① (a) $\boxed{\text{Sketch submitted as File 1a.png}}$ Huber loss becomes linear after a threshold reducing sensitivity to outliers (linear growth rather than quadratic) Therefore the loss introduced by an outlier will not aggitate the system significantly.

① (b)

$$H'(a) = \begin{cases} a & \text{if } |a| \leq \delta \\ \delta & \text{else} \end{cases} \qquad y = w^T x + b$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \frac{dy}{dw} = H'(y-t) \cdot x$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y} \frac{dy}{db} = H'(y-t)$$

② (a)

$$w^* = \operatorname{argmin} \frac{1}{2} \sum_i a^{(i)} (y^{(i)} - w^T x^{(i)})^2 + \frac{\lambda}{2} \|w\|^2$$

$$= \frac{1}{2} \sum_i a^{(i)} (y^{(i)2} + w^{T2} x^{(i)2} - 2 y^{(i)} w^T x^{(i)}) + \frac{\lambda}{2} \|w\|^2$$

$$= \frac{1}{2} (y^T A y + w^T X^T A X w - 2 w^T X^T A y) + \frac{\lambda}{2} \|w\|^2$$

gradient $= 0 = \nabla_{w^*} = \frac{1}{2} (2 X^T A X w^* - 2 X^T A y) + \lambda I w^*$

$$\cancel{\lambda w + X^T A X w = X^T}$$

$$\Rightarrow \lambda I w^* + X^T A X w^* = X^T A y$$

$$(\lambda I + X^T A X) w^* = X^T A y$$

$$\Rightarrow w^* = (\lambda I + X^T A X)^{-1} X^T A y$$

Mahsa Naserifar
1002330772
13-CSC411

② (C) $\boxed{\text{Plot submitted as file 2C.png}}$

② (d)

As τ ~~increases~~ decreases it is expected that the model is more sensitive to outliers. So at $\tau \to 0$, we will have overfitting and the validation loss is expected to be high.
At a higher τ, it is expected that the model loses sensitivity to outliers. As $\tau \to \infty$, we will have under fitting.

The loss in validation for low τ is as expected large. ~~However~~ the training loss increases as τ increases confirming the fact that the outliers are being ignored. And as expected the higher τ results in a more general model and less loss for validation.