Convex Optimization Project 2

Louis Arbogast, Jingwen Wang

November 2023

Question 1

The Huber loss function is given by:

$$L_\delta(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \delta \\ \delta|z| - \frac{1}{2}\delta^2 & \text{if } |z| > \delta \end{cases} \tag{1}$$

We want to prove that the following function equals the Huber loss:

$$h(z) = \inf_{t\in\mathbb{R}} \delta|t| + \frac{1}{2}(z-t)^2 = \inf_{t\in\mathbb{R}} \begin{cases} \frac{1}{2}t^2 - (z+\delta)t + \frac{1}{2}z^2, & \text{if } t < 0 \\ \frac{1}{2}z^2, & \text{if } t = 0 \\ \frac{1}{2}t^2 - (z-\delta t)t + \frac{1}{2}z^2, & \text{if } t > 0 \end{cases} \tag{2}$$

We treat the cases $t < 0$, $t = 0$ and $t > 0$ separately.

For $t < 0$, where $h(z) = \frac{1}{2}t^2 - (z+\delta)t + \frac{1}{2}z^2$

We find the minimum of the function by setting to 0 the derivative of $h(z)$ with respect to t:

$$\frac{\partial}{\partial t}\left[\frac{1}{2}t^2 - (z+\delta)t + \frac{1}{2}z^2\right] = t - z - \delta = 0 \Rightarrow t_{min} = z + \delta$$

If the minimum of the function is inside the treated domain, i.e. $t_{min} < 0 \Rightarrow z + \delta < 0 \Rightarrow z < -\delta$:

$$\inf_{t<0}\left(\frac{1}{2}t^2 - (z+\delta)t + \frac{1}{2}z^2\right) = -\delta z - \frac{\delta^2}{2}$$

Otherwise, if $t_{min} \geq 0 \Rightarrow z + \delta \geq 0 \Rightarrow z \geq -\delta$ the infimum on the domain $t < 0$ is located at $t = 0$ (because the function is quadratic with minimum in $t > 0$ thus it is decreasing on $t < 0$):

$$\inf_{t<0}\left(\frac{1}{2}t^2 - (z+\delta)t + \frac{1}{2}z^2\right) = \frac{1}{2}z^2$$

For $t > 0$, where $h(z) = \frac{1}{2}t^2 - (z-\delta)t + \frac{1}{2}z^2$

$$\frac{\partial}{\partial t}\left[\frac{1}{2}t^2 - (z-\delta)t + \frac{1}{2}z^2\right] = t - z + \delta = 0 \Rightarrow t_{min} = z - \delta$$

If the minimum of the function is inside the domain, i.e. $t_{min} > 0 \Rightarrow z - \delta > 0 \Rightarrow z > \delta$:

$$\inf_{t>0}\left(\frac{1}{2}t^2 - (z-\delta)t + \frac{1}{2}z^2\right) = \delta z - \frac{\delta^2}{2}$$

Otherwise, if $t_{min} \leq 0 \Rightarrow z - \delta \leq 0 \Rightarrow z \leq \delta$, the infimum on the domain $t > 0$ is located at $t = 0$ (because the function is quadratic with minimum in $t < 0$ thus it is increasing on $t < 0$):

$$\inf_{t>0}\left(\frac{1}{2}t^2 - (z-\delta)t + \frac{1}{2}z^2\right) = \frac{1}{2}z^2$$

Overall, if $z < -\delta$ and $z > \delta$ ($\Leftrightarrow |z| > \delta$):

$$h(z) = \inf_{t\in\mathbb{R}} \delta|t| + \frac{1}{2}(z-t)^2 = \delta|z| - \frac{\delta^2}{2}$$

And if $-\delta \leq z$ and $z \leq \delta$ ($\Leftrightarrow |z| \leq \delta$):

$$h(z) = \inf_{t \in \mathbb{R}} \delta|t| + \frac{1}{2}(z-t)^2 = \frac{1}{2}z^2$$

Thus we showed that $h(z) = L_\delta(z)$.

Our problem is given by:

$$\underset{w \in \mathbb{R}^N, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{m} L_\delta\left(w^\top x_i + b - y_i\right) + \frac{\rho}{2}\|w\|_2^2$$

By replacing the Huber loss by the function given in (2) and introducing $z_i = w^\top x_i + b - y_i$, the problem becomes:

$$\underset{w \in \mathbb{R}^N, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{m} \inf_{t_i \in \mathbb{R}} \delta\,|t_i| + \frac{1}{2}\left(z_i - t_i\right)^2 + \frac{\rho}{2}\|w\|_2^2$$

We then do the following change of variable: $t_i' = z_i - t_i \Rightarrow t_i = z_i - t_i'$
Since this is an affine transformation it preserves convexity. It also doesn't affect the infimum of the function since we only rescale its input and don't modify the functions expression. We obtain:

$$\underset{w \in \mathbb{R}^N, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{m} \inf_{t_i' \in \mathbb{R}} \delta\,|z_i - t_i'| + \frac{1}{2}t_i'^2 + \frac{\rho}{2}\|w\|_2^2$$

We can remove the infimum if we put $t \in \mathbb{R}^m$ in the decision variables since it is a minimization problem. By replacing z_i we get:

$$\underset{w \in \mathbb{R}^N, b \in \mathbb{R}, t \in \mathbb{R}^m}{\text{minimize}} \quad \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^{m} \delta\left|w^\top x_i + b - y_i - t_i\right| + \frac{\rho}{2}\|w\|_2^2$$

We can remove the absolute value and the sum by introducing 2 epigraphical variables for each term of the sum, which results in the following optimization problem:

$$\begin{array}{ll} \underset{\substack{w \in \mathbb{R}^n, b \in \mathbb{R}, t \in \mathbb{R}^m \\ r^+, r^- \in \mathbb{R}_+^m}}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \delta \mathbf{1}^\top\left(r^+ + r^-\right) + \frac{\rho}{2}\|w\|_2^2 \\ \text{subject to} & w^\top x_i + b - y_i - t_i \leq r_i^+ \quad \forall i = 1, \ldots, m \\ & y_i - w^\top x_i - b + t_i \leq r_i^- \quad \forall i = 1, \ldots, m \end{array}$$

Question 2

The predicted output is plotted in Figure 1. It shows that Huber loss is less affected by outliers compared to least squares regression. This happens because Huber loss uses absolute loss when the error is larger than a certain amount ($|z| > \delta$). Unlike least squares, Huber loss doesn't increase as much when the prediction error gets really big, especially with outliers. In this data set, there are two outliers—one at the top left and another at the bottom right. These outliers would usually make the regression line flatter. However, with Huber loss, the impact of these outliers is reduced, resulting in a steeper line that better follows the trend of the majority of the data.

Question 3

The regression problem in higher dimension is given by:

$$\begin{array}{ll} \underset{\substack{w \in \mathbb{R}^N, b \in \mathbb{R}, t \in \mathbb{R}^m \\ r^+, r^- \in \mathbb{R}_+^m}}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \delta \mathbf{1}^\top\left(r^+ + r^-\right) + \frac{\rho}{2}\|w\|_2^2 \\ \text{subject to} & w^\top \phi\left(x_i\right) + b - y_i - t_i \leq r_i^+ \quad \forall i = 1, \ldots, m \\ & y_i - w^\top \phi\left(x_i\right) - b + t_i \leq r_i^- \quad \forall i = 1, \ldots, m \end{array}$$

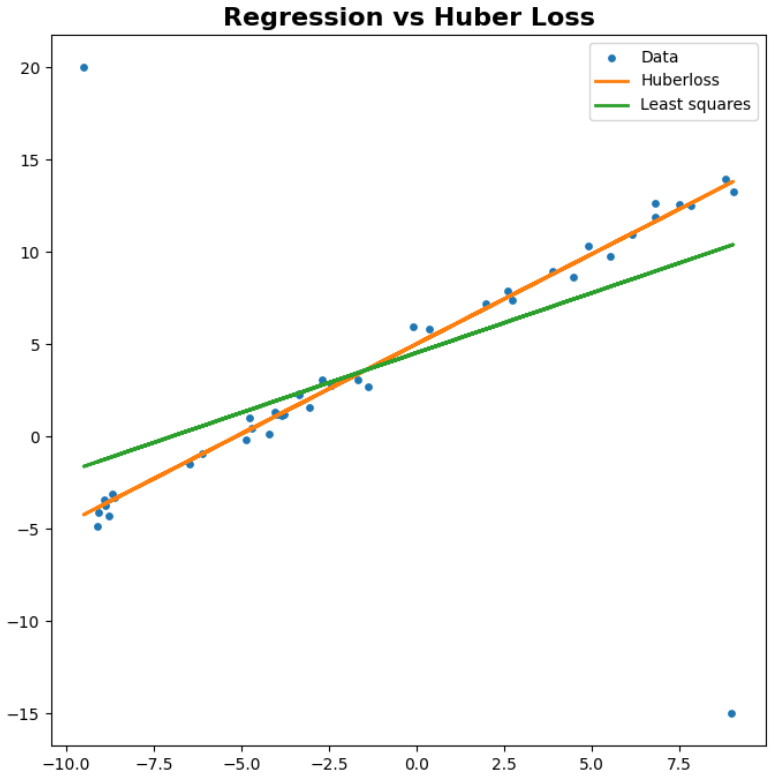

Figure 1: Regression vs Huber Loss

3.1

Let $\lambda^- \in \mathbb{R}_+^m$ and $\lambda^+ \in \mathbb{R}_+^m$ be the Lagrange multipliers associated to the inequality constraints. The Lagrangian can be expressed as:

$$
\begin{aligned}
L\left(w, b, t, r^+, r^-, \lambda^+, \lambda^-\right) = & \frac{1}{2}\|t\|_2^2 + \delta \overrightarrow{1}^\top \left(r^+ + r^-\right) + \frac{\rho}{2}\|w\|_2^2 \\
& + \sum_{i=1}^m \lambda_i^+ \left(w^\top \phi\left(x_i\right) + b - y_i - t_i - r_i^+\right) \\
& + \sum_{i=1}^m \lambda_i^- \left(y_i - w^\top \phi\left(x_i\right) - b + t_i - r_i^-\right) \\
& = \frac{\rho}{2}\|w\|_2^2 + w^\top \sum_{i=1}^m \phi\left(x_i\right)\left[\lambda_i^+ - \lambda_i^-\right] \\
& + b \sum_{i=1}^m \left(\lambda_i^+ - \lambda_i^-\right) \\
& + \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^m \left(\lambda_i^- - \lambda_i^+\right) t_i \\
& + \left(\delta \overrightarrow{1} - \lambda^+\right)^\top r^+ + \left(\delta \overrightarrow{1} - \lambda^-\right)^\top r^- \\
& + \sum_{i=1}^m y_i \left(\lambda_i^- - \lambda_i^+\right)
\end{aligned}
$$

3.2

The dual function is given as:

$$g(\lambda^+, \lambda^-) = \inf_{w,b,t,r^+,r^-} L\left(w, b, t, r^+, r^-, \lambda^+, \lambda^-\right)$$

Since the Lagrangian is continuous and convex with respect to w, b, t, we set the associated partial derivatives to 0 to find the infimum:

$$\begin{aligned}
\nabla_w L &= \rho w + \sum_{i=1}^{m} \phi(x_i)\left(\lambda_i^+ - \lambda_i^-\right) = 0 \Rightarrow w = \frac{1}{\rho}\sum_{i=1}^{m} \phi(x_i)\left(\lambda_i^- - \lambda_i^+\right) \\
\frac{\partial L}{\partial b} &= \sum_{i=1}^{m} \lambda_i^+ - \lambda_i^- = 0 \\
\nabla_t L &= t + \left(\lambda^- - \lambda^+\right) = 0 \Rightarrow t = \lambda^+ - \lambda^-
\end{aligned}$$

The Lagrangian is linear with respect to the variables $r_i^+, r_i^- \in \mathbb{R}_+^m$, which are positive. In order to have an infimum which is not $-\infty$ but instead 0, the slope associated to these variables must be positive, such that for all $i = 1, \ldots, m$:

$$\begin{aligned}
&\delta - \lambda_i^+ \geq 0 \Rightarrow 0 \leq \lambda_i^+ \leq \delta \\
&\delta - \lambda_i^- \geq 0 \Rightarrow 0 \leq \lambda_i^- \leq \delta \\
&\Rightarrow -\delta \leq \lambda_i^- - \lambda_i^+ = \beta_i \leq \delta
\end{aligned}$$

Replacing $\lambda_i^- - \lambda_i^+$ by β_i and using the previous results let us express the dual function as:

$$\begin{aligned}
g\left(\lambda^-, \lambda^+\right) =& \frac{1}{2\rho}\left(\sum_{i=1}^{m} \phi\left(x_i\right)\beta_i\right)^\top \left(\sum_{i=1}^{m} \phi\left(x_i\right)\beta_i\right) + \frac{1}{\rho}\left(\sum_{i=1}^{m} \phi\left(x_i\right)\beta_i\right)^\top \left(-\sum_{i=1}^{m} \phi\left(x_i\right)\beta_i\right) \\
&+ \frac{1}{2}\sum_{i=1}^{m} \beta_i^2 + \sum_{i=1}^{m} \beta_i\left(-\beta_i\right) + \sum_{i=1}^{m} y_i\beta_i \\
=& -\frac{1}{2\rho}\sum_{i=1}^{m}\sum_{j=1}^{m} \beta_i\phi(x_i)^\top \phi\left(x_j\right)\beta_j - \frac{1}{2}\sum_{i=1}^{m} \beta_i^2 + \sum_{i=1}^{m} \beta_i y_i
\end{aligned}$$

Therefore, we can express the dual problem as follows:

$$\begin{aligned}
\underset{\beta \in \mathbb{R}^m}{\text{maximize}} \quad & -\frac{1}{2\rho}\sum_{i=1}^{m}\sum_{j=1}^{m} \beta_i\phi(x_i)^\top \phi\left(x_j\right)\beta_j - \frac{1}{2}\sum_{i=1}^{m} \beta_i^2 + \sum_{i=1}^{m} \beta_i y_i \\
\text{subject to} \quad & \sum_{i=1}^{m} \beta_i = 0, \quad -\delta \leq \beta_i \leq \delta \quad \forall i = 1, \ldots, m
\end{aligned}$$

3.3

The stationarity KKT condition states that at the optimal point, the gradient of the Lagrangian is 0. Therefore from the partial derivative with respect to w (calculated in 3.2) at the optimal point (w^*, β^*) we have:

$$w^* = \frac{1}{\rho}\sum_{i=1}^{m} \phi(x_i)\beta_i^*$$

Such that the j^{th} component can be expressed as:

$$w_j^\star = \frac{1}{\rho}\sum_{i=1}^{m} \beta_i^\star \phi_j\left(x_i\right) \quad \forall j = 1, \ldots, N$$

From the partial derivative with respect to t evaluated in (t^*, β^*) we get:

$$t^* = (\lambda^+)^* - (\lambda^-)^* = -\beta^* \iff t_i^\star = -\beta_i^\star \quad \forall i = 1, \ldots, m$$

3.4

Introduce μ_i^+, μ_i^- and σ as lagrangian variables corresponding to $-\delta + \beta_i \leq 0$, $-\delta - \beta_i \leq 0$ and $\sum_{i=1}^m \beta_i = 0$ respectively. Lagrangian function for the dual problem

$$\begin{aligned} L(\beta, \mu, \sigma) = &- \frac{1}{2}\sum_{i=1}^m \beta_i^2 - \frac{1}{2\rho}\sum_{i=1}^m \sum_{i'=1}^m \beta_i \phi(x_i)^\top \phi(x_{i'})\beta_{i'} + \sum_{i=1}^m \beta_i y_i \\ &+ \sum_{i=1}^m \mu_i^+ (\beta_i - \delta) + \sum_{i=1}^m \mu_i^- (-\delta - \beta_i) + \sigma \sum_{i=1}^m \beta_i \end{aligned}$$

Step 1. Prove the strong duality.
The dual problem is concave as the equation,

$$-\frac{1}{2}\sum_{i=1}^m \beta_i^2 - \frac{1}{2\rho}\sum_{i=1}^m \sum_{i'=1}^m \beta_i \phi(x_i)^\top \phi(x_{i'})\beta_{i'}$$

are quadratic equations of β_i with negative coefficients and $\sum_{i=1}^m \beta_i y_i$ is affine.
Slater condition holds as,

$$\begin{aligned} \exists \beta \text{ such that } &\sum_{i=1}^m \beta_i = 0, -\delta \leq \beta_i \leq \delta, \forall i = 1, \ldots, m \\ &\text{for example, } \beta = \vec{0} \text{ (a zero vector).} \end{aligned}$$

Thus, strong duality holds. We can infer that the dual problem of the dual problem is the original primal problem.

Step 2. Show that σ corresponds to b
The original Lagrange problem

$$\begin{aligned} L\left(w, b, t, r^+, r^-, \lambda^+, \lambda^-\right) = &\frac{\rho}{2}\|w\|_2^2 - \sum_{i=1}^m \beta_i w^\top \phi(x_i) \\ &- b\sum_{i=1}^m \beta_i \\ &+ \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^m \beta_i t_i \\ &+ \delta \mathbf{1}^\top r^+ - \sum_{i=1}^m \lambda_i^+ r_i^+ \\ &+ \delta \mathbf{1}^\top r^- - \sum_{i=1}^m \lambda_i^- r_i^- \\ &+ \sum_{i=1}^m \beta_i y_i \end{aligned}$$

The two Lagrange problems should be the same as we proved from step 1. σ and b only appear once, both in first order. They are both the coefficients with respect to $\sum_{i=1}^m \beta_i$ We can infer that σ and b are the same variables.

Step 3. KKT Condition
Follow KKT conditions for optimality

$$\nabla_\beta L(\beta, \mu, \sigma) = \beta^* + \frac{1}{\rho}\phi(x)^\top \sum_{i=1}^m \phi(x_i)\beta_i^* - y - \mu^{-*} + \mu^{+*} + \sigma^* = 0 \quad \text{stationarity}$$

$$\begin{aligned} &\mu_i^{+*}(\beta_i^* - \delta) = 0 \\ &\mu_i^{-*}(-\beta_i^* - \delta) = 0 \end{aligned} \quad \begin{aligned} &\mu_i^{+*} = 0 \\ \Rightarrow\ &\mu_i^{-*} = 0 \\ &\forall i = 1 \ldots m \end{aligned} \quad \text{complementary slackness}$$

$$\Rightarrow \sigma^* = -\beta^* - \frac{1}{\rho}\phi(x)^\top \sum_{i=1}^{m} \phi(x_i)\beta_i^* + y$$

$$b = \sigma_k^* = -\beta_k + y_k - \frac{1}{\rho}\sum_{i=1}^{m} \phi\left(x_k\right)^\top \phi\left(x_i\right)\beta_i^* \quad \forall k = 1 \ldots m$$

3.5

If we can show that $K = B^\top B$ then we can prove that K is positive semidefinite

$$K = \begin{bmatrix} \phi\left(x_1\right)^T \phi\left(x_1\right) & \phi\left(x_1\right)^T \phi\left(x_2\right) & \ldots & \phi\left(x_1\right)^T \phi\left(x_m\right) \\ \phi\left(x_2\right)^T \phi\left(x_1\right) & \phi\left(x_2\right)^T \phi\left(x_2\right) & \ldots & \phi\left(x_2\right)^T \phi\left(x_m\right) \\ \vdots & \vdots & \ldots & \vdots \\ \phi\left(x_m\right)^T \phi\left(x_1\right) & \phi\left(x_m\right)^T \phi\left(x_2\right) & \ldots & \phi\left(x_m\right)^T \phi\left(x_m\right) \end{bmatrix} = \Phi(x)^T \Phi(x)$$

$$\text{Where } \Phi(x) = \begin{bmatrix} \phi\left(x_1\right) \\ \phi\left(x_2\right) \\ \vdots \\ \phi\left(x_m\right) \end{bmatrix}$$

3.6

In Figure 2, we see that using the kernel trick, which involves lifting the input data to a higher dimensional space, helps the prediction model follow the data's trend more closely. Regular least squares regression assumes a straight-line relationship and tends to miss subtle features like peaks and valleys in the data. However, these features are captured with the kernel trick, showing that the data behaves more like a Gaussian distribution. This approach provides a more accurate and detailed understanding of the data's pattern.

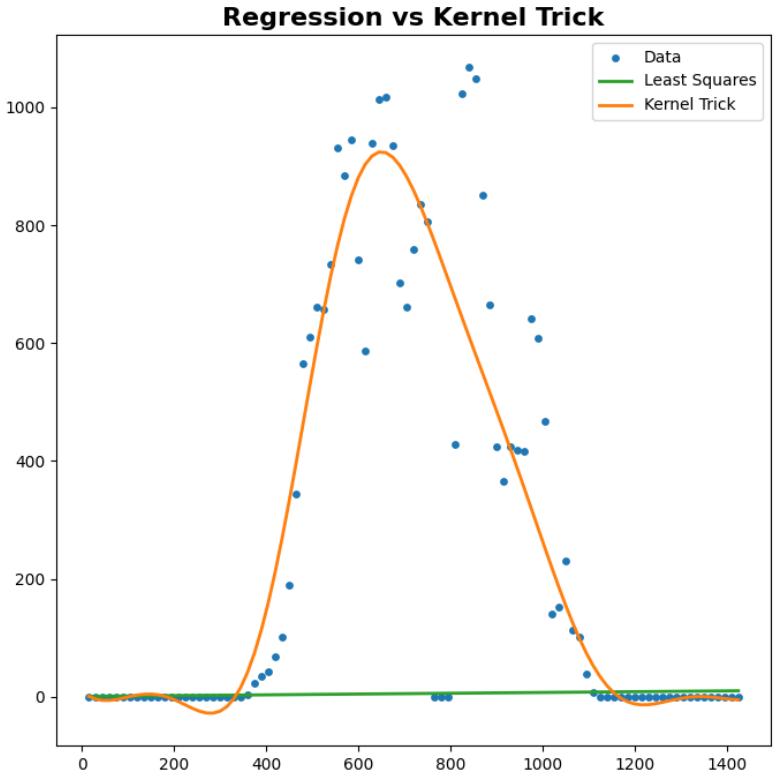

Figure 2: Regression vs Kernel Trick