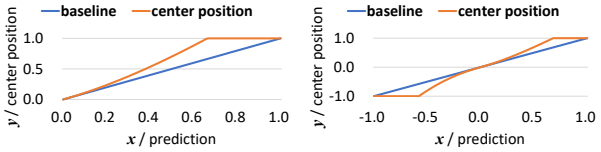
# Supplementary Material of CertPri

## A CENTER POSITION OF DEFINITION 1

To illustrate the center position of classification model and regression model in **Definition 1**, we draw a one-dimensional center position curve, where the $x$-axis is the model prediction output and the $y$-axis is the corresponding class center position. In Figure A.1, the baseline represents the $y = x$ line.




(a) Classification (b) Regression

**Figure A.1: Center position visualization of classification and regression in Definition 1.**

## B MORE DETAILS ABOUT CERTPRI

### B.1 Proof of Theorem 2

**Theorem 2** (Formal guarantee on lower bound $\gamma_L$ of inverse perturbation for regression model). *Let $\boldsymbol{x}_0 \in \mathbb{R}^{d_1}$ and $f^R : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be a regression model with continuously differentiable components. For all $\boldsymbol{\mu} \in \mathbb{R}^{d_1}$ with $||\boldsymbol{\mu}||_p \leq \min \frac{\sum_i |f^R_{i,+}(\boldsymbol{x}_0) - f^R_i(\boldsymbol{x}_0)|}{d_2 \times L^r_q}$, $\frac{1}{d_2} \sum |r(\boldsymbol{x}_0 + \boldsymbol{\mu}) - r(\boldsymbol{x}_0)| \leq \delta$ holds with $\frac{1}{p} + \frac{1}{q} = 1, 1 \leq \{p, q\} \leq \infty$ and $L^r_q$ is the Lipschitz constant for the function $\frac{\sum_i |f^R_{i,+}(\boldsymbol{x}) - f^R_i(\boldsymbol{x})|}{d_2}$ in $l_p$ norm. In other word, $\gamma_L = \min \frac{\sum_i |f^R_{i,+}(\boldsymbol{x}_0) - f^R_i(\boldsymbol{x}_0)|}{d_2 \times L^r_q}$ is a lower bound of minimum inverse perturbation.*

***Proof***. According to Lemma 1, the assumption that the function $h(\boldsymbol{x}) := \frac{\sum_i |f^R_{i,+}(\boldsymbol{x}) - f^R_i(\boldsymbol{x})|}{d_2}$ is Lipschitz continuous with Lipschitz constant $L^r_q$ gives:

$$|h(\boldsymbol{a}) - h(\boldsymbol{b})| \leq L^r_q ||\boldsymbol{a} - \boldsymbol{b}||_p. \tag{B.1}$$

Let $\boldsymbol{a} = \boldsymbol{x}_0 + \boldsymbol{\mu}$ and $\boldsymbol{b} = \boldsymbol{x}_0$, we get:

$$|h(\boldsymbol{x}_0 + \boldsymbol{\mu}) - h(\boldsymbol{x}_0)| \leq L^r_q ||\boldsymbol{\mu}||_p, \tag{B.2}$$

which can be rearranged as follows:

$$\begin{aligned} & -L^r_q ||\boldsymbol{\mu}||_p \leq h(\boldsymbol{x}_0 + \boldsymbol{\mu}) - h(\boldsymbol{x}_0) \leq L^r_q ||\boldsymbol{\mu}||_p, \\ \Rightarrow \quad & h(\boldsymbol{x}_0) - L^r_q ||\boldsymbol{\mu}||_p \leq h(\boldsymbol{x}_0 + \boldsymbol{\mu}) \leq h(\boldsymbol{x}_0) + L^r_q ||\boldsymbol{\mu}||_p. \end{aligned} \tag{B.3}$$

When $h(\boldsymbol{x}_0 + \boldsymbol{\mu}) = 0$, the inversely perturbed test input is moved to the regression center. As represented by Eq. (B.3), $h(\boldsymbol{x}_0) - L^r_q ||\boldsymbol{\mu}||_p$ is the lower bound of $h(\boldsymbol{x}_0 + \boldsymbol{\mu})$. If $h(\boldsymbol{x}_0) - L^r_q ||\boldsymbol{\mu}||_p \geq 0$ for sufficiently small inverse perturbation $||\boldsymbol{\mu}||_p$, the inversely perturbed test input cannot reach the regression center, i.e.,

$$\begin{aligned} & h(\boldsymbol{x}_0) - L^r_q ||\boldsymbol{\mu}||_p \geq 0, \\ \Rightarrow \quad & ||\boldsymbol{\mu}||_p \leq \frac{h(\boldsymbol{x}_0)}{L^r_q}, \\ \Rightarrow \quad & ||\boldsymbol{\mu}||_p \leq \frac{\sum_i |f^R_{i,+}(\boldsymbol{x}_0) - f^R_i(\boldsymbol{x}_0)|}{d_2 \times L^c_q}. \end{aligned} \tag{B.4}$$

### B.2 Formal Guarantee for ReLU

**Lemma 2** (Formal guarantee on $\gamma_L$ for ReLU activation). *Let $h(\boldsymbol{x}) = \sigma(W_l \sigma(W_{l-1} ... \sigma(W_1 \boldsymbol{x})))$ be a $l$-layer neural network with ReLU activation $\sigma(\boldsymbol{x}) = \max(0, \boldsymbol{x})$, where $W_i$ is the weights of the $i$-th layer. Here we do not consider bias terms due to their constant nature. Let $D \subset \mathbb{R}^d$ be a convex bounded closed set, then Equation* (1) *in the manuscript holds with $L_q = \sup_{\boldsymbol{x} \in D} \{|\sup_{||\boldsymbol{e}||_p = 1} h'(\boldsymbol{x}; \boldsymbol{e})|\}$ where $h'(\boldsymbol{x}; \boldsymbol{e}) := \lim_{\varepsilon \to 0^+} \frac{h(\boldsymbol{x} + \varepsilon \boldsymbol{e}) - h(\boldsymbol{x})}{\varepsilon}$ is the one-sided directional derivative, then* **Theorems 1** *and* **2** *still hold.*

***Proof***. For any $\{\boldsymbol{a}, \boldsymbol{b}\} \in D$, let $\boldsymbol{e} = \frac{\boldsymbol{b} - \boldsymbol{a}}{||\boldsymbol{b} - \boldsymbol{a}||_p}$ be the unit vector pointing from $\boldsymbol{a}$ to $\boldsymbol{b}$, and $r = ||\boldsymbol{a} - \boldsymbol{b}||_p$ is a $l_p$-norm. Define univariate function $\varphi(\varepsilon) = h(\boldsymbol{a} + \varepsilon \boldsymbol{e})$, then $\varphi(0) = h(\boldsymbol{a})$ and $\varphi(r) = h(\boldsymbol{b})$ and observe that $h'(\boldsymbol{a} + \varepsilon \boldsymbol{e}; \boldsymbol{e})$ and $h'(\boldsymbol{a} + \varepsilon \boldsymbol{e}; -\boldsymbol{e})$ are the right and left derivatives of $\varphi(\varepsilon)$, we have

$$\varphi'(\varepsilon) = \begin{cases} h'(\boldsymbol{a} + \varepsilon \boldsymbol{e}; \boldsymbol{e}) \leq L_q & \text{if } h'(\boldsymbol{a} + \varepsilon \boldsymbol{e}; \boldsymbol{e}) = h'(\boldsymbol{a} + \varepsilon \boldsymbol{e}; -\boldsymbol{e}) \\ \text{None} & \text{otherwise} \end{cases} \tag{B.5}$$

For a neural network with ReLU activation, there can be at most finite number of points in $\varepsilon \in (0, r)$ such that $\varphi'(\varepsilon)$ is none. This can be shown because each discontinuous $\varepsilon$ is caused by some ReLU activation, and there are only finite combinations. Let $0 = \varepsilon_0 < \varepsilon_1 < ... < \varepsilon_{k-1} < \varepsilon_k = 1$ be those points. Then, using the fundamental theorem of calculus on each interval separately, there exists $\overline{\varepsilon_i} \in (\varepsilon_i, \varepsilon_{i-1})$ for each $i$ such that

$$\begin{aligned} \varphi(r) - \varphi(0) &\leq \textstyle\sum_{i=0}^k |\varphi(\varepsilon_i) - \varphi(\varepsilon_{i-1})| \leq \sum_{i=0}^k |\varphi'(\overline{\varepsilon_i})(\varepsilon_i - \varepsilon_{i-1})| \\ &\leq \textstyle\sum_{i=0}^k L_q |\varepsilon_i - \varepsilon_{i-1}|_p = L_q ||\boldsymbol{b} - \boldsymbol{a}||_p, \end{aligned} \tag{B.6}$$

i.e., $h(\boldsymbol{b}) - h(\boldsymbol{a}) \leq L_q ||\boldsymbol{b} - \boldsymbol{a}||_p$. Therefore, **Theorems 1** and **2** still hold.

### B.3 Generalized Extreme Value Theory

Here we give the PDF and CDF of the three extreme value distributions.

**Gumbel** (Type I). The PDF and CDF with $\xi = 0$ are as follows:

$$\begin{aligned} g_\xi(z) &= \exp(-(z + \exp(-z))), \\ G_\xi(z) &= \exp(-\exp(-z)). \end{aligned} \tag{B.7}$$

**Fréchet** (Type II). The PDF and CDF with $\xi > 0$ are as follows:

$$\begin{aligned} g_\xi(z) &= \begin{cases} \exp(-(1 + \xi z)^{-\frac{1}{\xi}}) \times (1 + \xi z)^{-\frac{1+\xi}{\xi}}, & z > -\frac{1}{\xi} \\ 0, & z \leq -\frac{1}{\xi} \end{cases} \\ G_\xi(z) &= \begin{cases} \exp(-(1 + \xi z)^{-\frac{1}{\xi}}), & z > -\frac{1}{\xi} \\ 0, & z \leq -\frac{1}{\xi} \end{cases} \end{aligned} \tag{B.8}$$

**Weibull** (Type III). The PDF and CDF with $\xi < 0$ are as follows:

$$\begin{aligned} g_\xi(z) &= \begin{cases} \exp(-(1 + \xi z)^{-\frac{1}{\xi}}) \times (1 + \xi z)^{-\frac{1+\xi}{\xi}}, & z \leq -\frac{1}{\xi} \\ 0, & z > -\frac{1}{\xi} \end{cases} \\ G_\xi(z) &= \begin{cases} \exp(-(1 + \xi z)^{-\frac{1}{\xi}}), & z \leq -\frac{1}{\xi} \\ 1, & z > -\frac{1}{\xi} \end{cases} \end{aligned} \tag{B.9}$$

## B.4 Prioritization in Black-box Scenarios

We can extend CertPri to a black-box scenario based on the gradient estimation. The gradient norm at line 6 in **Algorithm 1** is computed by back propagation, which requires internal details of the model. When we replace it with the gradient estimation (as shown in **Algorithm 2**), CertPri can be implemented in a black-box scenario. We conduct a preliminary study based on a small dataset, and find that $T$>5 for Algorithm 2 is effective in general. To guarantee CertPri's effectiveness in the black-box scenario, we follow the double-minimum strategy, i.e., $T$=10, $\epsilon$=0.005max($\boldsymbol{x}$) for Algorithm 2.

**Algorithm 2**: Gradient estimation in the black-box scenario.

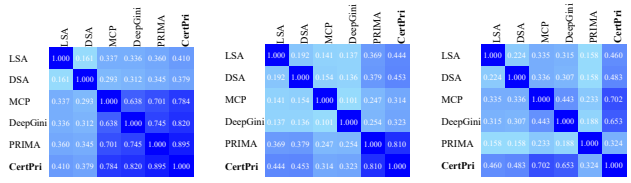
**Input**: A test input $\boldsymbol{x}_0$, a loss function $Loss(y_{true}, y_{pred})$ of $f(\boldsymbol{x})$, iterations $T$, a small constant $\epsilon$.
**Output**: The estimated gradient norm $\hat{g}$.

1 $\hat{g}_0 = 0$, noise mean $u_n = 0$, noise variance $\sigma_n = 1$.
2 **For** $i = 1 : T$ **do**
3 Randomly sample Gaussian noise vector $\boldsymbol{n}_i$ with $u_n$ and $\sigma_n$ in the same dimension as $\boldsymbol{x}_0$.
4 $\boldsymbol{x}_i^+ = \boldsymbol{x}_0 + \epsilon \times \boldsymbol{n}_i, \quad \boldsymbol{x}_i^- = \boldsymbol{x}_0 - \epsilon \times \boldsymbol{n}_i$.
5 $l_i^+ = Loss\left(f(\boldsymbol{x}_0), f(\boldsymbol{x}_i^+)\right), \quad l_i^- = Loss\left(f(\boldsymbol{x}_0), f(\boldsymbol{x}_i^-)\right)$.
6 $\hat{g}_i = \hat{g}_{i-1} + ||(l_i^+ - l_i^-) \times \boldsymbol{n}_i||_q$.
7 **End For**
8 $\hat{g} = \frac{\hat{g}_i}{2\epsilon \times T}$.

# C MORE EXPERIMENTAL RESULTS

## C.1 Correlation of Robustness

To interpret the utility of robustness, taking adaptive attacks on ImageNet dataset as an example (ID: 22-24), we first calculate the Pearson correlation of empirical movement costs for adaptive attacked test inputs prioritized by different methods. Then we illustrate them as heatmaps shown in Figure C.1.

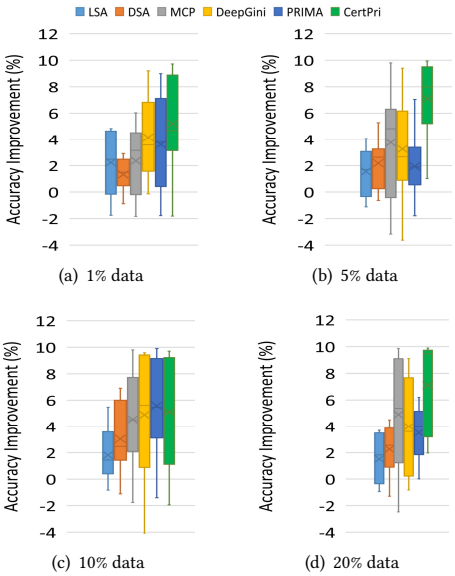



(a) +AdapS (ID: 22) (b) +AdapC (ID: 23) (c) +AdapM (ID: 24)

**Figure C.1: Pearson correlation of empirical movement costs for adaptive attacked test inputs prioritized by different methods on ImageNet dataset.**

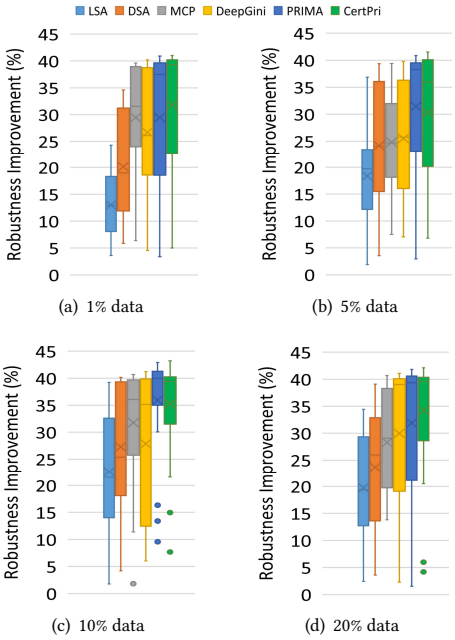
## C.2 Guidance of CertPri

The evaluation results of accuracy improvement and robustness improvement for DNNs on CIFAR10 dataset are illustrated as boxplots shown in Figure C.2 and Figure C.3. In terms of performance, we sample original data prioritized at the front 1%, 5%, 10% and 20% in the training set. We set epoch=5 for retraining due to the small number of data. We compare the test accuracy. In terms of robustness, we sample adversarial data prioritized at the front 1%, 5%, 10% and 20% in the adversarial set. We mix the sampled data with the original training set, and set epoch=2 for retraining due to the large number of data. Repeat above operations 5 times.




(a) 1% data (b) 5% data (c) 10% data (d) 20% data

**Figure C.2: The boxplots of accuracy improvement for different methods under first 1%, 5%, 10% and 20% data sampling.**




(a) 1% data (b) 5% data (c) 10% data (d) 20% data

**Figure C.3: The boxplots of robustness improvement for different methods under first 1%, 5%, 10% and 20% data sampling.**