WORKSHEET 2

PROBLEM 1

Find optimal parameter value, $\hat{\underline{w}}$, for

$$\mathcal{L} = \sum_{n=1}^{N} \left(t_n - \underline{w}^T \underline{x}_n \right)^2$$

Note: t_n is a scalar quantity

$$\underline{w} = (w_0, \ldots, w_D)^T$$

$$\underline{x} = (x_0, \ldots, x_D)^T$$

To find $\hat{\underline{w}}$, set $\frac{\partial \mathcal{L}}{\partial \underline{w}} = 0$

Find $\frac{\partial \mathcal{L}}{\partial \underline{w}}$ by application of the chain rule:

$$\frac{d}{dx}\left(f(g(x))\right) = \frac{d\, f(g(x))}{d\, g(x)} \cdot \frac{d\, g(x)}{dx}$$

or by substitution:

Let $u_n = t_n - \underline{w}^T \underline{x}_n$

$$\Rightarrow \frac{\partial u_n}{\partial \underline{w}} = -\underline{x}_n$$

$$\Rightarrow \mathcal{L} = \sum_{n=1}^{N} u_n^2$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial u} = 2 \sum_{n=1}^{N} u_n = 2 \sum_{n=1}^{N} \left(t_n - w^T x_n \right)$$

Then $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial u} \cdot \frac{\partial u}{\partial w} = 2 \sum_{n=1}^{N} (t_n - w^T x_n)(-x_n)$

Note that $w^T x_n = x_n^T w$

and $(t_n - w^T x_n)$ is a scalar

$$\Rightarrow \frac{\partial L}{\partial \underline{w}} = 2 \sum_{n=1}^{N} (t_n - x_n^T w)(-x_n)$$

$$= -2 \sum_{n=1}^{N} (x_n)(t_n - x_n^T w)$$

$$= -2 \left(\left(\sum_{n=1}^{N} x_n t_n \right) - \left(\sum_{n=1}^{N} x_n x_n^T w \right) \right)$$

Setting $\frac{\partial L}{\partial \underline{w}} = 0$

$$\Rightarrow \left(\sum_{n=1}^{N} x_n x_n^T \right) \hat{\underline{w}} = \sum_{n=1}^{N} x_n t_n$$

$$\hat{\underline{w}} = \left(\sum_{n=1}^{N} x_n x_n^T \right)^{-1} \sum_{n=1}^{N} x_n t_n$$

Note that in matrix / vector notation

$$L = (t - X\underline{w})^T (t - X\underline{w})$$

where $\underline{t} = (t_1, \ldots, t_N)^T \quad X = (\underline{x}_1, \ldots, \underline{x}_N)^T$

Then $\frac{\partial L}{\partial \underline{w}} = 2(-X)^T (t - X\underline{w})$

Setting $\frac{\partial L}{\partial \underline{w}} = 0 \Rightarrow X^T X \hat{\underline{w}} = X^T \underline{t}$

$$\hat{\underline{w}} = (X^T X)^{-1} X^T \underline{t}$$

as above.

$\hat{w}$ provides a critical point for L.

Use the second derivative test to check whether this critical point corresponds to a maximum or a minimum point

$$\frac{\partial^2 L}{\partial \underline{w}^2} = -2\left(-\sum_{n=1}^{N} x_n x_n^T\right) = 2\sum_{n=1}^{N} x_n x_n^T$$

$$= 2\,X^T X \quad .$$

So, if $X^T X$ is positive definite

(which as a sum of squares it will be)
when X is full rank, $N \geq M$

then $\hat{w}$ corresponds to a minimum

and the squared loss function L is

minimised at $\hat{w}$.

Scaling $\mathcal{L}$ by a constant will have no impact on the optimal parameter $\hat{w}$, thus the $\hat{w}$ remains unchanged for both $\mathcal{L}$ and $\frac{1}{N}\mathcal{L}$.

PROBLEM 2

Consider the weighted average loss

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \alpha_n \left(t_n - \underline{w}^T \underline{x}_n\right)^2$$

where t_n, $\underline{w}$, $\underline{x}_n$ are defined as in Problem 1, and the weight / importance of the n^{th} data point, $\{t_n, \underline{x}_n\}$ is given by α_n

Find the optimal least squares parameter estimate by setting $\frac{d\mathcal{L}}{d\underline{w}} = 0$

$$\frac{d\mathcal{L}}{d\underline{w}} = \frac{1}{N} \sum_{n=1}^{N} 2\alpha_n \left(t_n - \underline{w}^T \underline{x}_n\right)\left(-\underline{x}_n\right)$$

which as in problem 1

$$= -\frac{2}{N} \left(\sum_{n=1}^{N} \alpha_n \underline{x}_n t_n - \sum_{n=1}^{N} \alpha_n \underline{x}_n \underline{x}_n^T \underline{w} \right)$$

setting $\frac{d\mathcal{L}}{d\underline{w}} = 0 \Rightarrow \hat{\underline{w}} = \dfrac{\sum_{n=1}^{N} \alpha_n \underline{x}_n t_n}{\sum_{n=1}^{N} \alpha_n \underline{x}_n \underline{x}_n^T}$

or in matrix notation

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \alpha_n \left(t_n - w^T x_n\right)^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left(t_n - w^T \underline{x}_n\right) \alpha_n \left(t_n - w^T \underline{x}_n\right)$$

$$= \frac{1}{N} \left(\underline{t} - Xw\right)^T A \left(\underline{t} - Xw\right)$$

where $A = \begin{pmatrix} \alpha_1 & & 0 \\ & \alpha_2 & \\ 0 & & \ddots \\ & & & \alpha_N \end{pmatrix}$ is the diagonal matrix of observation weights.

then $$\frac{d\mathcal{L}}{dw} = \frac{1}{N} \cdot 2 \cdot (-X)^T \cdot A \cdot \left(\underline{t} - Xw\right)$$

$$= -\frac{2}{N} \left(X^T A \underline{t} - X^T A X w\right)$$

$$\Rightarrow \quad \hat{w} = \left(X^T A X\right)^{-1} X^T A \underline{t}$$

as above.

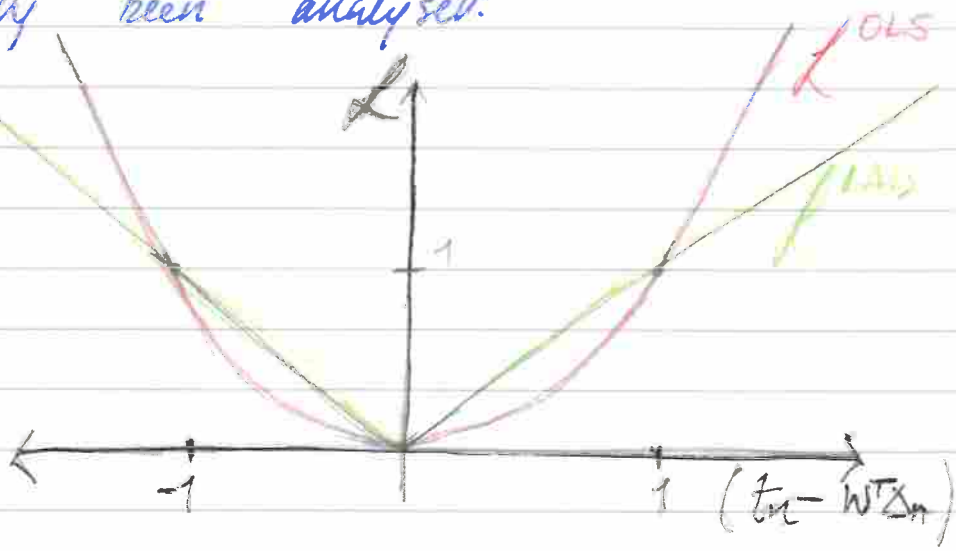
Note: For a guide to matrix / vector calculus see the Matrix Cookbook, for which there is a link on the module webpage.

PROBLEM 3

Consider the loss function

$$\mathcal{L}^{LAD} = \sum_{n=1}^{N} |t_n - w^T x_n|$$

versus the $\mathcal{L}^{OLS} = \sum_{n=1}^{N} (t_n - w^T x_n)^2$ which has already been analysed.



Examining the plot above illustrates that minimising $\mathcal{L}^{LAD}$ places equal weight on all observations, while minimising $\mathcal{L}^{OLS}$ places a greater weight on those observations for which there are large errors.

Thus $\mathcal{L}^{OLS}$ is more heavily influenced by outliers when estimating the optimal parameter values, making $\mathcal{L}^{LAD}$ more robust to these outliers

1

To find the optimal parameter values for L^{LAD} set $\frac{\partial L^{LAD}}{\partial \underline{w}} = 0$

For $t_n - \underline{w}^T \underline{x}_n > 0$

$$L^{LAD} = t_n - \underline{w}^T \underline{x}_n$$

$$\frac{\partial L^{LAD}}{\partial \underline{w}} = -\underline{x}_n \neq 0$$

For $t_n - \underline{w}^T \underline{x}_n < 0$

$$L^{LAD} = -(t_n - \underline{w}^T \underline{x}_n)$$

$$\frac{\partial L^{LAD}}{\partial \underline{w}} = \underline{x}_n \neq 0$$

For $t_n - \underline{w}^T \underline{x}_n = 0$

$$L^{LAD} = \mp (t_n - \underline{w}^T \underline{x}_n)$$

$$\frac{\partial L^{LAD}}{\partial \underline{w}} = \mp \underline{x}_n \neq 0$$

Thus L^{LAD} has no critical points and there is no analytic solution for optimal parameter values.