MTH786, Semester A, 2023/2024
Solutions of coursework 2 **N. Perra**

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 $\left(x^{(1)}, y^{(1)}\right) = (-2, 1)$, $\left(x^{(2)}, y^{(2)}\right) = (-1, 2)$, $\left(x^{(3)}, y^{(3)}\right) = (0, 3)$, $\left(x^{(4)}, y^{(4)}\right) = (1, 4)$.

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

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

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 $\left(x^{(5)}, y^{(5)}\right) = (2, 20)$ is added. Evaluate new error measure and corresponding minimiser.

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

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

Some help: $19^2 = 361, 18^2 = 324, 17^2 = 289, 16^2 = 256, 15^2 = 225, 14^2 = 196, 13^2 = 169$.

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

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

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

Solutions:

1. • Filling the missing entries for the one-parameter MSE yields the following table:

	$w^{(0)}=1$	$w^{(0)}=2$	$w^{(0)}=3$	$w^{(0)}=4$	$w^{(0)}=5$	$w^{(0)}=6$	$w^{(0)}=7$
$y^{(1)}=1$	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
$\mathrm{MSE}(w)\cdot 2s$	14	6	6	14	30	54	86

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

- 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
MSE$(w)\cdot 2s$	375	330	295	270	255	250	255

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

2. - Filling the missing entries for the one-parameter MAE yields 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	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
MAE$(w)\cdot s$	6	4	4	6	10	14	18

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

   - 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
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. 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.

Solutions:

1. The likelihood is given by

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

2. Based on the solution of the previous exercise, the negative log-likelihood is given by

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

The partial derivative of (1) with respect to μ therefore is

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

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

3. The derivative of (1) with respect to σ^2 is

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

Setting this expression to zero and replacing the unknown quantity μ by the estimate $\hat{\mu}$, we obtain $\hat{\sigma}^2 = \frac{1}{s}\sum_{n=1}^{s}(x_n - \hat{\mu})^2$.

4. From the chain rule we observe $-\nabla \log(\rho(w)) = -\frac{1}{\rho(w)}\nabla\rho(w) = -\frac{\nabla\rho(w)}{\rho(w)}$, or

$$\nabla\rho(w) = \rho(w)\nabla\log(\rho(w)).$$

Hence, if $\nabla\log(\rho(w)) = 0$ is satisfied, we automatically observe $\nabla\rho(w) = 0$.

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

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

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

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

Solutions: 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

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

The gradient of MSE is a vector of partial derivatives of the form $\frac{\partial}{\partial w^{(p)}{}_q}\mathbf{W}$. These derivatives can be evaluated by using the chain and the sum rules as follows

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

The derivative $\frac{\partial}{\partial w_q^{(p)}} w_j^{(k)}$ is equal to either 0 or 1. And it is equal to 1 if and only if $p = k$, $j = q$. Thus,

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

The gradient $\nabla MSE(\mathbf{W}^*)$ is a vector of all partial derivatives, but if written in a matrix form this can be represented as

$$\nabla MSE(\mathbf{W}^*) = \frac{1}{s}\left(\mathbf{X}^{\top}\mathbf{X}\mathbf{W} - \mathbf{X}^{\top}\mathbf{Y}\right).$$

Problem 4. Compute the solution of the polynomial regression problem

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

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. $d = 3$.

Solutions: From the lecture notes we know that the solution of (2) can be computed by solving the normal equations

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

for

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

We further compute

$$\left(\Phi(\mathbf{X})^\top \Phi(\mathbf{X})\right)_{jk} = \sum_{i=1}^{s} \left(x^{(i)}\right)^{j+k-2} \qquad \text{and} \qquad \left(\Phi(\mathbf{X})^\top \mathbf{Y}\right)_j = \sum_{i=1}^{s} \left(x^{(i)}\right)^{j-1} y^{(i)},$$

for $j,k \in \{1,\dots,d{+}1\}$ and $i \in \{1,\dots,s\}$. Hence, for the given values $\{(x_i,y_i)\}_{i=1}^5$ we can compute $\hat{w}$ for $d=1$, $d=2$ and $d=3$ via

$$\begin{pmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{15}{8} \end{pmatrix} \hat{\mathbf{w}}_{d=1} = \begin{pmatrix} 0 \\ -\frac{1}{2} \end{pmatrix}, \qquad \begin{pmatrix} 5 & \frac{5}{2} & \frac{15}{8} \\ \frac{5}{2} & \frac{15}{8} & \frac{25}{16} \\ \frac{15}{8} & \frac{25}{16} & \frac{177}{128} \end{pmatrix} \hat{\mathbf{w}}_{d=2} = \begin{pmatrix} 0 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix},$$

and

$$\begin{pmatrix} 5 & \frac{5}{2} & \frac{15}{8} & \frac{25}{16} \\ \frac{5}{2} & \frac{15}{8} & \frac{25}{16} & \frac{177}{128} \\ \frac{15}{8} & \frac{25}{16} & \frac{177}{128} & \frac{325}{256} \\ \frac{25}{16} & \frac{177}{128} & \frac{325}{256} & \frac{2445}{2048} \end{pmatrix} \hat{\mathbf{w}}_{d=3} = \begin{pmatrix} 0 \\ -\frac{1}{2} \\ -\frac{1}{2} \\ -\frac{13}{32} \end{pmatrix}.$$

with solutions

$$\hat{\mathbf{w}}_{d=1} = \frac{1}{5}\begin{pmatrix} 2 \\ -4 \end{pmatrix} \quad \hat{\mathbf{w}}_{d=2} = \frac{1}{5}\begin{pmatrix} 2 \\ -4 \\ 0 \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{w}}_{d=3} = \frac{1}{3}\begin{pmatrix} 0 \\ 32 \\ -96 \\ 64 \end{pmatrix}.$$