Honours Math for Machine Learning HW1

Saina Koukpari - McGill University

Problem 1.1

(1) We set the derivative of $\hat{L}$ to zero to find the critical points: $\hat{L}'(w) = \frac{1}{m}\sum_{i=1}^{m} q'(e) = 0$ where $q'(e) = (wx_i - y_i)x_i$ for $X = [1,2,3]^T$ and $Y = [2,4,5]^T$. Rearranging and plugging in the variables we have,

$$\begin{aligned}
\hat{L}'(w) &= \frac{1}{m}\sum_{i=1}^{m}(wx_i - y_i)x_i \\
&= \frac{1}{m}\sum_{i=1}^{m} wx_i^2 - y_i x_i \\
&= \frac{1}{3}(w - 2 + 4w - 8 + 9w - 15) \\
&= \frac{1}{3}(14w - 25) = 0 \\
\implies w &= 25/14
\end{aligned}$$

Hence the critical point is at $w = 25/14$ and

$$\hat{L}(25/14) = \frac{1}{3}\left(\frac{(25/14 - 2)^2}{2} + \frac{(2(25/14) - 4)^2}{2} + \frac{(3(25/14) - 5)^2}{2}\right) = 35/196$$

(2) $\hat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(wx_i - y_i) = \frac{1}{2}(|w-1| + |2w-3|)$ given that $X = \{1,2\}$, $Y = \{1,2\}$ and $h = wx$. We can plot the piece-wise function by the following three cases,

1. $w \le 1 \implies (1 - w) + (3 - 2w) = 4 - 3w$
2. $1 < w \le 3/2 \implies (w - 1) + (3 - 2w) = 2 - w$
3. $w > 3/2 \implies (w - 1) + (2w - 3) = 3w - 4$

From this we plot the graph,

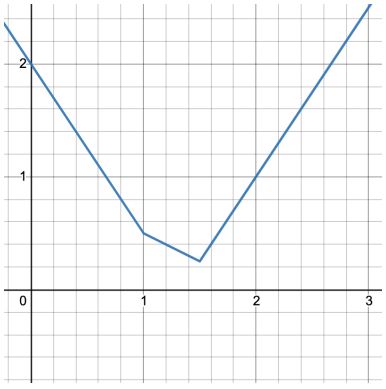

Figure 1: Graphical Representation of $\hat{L}(w)$

where the minimum is at $(3/2, 1/4)$. So $\ell_1 = \frac{1}{2}(|(3/2) - 1| + |2(3/2) - 3|) = 1/2(1/2) = 1/40$.

(3) Given $\hat{L}(w) = \frac{1}{m}\sum_i |w - y_i|$ for case 1 with $y = [2, 4, 5]^T$ and case 2 with $y = [3, 6]$, we can find the piece-wise functions as follows,

Case 1:

1. $w \leq 2 \implies (2 - w) + (4 - w) + (5 - w) = 11 - 3w$
2. $2 < w \leq 4 \implies (w - 2) + (4 - w) + (5 - w) = 7 - w$
3. $4 < w \leq 5 \implies (w - 2) + (w - 4) + (5 - w) = w - 1$
4. $5 < w \implies (w - 2) + (w - 4) + (w - 5) = 3w - 11$

Case 2:

1. $w \leq 3 \implies (3 - w) + (6 - w) = 9 - 2w$
2. $3 < w \leq 6 \implies (w - 3) + (6 - w) = 3$
3. $6 < w \implies (w - 3) + (w - 6) = 2w - 9$

Then we obtain the following graphs respectively,

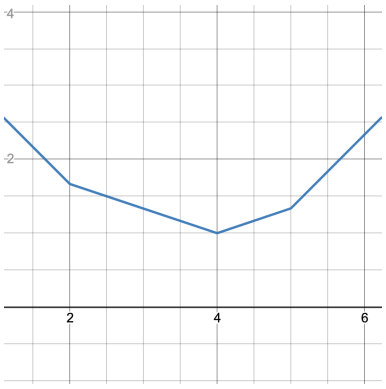

Figure 2: Graphical Representation of $\hat{L}(w)$ for case $y = [2, 4, 5]^T$

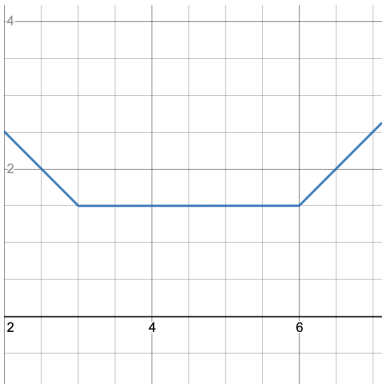

Figure 3: Graphical Representation of $\hat{L}(w)$ for case $y = [3, 6]$

From the above we find that the minimizer for case 1 is $w = 4$ and the minimizers for case 2 are $w \in [3, 6]$.

(4) i) the data matrix F is a 11×2 matrix given by the data points of $f(1, x)$. Then we have

$$\begin{bmatrix} 1 & 0.0 \\ 1 & 0.1 \\ 1 & 0.2 \\ 1 & 0.3 \\ 1 & 0.4 \\ 1 & 0.5 \\ 1 & 0.6 \\ 1 & 0.7 \\ 1 & 0.8 \\ 1 & 0.9 \\ 1 & 1 \end{bmatrix}$$

From this we know that the size of F^T is 2×11 so the size of $F^T F$ is 2×2. Using the matrix equation to solve for w where $F^T F w = F^T y$ for y of size 11×1, the size of w is found to be 2×1.

Writing a program (in python) to generate random values of y, I found the following,

Code:

```
import random
import math

x = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
y = [round((math.sin(2*math.pi * x[i]) + random.random()*0.1), 2) for i in range(11)]

print(x) print(y)
```

Output:

```
[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
[0.0, 0.65, 0.95, 1.02, 0.64, 0.02, -0.57, -0.92, -0.95, -0.52, 0.09]
```

Using the above values as an example we can solve for w and plot $y = wx$.

$$\begin{aligned} w &= (F^T F)^{-1} F^T y \\ &= \left(\begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 0.0 & 0.1 & \dots & 0.9 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.0 \\ 1 & 0.1 \\ 1 & 0.2 \\ 1 & 0.3 \\ 1 & 0.4 \\ 1 & 0.5 \\ 1 & 0.6 \\ 1 & 0.7 \\ 1 & 0.8 \\ 1 & 0.9 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 0.0 & 0.1 & \dots & 0.9 & 1 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.65 \\ 0.95 \\ 1.02 \\ 0.64 \\ 0.02 \\ -0.57 \\ -0.92 \\ -0.95 \\ -0.52 \\ 0.09 \end{bmatrix} \\ &= \left(\begin{bmatrix} 11 & 5.5 \\ 5.5 & 3.85 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.41 \\ -1.297 \end{bmatrix} \\ &= \begin{bmatrix} 0.318 & -0.454 \\ -0.454 & 0.909 \end{bmatrix} \begin{bmatrix} 0.41 \\ -1.297 \end{bmatrix} \\ &= \begin{bmatrix} 0.719 \\ -1.365 \end{bmatrix} \end{aligned}$$

$$y = Fw = \begin{bmatrix} 1 & 0.0 \\ 1 & 0.1 \\ 1 & 0.2 \\ 1 & 0.3 \\ 1 & 0.4 \\ 1 & 0.5 \\ 1 & 0.6 \\ 1 & 0.7 \\ 1 & 0.8 \\ 1 & 0.9 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.719 \\ -1.365 \end{bmatrix} = \begin{bmatrix} 0.719 \\ 0.5825 \\ 0.446 \\ 0.3095 \\ 0.173 \\ 0.0365 \\ -0.1 \\ -0.2365 \\ -0.373 \\ -0.5095 \\ -0.646 \end{bmatrix}$$

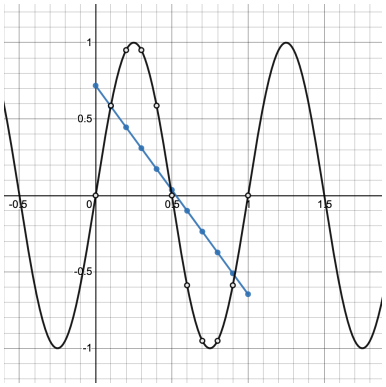

Figure 4: Graphical Representation of $y = Fw$ for $f(x) = (1, x)$

The above shows the data with error in blue and the solution $y = \sin(2\pi x)$ in black. It is not a good fit.

ii) the data matrix F is a 11×4 matrix given by the data points of $f(1, x, x^2, x^3)$. Then we have

$$\begin{bmatrix} 1 & 0.0 & 0.0 & 0.0 \\ 1 & 0.1 & 0.01 & 0.001 \\ 1 & 0.2 & 0.04 & 0.008 \\ 1 & 0.3 & 0.09 & 0.027 \\ 1 & 0.4 & 0.16 & 0.064 \\ 1 & 0.5 & 0.25 & 0.125 \\ 1 & 0.6 & 0.36 & 0.216 \\ 1 & 0.7 & 0.49 & 0.343 \\ 1 & 0.8 & 0.64 & 0.512 \\ 1 & 0.9 & 0.81 & 0.729 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

From this we know that the size of F^T is 4×11 so the size of $F^T F$ is 4×4. Using the matrix equation to solve for w where $F^T Fw = F^T y$ for y of size 11×1, the size of w is found to be 4×1.

Using the same values generated in (i), we find w to be,

$$
\begin{aligned}
w &= (F^T F)^{-1} F^T y \\
&= \left(\begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 0.0 & 0.1 & \dots & 0.9 & 1 \\ 0.0 & 0.01 & \dots & 0.81 & 1 \\ 0.0 & 0.001 & \dots & 0.729 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.0 & 0.0 & 0.0 \\ 1 & 0.1 & 0.01 & 0.001 \\ 1 & 0.2 & 0.04 & 0.008 \\ 1 & 0.3 & 0.09 & 0.027 \\ 1 & 0.4 & 0.16 & 0.064 \\ 1 & 0.5 & 0.25 & 0.125 \\ 1 & 0.6 & 0.36 & 0.216 \\ 1 & 0.7 & 0.49 & 0.343 \\ 1 & 0.8 & 0.64 & 0.512 \\ 1 & 0.9 & 0.81 & 0.729 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 0.0 & 0.1 & \dots & 0.9 & 1 \\ 0.0 & 0.01 & \dots & 0.81 & 1 \\ 0.0 & 0.001 & \dots & 0.729 & 1 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.65 \\ 0.95 \\ 1.02 \\ 0.64 \\ 0.02 \\ -0.57 \\ -0.92 \\ -0.95 \\ -0.52 \\ 0.09 \end{bmatrix} \\
&= \left(\begin{bmatrix} 11 & 5.5 & 3.85 & 3.025 \\ 5.5 & 3.85 & 3.025 & 2.5333 \\ 3.85 & 3.025 & 2.5333 & 2.20825 \\ 3.025 & 2.5333 & 2.20825 & 1.978405 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.41 \\ -1.297 \\ -1.3515 \\ -1.13491 \end{bmatrix} \\
&= \begin{bmatrix} 0.7902 & -5.53613 & 10.48951 & -5.8275 \\ -5.53613 & 65.527065 & -150.5439 & 92.59259 \\ 10.48951 & -150.5439 & 375.874125 & -242.812742 \\ -5.8275 & 92.59259 & -242.812742 & 161.875161 \end{bmatrix} \begin{bmatrix} 0.41 \\ -1.297 \\ -1.3515 \\ -1.13491 \end{bmatrix} \\
&= \begin{bmatrix} -0.05853146 \\ 11.1174048 \\ -32.867132 \\ 21.965811 \end{bmatrix}
\end{aligned}
$$

$$
y = Fw = \begin{bmatrix} 1 & 0.0 & 0.0 & 0.0 \\ 1 & 0.1 & 0.01 & 0.001 \\ 1 & 0.2 & 0.04 & 0.008 \\ 1 & 0.3 & 0.09 & 0.027 \\ 1 & 0.4 & 0.16 & 0.064 \\ 1 & 0.5 & 0.25 & 0.125 \\ 1 & 0.6 & 0.36 & 0.216 \\ 1 & 0.7 & 0.49 & 0.343 \\ 1 & 0.8 & 0.64 & 0.512 \\ 1 & 0.9 & 0.81 & 0.729 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -0.05853146 \\ 11.1174048 \\ -32.867132 \\ 21.965811 \end{bmatrix} = \begin{bmatrix} -0.05853146 \\ 0.746503511 \\ 1.025990708 \\ 0.911724997 \\ 0.535501244 \\ 0.029114315 \\ -0.475640924 \\ -0.846969607 \\ -0.953076868 \\ -0.662167841 \\ 0.15755234 \end{bmatrix}
$$

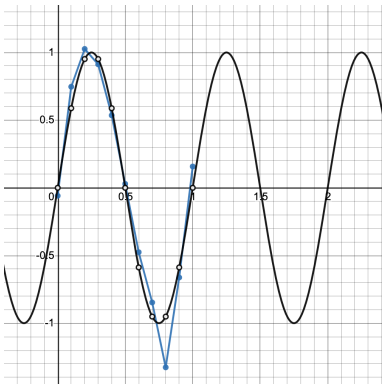

Figure 5: Graphical Representation of $y = Fw$ for case $f(x) = (1, x, x^2, x^3)$

We see in Figure 5 the data with error in blue and the solution $y = \sin(2\pi x)$ in black. It is a much better fit than Figure 4.

(5) i) We have

$$\hat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(wx_i - y_i) = \frac{1}{m}\sum_{i=1}^{m} \frac{(wx_i - y_i)^2}{2} = \frac{1}{3}\left(\frac{(3w_1 - 6)^2}{2} + \frac{(2w_2 - 2)^2}{2} + \frac{(w_1 + w_2 - 5)^2}{2}\right)$$

Taking the derivatives, we find the gradient to be,

$$\begin{aligned}
\nabla \hat{L} &= \left[\frac{\partial \hat{L}}{\partial w_1}, \frac{\partial \hat{L}}{\partial w_2}\right] \\
&= \left[\frac{1}{3}\left(\frac{2 \cdot 3(3w_1 - 6)}{2} + \frac{2(w_1 + w_2 - 5)}{2}\right), \frac{1}{3}\left(\frac{2 \cdot 2(2w_2 - 2)}{2} + \frac{2(w_1 + w_2 - 5)}{2}\right)\right] \\
&= \left[\frac{1}{3}\left(3(3w_1 - 6) + (w_1 + w_2 - 5)\right), \frac{1}{3}\left(2(2w_2 - 2) + 2(w_1 + w_2 - 5)\right)\right] \\
&= \left[\frac{1}{3}\left(10w_1 + w_2 - 23)\right), \frac{1}{3}\left(w_1 + 5w_2 - 9\right)\right]
\end{aligned}$$

Setting the above equal to zero, we obtain the system of equations,

$$\begin{aligned}
& 10w_1 + w_2 - 23 = 0; \;\; w_1 + 5w_2 - 9 = 0 \\
\implies & w_1 = 9 - 5w_2 \\
\implies & 10(9 - 5w_2) + w_2 - 23 = 67 - 49w_2 \\
\implies & w_2 = 67/49 \\
\implies & w_1 = 9 - 5(67/49) = 106/49
\end{aligned}$$

ii) Setting up the matrix equation we obtain,

$$\begin{aligned} w &= (X^T X)^{-1} X^T y \\ &= \left(\begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 2 \\ 5 \end{bmatrix} \\ &= \left(\begin{bmatrix} 10 & 1 \\ 1 & 5 \end{bmatrix} \right)^{-1} \begin{bmatrix} 23 \\ 9 \end{bmatrix} \\ &= \frac{1}{49} \begin{bmatrix} 5 & -1 \\ -1 & 10 \end{bmatrix} \begin{bmatrix} 23 \\ 9 \end{bmatrix} \\ &= \frac{1}{49} \begin{bmatrix} 106 \\ 67 \end{bmatrix} = \begin{bmatrix} 106/49 \\ 67/49 \end{bmatrix} \end{aligned}$$

(6) The 10% Winsorized mean with 10 numbers is given by,

$$\frac{y_2 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_9}{10}$$

Choosing all data points equal to zero we have the Winsorized mean equal to zero and the minimizer of the Huber loss equal to zero.

The main difference between the Winsorized mean and the minimizer of the Huber loss is that the Huber loss has a scale δ which determines the outliers, but the Winsorized mean has a fraction of values :) Hence the Winsorized mean is more sensitive to outliers.

(7) For linear regression loss $\ell(w, y_i) = |w - y_i|$,

$$\frac{1}{10} \sum_{i=1}^{10} |w - y_i| = \frac{1}{10} (9|w| + |w - y|)$$

For quadratic loss $\ell(w, y_i) = \frac{1}{2}(w - y_i)^2$,

$$\frac{1}{10} \sum_{i=1}^{10} \frac{1}{2} (w - y_i)^2 = \frac{1}{10} \left(\frac{(9w)^2}{2} + \frac{(w - y)^2}{2} \right) = \frac{(9w)^2 + (w - y)^2}{20}$$

And for Huber loss we have

$$\ell(e) = \begin{cases} \frac{1}{10}(9|w| + |w - y|) & |y| \le 1 \\ \frac{1}{10}(9|w| + |w - y| - 1) & |y| \ge 1 \end{cases}$$

Problem 1.2

(1) For theoretical **x**:

$$\begin{aligned} w &= (X^T X)^{-1} X^T y \\ &= \left(\begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ &= \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{x_1 x_1 + x_2 x_2 + \cdots + x_n x_n} \end{aligned}$$

Taking (1.1.1) as an example, we have $X = [1,2,3]^T$ and $Y = [2,4,5]^T$. Solving for w using vector notion we obtain,

$$\begin{aligned} w &= (X^T X)^{-1} X^T y \\ &= \left(\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix} \\ &= \left(\begin{bmatrix} 14 \end{bmatrix} \right)^{-1} \begin{bmatrix} 25 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{14} \end{bmatrix} \begin{bmatrix} 25 \end{bmatrix} = 25/14 \end{aligned}$$

And the value from our equation is,

$$\frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{x_1 x_1 + x_2 x_2 + \cdots + x_n x_n} = \frac{2 + 8 + 15}{1 + 4 + 9} = \frac{25}{149}$$

(2) We wish to minimize error ϵ where $y = wX + \epsilon \implies \epsilon = wX - y$. Taking the sum of squared errors we have $\min \epsilon^2 = \min \|wX - y\|^2$. Then using the matrix theory fact we derive as follows,

$$\begin{aligned} \|wX - y\|^2 &= (wX - y)^T (wX - y) \\ &= ((wX)^T - y^T)(wX - y) \\ &= (wX)^T (wX) - 2wX^T y + y^T y \\ &\quad \text{taking the derivative to find minimum:} \\ &\quad 2(X^T X)w - 2X^T y = 0 \\ \implies & (X^T X)w = X^T y \end{aligned}$$

(3) We know that both $e^2/2$ and $\delta(|e| - \delta/2))$ are continuous functions where they are defined. It remains to show that the Huber loss is continuous where the former changes to the latter function. Suppose we approach $+\delta$, i.e. from $e^2/2$ going towards $\delta(|e| - \delta/2))$. Then,

$$\lim_{e \to \delta^-} e^2/2 = \delta^2/2 = \delta(\delta - \delta/2) = \lim_{e \to \delta^+} \delta(|e| - \delta/2))$$

By symmetry, the same applies for the point at $-\delta$, so the Huber loss function is continuous at all points.

To check differentiability, we have that for h positive,

$$\begin{aligned} \lim_{h \to 0} \frac{q(\delta + h) - q(\delta)}{h} &= \lim_{h \to 0} \frac{\delta(|\delta + h| - \delta/2)) - \delta^2/2}{h} \\ &= \lim_{h \to 0} \frac{\delta|\delta + h| - \delta^2/2 - \delta^2/2}{h} \\ &= \lim_{h \to 0} \frac{\delta^2 + \delta h - \delta^2}{h} \\ &= \delta \end{aligned}$$

Therefore, the Huber loss is differentiable since the above exists.

The second derivative of the Huber loss is found as follows,

$$q'(e) = \begin{cases} e & |e| \le \delta \\ \delta * \text{sign}(e) & |e| \ge \delta \end{cases}$$

$$q''(e) = \begin{cases} 1 & |e| \le \delta \\ 0 & |e| \ge \delta \end{cases}$$

(5) A function is convex iff $f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2)$ for $0 \le t \le 1$.

The linear loss, $\ell(x_i, y_i) = |x_i - y_i|$, is convex since taking $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x_2} = (x_2, y_2)$, we have

$$\begin{aligned} \ell(t(x_1, y_1) + (1-t)(x_2, y_2)) &= \ell(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \\ &= |tx_1 + (1-t)x_2 - ty_1 - (1-t)y_2| \\ &= |t(x_1 - y_1) + (1-t)(x_2 - y_2)| \\ &\le t|x_1 - y_1| + (1-t)|x_2 - y_2| \quad \text{(triangle inequality)} \\ &= t\ell(x_1, y_1) + (1-t)\ell(x_2, y_2) \quad \square \end{aligned}$$

The quadratic loss, $\ell(x_i, y_i) = \frac{1}{2}(x_i - y_i)^2$, is convex since taking $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x_2} = (x_2, y_2)$, we have

$$\begin{aligned} \ell(t(x_1, y_1) + (1-t)(x_2, y_2)) &= \ell(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \\ &= \frac{1}{2}\left(tx_1 + (1-t)x_2 - ty_1 - (1-t)y_2\right)^2 \\ &= \frac{1}{2}\left(t(x_1 - y_1) + (1-t)(x_2 - y_2)\right)^2 \\ &\le \frac{1}{2}(t(x_1 - y_1))^2 + ((1-t)(x_2 - y_2))^2 \quad \text{(triangle inequality)} \\ &\le \frac{1}{2}t((x_1 - y_1))^2 + (1-t)((x_2 - y_2))^2 \quad \text{(because for } 0 \le t \le 1,\ t^2 \le t) \\ &= t\ell(x_1, y_1) + (1-t)\ell(x_2, y_2) \quad \square \end{aligned}$$

Finally, taking $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x_2} = (x_2, y_2)$, we show that the Huber loss is convex.

$$\begin{aligned} \ell(t(x_1, y_1) + (1-t)(x_2, y_2)) &= \ell(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \\ &= \begin{cases} (tx_1 + (1-t)x_2 - ty_1 - (1-t)y_2))^2/2 & |e| \le \delta \\ \delta(|tx_1 + (1-t)x_2 - ty_1 - (1-t)y_2)| - \delta/2) & |e| \ge \delta \end{cases} \end{aligned}$$

When $|e| \le \delta$, the function is convex as shown in the case for quadratic loss. Now consider the case where $|e| \ge \delta$,

$$
\begin{aligned}
\delta(|tx_1 + (1-t)x_2 - ty_1 - (1-t)y_2)| - \delta/2) &= \delta(|t(x_1 - y_1) - (1-t)(x_2 - y_2)| - \delta/2) \\
&= \delta|t(x_1 - y_1) - (1-t)(x_2 - y_2)| - \delta^2/2 \\
&\leq \delta(|t(x_1 - y_1)| - |(1-t)(x_2 - y_2)|) - \delta^2/2 \quad \text{(triangle inequality)} \\
&= \delta|t(x_1 - y_1)| - \delta|(1-t)(x_2 - y_2)| - \delta^2/2 \\
&= \delta|t(x_1 - y_1)| - \delta|(1-t)(x_2 - y_2)| - \delta^2/2 + \frac{1}{2}t\delta^2 - \frac{1}{2}t\delta^2 \\
&= \delta t|x_1 - y_1| - \frac{1}{2}t\delta^2 + \delta(1-t)|x_2 - y_2| - \frac{1}{2}(1-t)\delta^2 \\
&= t\Big(\delta(|x_1 - y_1| - \frac{1}{2}\delta)\Big) + (1-t)\Big(\delta(|x_2 - y_2| - \frac{1}{2}\delta)\Big) \\
&= t\ell(x_1, y_1) + (1-t)\ell(x_2, y_2) \quad \square
\end{aligned}
$$

Problem 1.3

(1) i) In order to build a scheme that is less sensitive to outliers, we could use the Huber loss function with δ equal to the maximum effect we want any outlier grades to have on our result. Finding the median (w^*) of the grades, we can find the inliers that are within δ of w^* and the outlier that are not within δ of w^*, to then apply the corresponding functions of Huber loss.

ii) Using Huber loss with $\delta = 0.2$ we have,

$$
\ell(e) \begin{cases} 1/2e^2 & |e| \leq 0.2 \\ 0.2(|e| - 0.1) & |e| \geq 0.2 \end{cases}
$$

Using the example where the scores are $0.9, 0.9, 0.9, 0.9, 0.0$, we find the median $w^* = 0.9$ where $0.9 - 0.2 = 0.7 \in [0.9 - 0.2, 0.9 + 0.2]$ and $0.0 - 0.2 = -0.2 \notin [0.9 - 0.2, 0.9 + 0.2]$. Then the loss is,

$$
\begin{aligned}
\frac{1}{5}\sum_{i=1}^{5} \ell(e) &= \frac{1}{5}(4(1/2(w^* - 0.9)^2) + 0.2(|w^* - 0| - 0.1) \\
&= \frac{1}{5}(2(0.9 - 0.9)^2 + 0.2(0.9 - 0.1) \\
&= \frac{1}{5}(0.16) = 0.032
\end{aligned}
$$

The above has a loss close to zero with a roughly 3 percent error given the grading scale, hence the Huber loss is effectively used for this example.

(2) For the quadratic to be continuously differentiable, we must have $\lim_{t \to \delta^-} \ell(t) = \lim_{t \to \delta^+} \ell(t)$. So,

$$
\lim_{t \to \delta^-} |t| = |\delta| = \frac{\delta^2}{|\delta|} = \lim_{t \to \delta^+} \frac{t^2}{|\delta|}
$$

Then the flipped Huber loss function is given by,

$$
\ell(t) = \begin{cases} |t| & |t| \leq \delta \\ t^2/|\delta| & |t| \geq \delta \end{cases}
$$

*Remark: Note that the function won't be continuously differentiable simply by modifying the quadratic loss since $|t|$ is not differentiable at 0.