum_{k=1}^{d+1} x_k^{(i)} \frac{\partial}{\partial w_q^{(p)}} w_j^{(k)} \right) \\ &= \frac{1}{s} \sum_{i=1}^s \left(\sum_{k=1}^{d+1} x_k^{(i)} w_q^{(k)} - y_q^{(i)} \right) x_p^{(i)}\end{aligned}$$



Problem 3

Only terms with $j=q$ and $k=p$ are 1!

$$\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) = \sum_{i=1}^s \sum_{j=1}^n \left(\sum_{k=1}^{d+1} x_k^{(i)} w_j^{(k)} - y_j^{(i)} \right) \left(\sum_{k=1}^{d+1} x_k^{(i)} \frac{\partial}{\partial w_q^{(p)}} w_j^{(k)} \right).$$

$$= \frac{1}{s} \sum_{i=1}^s \left(\sum_{k=1}^{d+1} x_k^{(i)} w_q^{(k)} - y_q^{(i)} \right) x_p^{(i)}$$

$$= \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} x_p^{(i)} x_k^{(i)} w_q^{(k)} - \sum_{i=1}^s x_p^{(i)} y_q^{(i)} \right)$$

Problem 3

$$\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) = \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} \mathbf{X}_{i,p} \mathbf{X}_{i,k} \mathbf{W}_{k,q} - \sum_{i=1}^s \mathbf{X}_{i,p} \mathbf{Y}_{i,q} \right)$$



Problem 3

$$\begin{aligned}\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) &= \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} \mathbf{X}_{i,p} \mathbf{X}_{i,k} \mathbf{W}_{k,q} - \sum_{i=1}^s \mathbf{X}_{i,p} \mathbf{Y}_{i,q} \right) \\ &= \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} \mathbf{X}_{p,i}^\top \mathbf{X}_{i,k} \mathbf{W}_{k,q} - \sum_{i=1}^s \mathbf{X}_{p,i}^\top \mathbf{Y}_{i,q} \right)\end{aligned}$$



Problem 3

$$\begin{aligned}\frac{\partial}{\partial w_q^{(p)}} MSE(\mathbf{W}) &= \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} \mathbf{X}_{i,p} \mathbf{X}_{i,k} \mathbf{W}_{k,q} - \sum_{i=1}^s \mathbf{X}_{i,p} \mathbf{Y}_{i,q} \right) \\ &= \frac{1}{s} \left(\sum_{i=1}^s \sum_{k=1}^{d+1} \mathbf{X}_{p,i}^\top \mathbf{X}_{i,k} \mathbf{W}_{k,q} - \sum_{i=1}^s \mathbf{X}_{p,i}^\top \mathbf{Y}_{i,q} \right) \\ &= \frac{1}{s} \left(\mathbf{X}^\top \mathbf{X} \mathbf{W} - \mathbf{X}^\top \mathbf{Y} \right)_{p,q}.\end{aligned}$$

Problem 4

Problem 4. Compute the solution of the polynomial regression problem

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \sum_{i=1}^s |\langle \phi(\mathbf{x}^{(i)}), \mathbf{w} \rangle - \mathbf{y}^{(i)}|^2 \right\} \quad (2)$$

by hand, for the data samples $(x^{(1)}, y^{(1)}) = (0, 0)$, $(x^{(2)}, y^{(2)}) = (1/4, 1)$, $(x^{(3)}, y^{(3)}) = (1/2, 0)$, $(x^{(4)}, y^{(4)}) = (3/4, -1)$ and $(x^{(5)}, y^{(5)}) = (1, 0)$ and choices

1. $d = 1$,
2. $d = 2$,
3. $d = 3$.



Problem 4

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) \hat{\mathbf{w}} = \Phi(\mathbf{X})^\top \mathbf{Y},$$



Problem 4

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) \hat{\mathbf{w}} = \Phi(\mathbf{X})^\top \mathbf{Y},$$

$$\Phi(\mathbf{X}) = \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \dots & (x^{(1)})^d \\ 1 & x^{(2)} & (x^{(2)})^2 & \dots & (x^{(2)})^d \\ 1 & x^{(3)} & (x^{(3)})^2 & \dots & (x^{(3)})^d \\ 1 & x^{(4)} & (x^{(4)})^2 & \dots & (x^{(4)})^d \\ 1 & x^{(5)} & (x^{(5)})^2 & \dots & (x^{(5)})^d \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \\ y^{(5)} \end{pmatrix}.$$

Problem 4

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) \hat{\mathbf{w}} = \Phi(\mathbf{X})^\top \mathbf{Y},$$

$$\phi^\top(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^3 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \quad \phi(x) = \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$

$$\phi^\top(x)\phi(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^3 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$



$$\phi^T(x)\phi(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^3 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$

$$(\phi^T(x)\phi(x))_{2,3} = x^{(1)} (x^{(1)})^2 + x^{(2)} (x^{(2)})^2 + x^{(3)} (x^{(3)})^2 + x^{(4)} (x^{(4)})^2 + x^{(5)} (x^{(5)})^2$$



$$\phi^\top(x)\phi(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^3 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$

$$\begin{aligned} (\phi^\top(x)\phi(x))_{2,3} &= x^{(1)} (x^{(1)})^2 + x^{(2)} (x^{(2)})^2 + x^{(3)} (x^{(3)})^2 + x^{(4)} (x^{(4)})^2 + x^{(5)} (x^{(5)})^2 \\ &= (x^{(1)})^3 + (x^{(2)})^3 + (x^{(3)})^3 + (x^{(4)})^3 + (x^{(5)})^3 \end{aligned}$$

$$\phi^T(x)\phi(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^3 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$

$$\begin{aligned} (\phi^T(x)\phi(x))_{2,3} &= x^{(1)} (x^{(1)})^2 + x^{(2)} (x^{(2)})^2 + x^{(3)} (x^{(3)})^2 + x^{(4)} (x^{(4)})^2 + x^{(5)} (x^{(5)})^2 \\ &= (x^{(1)})^3 + (x^{(2)})^3 + (x^{(3)})^3 + (x^{(4)})^3 + (x^{(5)})^3 \end{aligned}$$

$$= \sum_{i=1}^s (x^{(i)})^{2+3-2}$$

$$\phi^\top(x)\phi(x) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^2 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 \\ 1 & x^{(2)} & (x^{(2)})^2 & (x^{(2)})^3 \\ 1 & x^{(3)} & (x^{(3)})^2 & (x^{(3)})^3 \\ 1 & x^{(4)} & (x^{(4)})^2 & (x^{(4)})^3 \\ 1 & x^{(5)} & (x^{(5)})^2 & (x^{(5)})^3 \end{pmatrix}$$

$$\begin{aligned} (\phi^\top(x)\phi(x))_{2,3} &= x^{(1)} (x^{(1)})^2 + x^{(2)} (x^{(2)})^2 + x^{(3)} (x^{(3)})^2 + x^{(4)} (x^{(4)})^2 + x^{(5)} (x^{(5)})^2 \\ &= (x^{(1)})^3 + (x^{(2)})^3 + (x^{(3)})^3 + (x^{(4)})^3 + (x^{(5)})^3 \end{aligned}$$

$$= \sum_{i=1}^s (x^{(i)})^{2+3-2} \quad \longrightarrow \quad (\phi(x)^\top \phi(x))_{j,k} = \sum_{i=1}^s (x^{(i)})^{j+k-2}$$

$$\phi^T(x)y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^2 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \\ y^{(5)} \end{pmatrix}$$



$$\phi^T(x)y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^2 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \\ y^{(5)} \end{pmatrix}$$

$$(\phi^T(x)y)_3 = (x^{(1)})^2 y^{(1)} + (x^{(2)})^2 y^{(2)} + (x^{(3)})^2 + y^{(3)} + (x^{(4)})^2 y^{(4)} + (x^{(5)})^2 y^{(5)}$$



$$\phi^T(x)y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^2 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \\ y^{(5)} \end{pmatrix}$$

$$(\phi^T(x)y)_3 = (x^{(1)})^2 y^{(1)} + (x^{(2)})^2 y^{(2)} + (x^{(3)})^2 + y^{(3)} + (x^{(4)})^2 y^{(4)} + (x^{(5)})^2 y^{(5)}$$

$$= \sum_{i=1}^s (x^{(i)})^{3-1} y^{(i)}$$

$$\phi^T(x)y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} \\ (x^{(1)})^2 & (x^{(2)})^2 & (x^{(3)})^2 & (x^{(4)})^2 & (x^{(5)})^2 \\ (x^{(1)})^3 & (x^{(2)})^2 & (x^{(3)})^3 & (x^{(4)})^3 & (x^{(5)})^3 \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \\ y^{(5)} \end{pmatrix}$$

$$(\phi^T(x)y)_3 = (x^{(1)})^2 y^{(1)} + (x^{(2)})^2 y^{(2)} + (x^{(3)})^2 + y^{(3)} + (x^{(4)})^2 y^{(4)} + (x^{(5)})^2 y^{(5)}$$

$$= \sum_{i=1}^s (x^{(i)})^{3-1} y^{(i)} \quad \longrightarrow \quad (\phi(x)^T y)_j = \sum_{i=1}^s (x^{(i)})^{j-1} y^{(i)}$$



$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$



$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)} = \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{1}{2} \\ 1 & \frac{3}{4} \\ 1 & 1 \end{pmatrix}$$

$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix}$$

$$\phi(x)_{(d=1)} = \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{1}{2} \\ 1 & \frac{3}{4} \\ 1 & 1 \end{pmatrix}$$



$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix} \quad \phi(x)_{(d=1)} = \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{1}{2} \\ 1 & \frac{3}{4} \\ 1 & 1 \end{pmatrix} \quad (\phi(x)^\top \phi(x))_{j,k} = \sum_{i=1}^s (x^{(i)})^{j+k-2}$$



$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix} \quad \phi(x)_{(d=1)} = \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{1}{2} \\ 1 & \frac{3}{4} \\ 1 & 1 \end{pmatrix} \quad (\phi(x)^\top \phi(x))_{j,k} = \sum_{i=1}^s (x^{(i)})^{j+k-2}$$

$$\phi(x)_{(d=1)}^\top \phi(x)_{(d=1)} = \begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix}$$

$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$



$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)}^\top y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}$$

$$\begin{aligned} (x^{(1)}, y^{(1)}) &= (0,0) & (x^{(2)}, y^{(2)}) &= \left(\frac{1}{4}, 1\right) & (x^{(3)}, y^{(3)}) &= \left(\frac{1}{2}, 0\right) & (x^{(4)}, y^{(4)}) &= \left(\frac{3}{4}, -1\right) \\ (x^{(5)}, y^{(5)}) &= (1,0) \end{aligned}$$

$$\phi(x)_{(d=1)}^\top y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}$$

$$\phi(x)_{(d=1)}^\top y = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \quad \longrightarrow \quad w_0 = -\frac{w_1}{2}$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \quad \longrightarrow \quad w_0 = -\frac{w_1}{2}$$

$$\frac{5}{2}w_0 + \frac{15}{8}w_1 = -\frac{1}{2}$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \quad \longrightarrow \quad w_0 = -\frac{w_1}{2}$$

$$\frac{5}{2}w_0 + \frac{15}{8}w_1 = -\frac{1}{2} \quad \longrightarrow \quad -10w_1 + 15w_1 = -4$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \longrightarrow w_0 = -\frac{w_1}{2}$$

$$\frac{5}{2}w_0 + \frac{15}{8}w_1 = -\frac{1}{2} \longrightarrow -10w_1 + 15w_1 = -4 \longrightarrow$$

$$w_1 = -\frac{4}{5}$$



$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \longrightarrow w_0 = -\frac{w_1}{2}$$

$$\frac{5}{2}w_0 + \frac{15}{8}w_1 = -\frac{1}{2} \longrightarrow -10w_1 + 15w_1 = -4 \longrightarrow$$

$$w_1 = -\frac{4}{5}$$

$$w_0 = -\frac{w_1}{2}$$

$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}$$

$$5w_0 + \frac{5}{2}w_1 = 0 \longrightarrow w_0 = -\frac{w_1}{2}$$

$$\frac{5}{2}w_0 + \frac{15}{8}w_1 = -\frac{1}{2} \longrightarrow -10w_1 + 15w_1 = -4 \longrightarrow$$

$$w_1 = -\frac{4}{5}$$

$$w_0 = -\frac{w_1}{2}$$

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \frac{2}{5} \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$