Contribution Report

Discussion and Solve Problems for Questions 1, 2, and 3:
Equal Contribution by both partners

Solutions Write-up for Questions 1, 2, and 3:
David Wu

Proof-reading for Questions 1, 2, and 3:
Shu Wang

Let $u, v \in \mathbb{R}^n$, $c \in R$, and y be any linear regression predictor.

Then: $y(u+v) = w^T(u+v)$

$$= \sum_i w_i (u+v)_i$$

$$= \sum_i w_i (u_i + v_i)$$

$$= \sum_i w_i u_i + w_i v_i$$

$$= (w_1 u_1 + w_1 v_1) + \dots + (w_n u_n + w_n v_n)$$

Rearranging the above gives:

$$y(u+v) = (w_1 u_1 + w_2 u_2 \dots + w_n u_n) + (w_1 v_1 + w_2 v_2 + \dots w_n v_n)$$

$$= \sum_i w_i u_i + \sum_i w_i v_i$$

$$= w^T u + w^T v$$

$$= y(u) + y(v).$$

Similarly, we have:

$$y(cv) = w^T(cv)$$

$$= \sum_i w_i (c v_i)$$

$$= w_1 c v_1 + w_2 c v_2 \dots w_n c v_n$$

Rearranging the above:

$$y(cv) = c w_1 v_1 + c w_2 v_2 \dots c w_n v_n$$

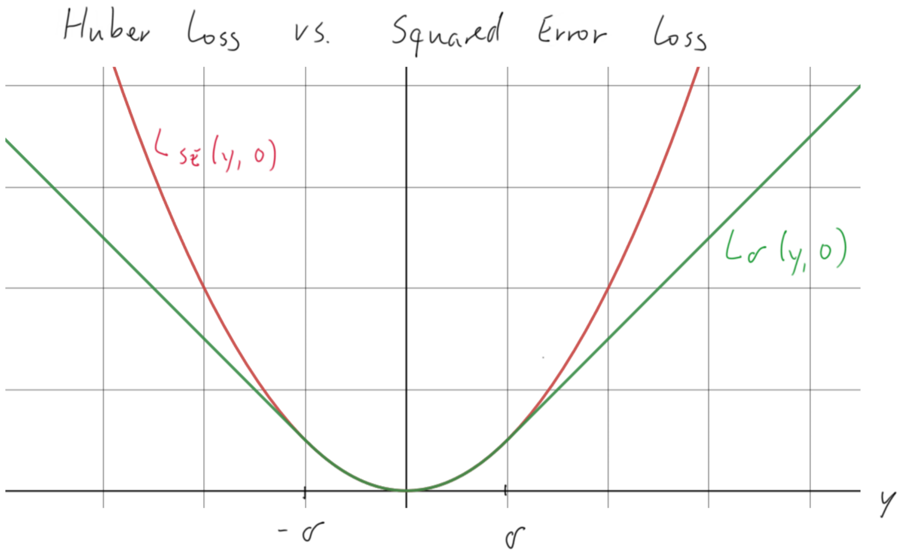
$$= c(w_1 v_1 + \dots + w_n v_n)$$

$$= c \sum w_i v_i$$

$$= c w^T v = c\, y(v)$$

Therefore all linear regression predictors are linear functions.

2. a.



Consider N data points with no outliers and y^* be the optimal predictor. Now consider additional outlier points $(x^{(i)}, t^{(i)})$. Since these points comes from a different conditional distribution, we expect that $y^*(x^{(i)}) - t^{(i)}$ to be large, meaning $\left|y^*\left(x^{(i)}\right) - t^{(i)}\right| > \delta$.

From the graphs, L_{SE} assigns a larger loss to $y^*(x^{(i)}) - t^{(i)}$ than L_δ. For a modelling fitting all the points (N + outliers), y^* is no longer the optimal predictor, but since we are trying to minimize average loss, a small value for $L(y^*(x^{(i)}) - t^{(i)})$ terms mean that we have to adjust y^* less. Therefore, the predictor under L_δ is affected by outliers less when compared to L_{SE}.

2.b.

$$H_\sigma(a) = \begin{cases} \sigma(-a - \frac{1}{2}\sigma) & a < -\sigma \\ \frac{1}{2}a^2 & -\sigma \leq a \leq \sigma \\ \sigma(a - \frac{1}{2}\sigma) & a > \sigma \end{cases} \Rightarrow H'_\sigma(a) = \begin{cases} -\sigma & a < -\sigma \\ a & |a| \leq \sigma \\ \sigma & a > \sigma \end{cases}$$

$$\vec{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} x^{(1)T} w \\ \vdots \\ x^{(n)T} w \end{pmatrix} = Xw.$$

$\frac{\partial \hat{R}}{\partial \vec{y}} = \left(\frac{\partial \hat{R}}{\partial y^{(1)}} \cdots \frac{\partial \hat{R}}{\partial y^{(n)}} \right)^T$. Since only the jth term of $\hat{R} = \frac{1}{N} \sum_{i=1}^{N} L_\sigma(y^{(i)}, t^{(i)})$

depends on $y^{(j)}$, $0 < j < N$, $\frac{\partial \hat{R}}{\partial y^{(j)}} = \frac{1}{N} H'_\sigma(y^{(j)} - t^{(j)})$

In vector notation $\boxed{\frac{\partial \hat{R}}{\partial \vec{y}} = \frac{1}{N} H'_\sigma(y - t)}$

Using the chain rule $\frac{\partial \hat{R}}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} L'_\sigma(w^T x^{(i)}, t)\, x_j^{(i)}$, where $x_j^{(i)}$ is

the jth element of $x^{(i)}$, and L' is derivative with respect to $y^{(i)}$

The above can be written as $\frac{\partial \hat{R}}{\partial w_j} = \frac{1}{N} \left(x_j^{(1)}\ x_j^{(2)} \cdots x_j^{(N)} \right) \begin{bmatrix} L'_\sigma(w^T x^{(1)}, t) \\ \vdots \\ L'_\sigma(w^T x^{(n)}, t) \end{bmatrix}$

Therefore $\boxed{\frac{\partial \hat{R}}{\partial w} = \frac{1}{N} X^T H'_\sigma(Xw - t)}$

c. See python file

d. See python file

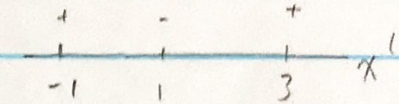
e. The training set loss is calculated by $\frac{1}{2N}\sum(\hat{y}-t)^2$, and when training the model we find weights that minimize $\frac{1}{N}\sum H_\delta(w^T x^{(i)}-t)$.

However, for larger δ, Huber loss becomes closer to squared error loss (see picture in part a), and the training model chooses weights to minimize a function that is closer to $\frac{1}{2N}\sum(\hat{y}-t)^2$. Therefore, we expect the training set loss to decrease for increasing δ.

Validation loss does not increase always increase, because the model generalizes poorly. For larger δ, the model fits squared errors loss better on the training set but also gets affected by the outliers more. The optimal δ needs to balance between these two factors.

Appendix: Output from python

```
linear regression validation loss: 28.44202896790706
delta: 0.1, valid. squared error loss: 4.909228730093678, train squared error loss: 38.041444425291125
delta: 0.5, valid. squared error loss: 4.018972761905316, train squared error loss: 38.311166891032116
delta: 1, valid. squared error loss: 4.200164017924187, train squared error loss: 37.993439863363456
delta: 5, valid. squared error loss: 23.636105018763637, train squared error loss: 28.84756778475285
delta: 10, valid. squared error loss: 26.420467729491516, train squared error loss: 27.79829743150903
```

3.a.  Suppose some weights perflectly classifies all points. Then it classifies $x^{(i)} = -1$ and $x^{(i)} = 3$ as 1.

From lecture, we know that the set of points that have the same prediction is a convex set, so the point $\frac{1}{2}(-1) + \frac{1}{2}(3) = -0.5 + 1.5 = 1$ must be classified with label 1. However, if $x^{(i)} = 1$, $t^{(i)} = 0$, so we have a contradiction and the data is not linearly seperable.

b.

ψ_1	ψ_2	$t^{(i)}$
-1	1	1
1	1	0
3	9	1

We want $w_1 \psi_1 + w_2 \psi_2 \geq 0$ only for points with $t^{(i)} = 1$. This gives:

$-w_1 + w_2 \geq 0$ — (1)

$w_1 + w_2 < 0$ — (2)

$3w_1 + 9w_2 \geq 0$ — (3)

A possible solution is $w_1 = -2$ and $w_2 = 1$. Then:

(1) becomes $-(-2) + 1 = 2 + 1 \geq 0$

(2) becomes $-2 + 1 = -1 < 0$

(3) becomes $3(-2) + 9(1) = -6 + 9 \geq 0$.

All 3 points are classified correctly.