A Appendix

The appendix contains additional material that could not be incorporated into the main paper due to page constraints. This includes detailed proofs of the theorems, supporting visual representations, and supplementary experimental data. The formal proofs, accompanied by illustrative diagrams that elucidate the basic concept behind each theorem, are discussed in sections B, C, and D. Section E delivers in-depth experimental methodologies that underpin the primary paper and some minor complementary experiments. These supplementary experiments comprise sensitivity analysis of the hyperparameters and a showcase of how Optuna's early stopping (pruning) algorithm has contributed to minimizing training time.

B Detailed Proofs for Theorem 1 and Accompanying Lemmas

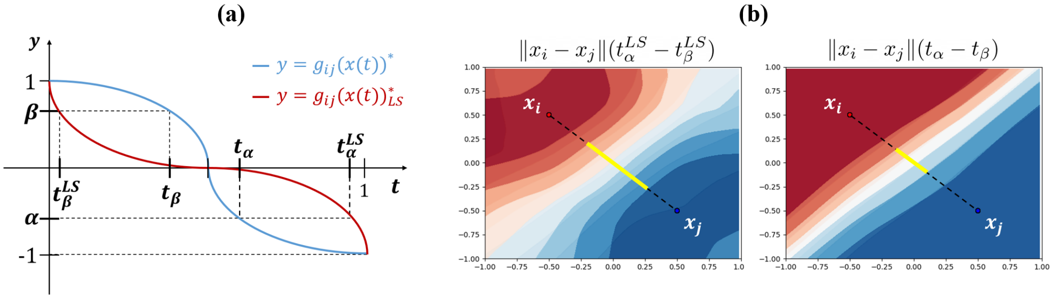

Figure 3: Depictions related to Theorem 1. **(a)** In simplified terms, the value of $g_{ij}(x(t))^*_{LS}$ for $t \in [0, 1/2]$ is lower than or equal to $g_{ij}(x(t))^*$, and the inverse holds true for $t \in [1/2, 1]$. This particular pattern gives rise to $t_\alpha^{LS} - t_\beta^{LS} \geq t_\alpha - t_\beta$, resulting in increased boundary thickness. **(b)** A simplified representation of how SPIDER's monotonically decreasing smoothing function assures increased boundary thickness. The yellow lines indicate the boundary thickness for SPIDER (on the left) and SPIDER without the smoothing function (on the right), respectively.

B.1 Proof on Theorem 1

Theorem 1. *Given any isotropic pdf $p_\epsilon(z) := p_\gamma(\|z\|)/S_n(\|z\|)$ with monotonically decreasing function $p_\gamma(\cdot)$ and a monotonically decreasing function $s(\cdot)$, the boundary thickness of f^*_{LS} is always greater than or equal to f^*, i.e.*

$$\forall -1 \leq \alpha \leq 0 < \beta \leq 1,\ \Theta(f^*, \alpha, \beta, x_i, x_j) \leq \Theta(f^*_{LS}, \alpha, \beta, x_i, x_j)$$

for datapoints $(x_i, y_i), (x_j, y_j)$ satisfying $y_i \neq y_j$ in the dataset.

Proof. Let $\mathcal{E}_{lin}$ denote the event that perturbed images $x_i + \epsilon_i$ and $x_j + \epsilon_j$ lies on segment $x(t)$. Given such event $\mathcal{E}_{lin}$, we can formulate the perturbed images as $x'_i := x_i + (x_j - x_i)\eta_i$ and $x'_j = x_j + (x_i - x_j)\eta_j$, where η_i and η_j are univariate random variables in $[0, 1]$. Since $\epsilon_i, \epsilon_j \sim p_\epsilon(\cdot)$ where $p_\epsilon(\cdot)$ is an isotropic decreasing function centered at origin, $\eta_i, \eta_j \sim p_\eta(\cdot)$ where $p_\eta(\cdot)$ is a 1D monotonically decreasing pdf defined in the domain $[0, 1]$. In addition, we assume that $\forall t \in [0, 1]$, $p_\eta(t) > 0$, i.e. the probabilities of perturbed images are nonzero at the segment $x(t)(t \in [0, 1])$ to make the calculation of f^* and f^*_{LS} possible. One justification for the assumption is that for small $h > 0$, $(1 - h)p_\eta(t) + h \cdot \text{Uniform}(0, 1) \approx p_\eta(t)$ in practice.

Now, we are ready to calculate the expected labels on the line segment. Since the primary interest here is boundary thickness which is parameterized by $\|x_i - x_j\|$ and $g_{ij}(x(t))$, we take a closer look at the ith and jth element of the one-hot encoded labels.

For any point $x(t)$ and its corresponding label $y = [y_0 \cdots y_i \cdots y_j \cdots y_c]$ w.l.o.g.,

$\mathbb{E}[y \mid x(t), \mathcal{E}_{lin}]$

$\propto \mathbb{E}[y \mid x_i + (x_j - x_i)\eta_i = x(t)] \cdot p_\eta(t) + \mathbb{E}[y \mid x_j + (x_i - x_j)\eta_j = x(t)] \cdot p_\eta(1 - t)$

$$= [0 \cdots 1 \cdots 0 \cdots 0] \cdot p_\eta(t) + [0 \cdots 0 \cdots 1 \cdots 0] p_\eta(1-t)$$

$$\mathbb{E}[y \mid x(t), \mathcal{E}_{lin}] = \left[0 \cdots \frac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)} \cdots \frac{p_\eta(1-t)}{p_\eta(t) + \_\eta(1-t)} \cdots 0 \right]$$

, i.e. $y_i = \dfrac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)}$, $y_j = \dfrac{p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$, and 0 elsewhere.

Let $f^* : \mathbb{R}^n \to \Delta_c$ be any function satisfying $f^*(x(t))_i = \dfrac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)}$ and $f^*(x(t))_j = \dfrac{p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$ for all $t \in [0, 1]$, where Δ_c is a probability simplex in $\mathbb{R}^c$. f^* is an optimal function that minimizes surrogate loss along the line segment $x(t), t \in [0, 1]$.

Analogously, we can formulate perturbed images with smoothed labels given $\mathcal{E}_{lin}$ as $(x_i + (x_j - x_i)\eta_i, \tilde{y}_i)$ and $(x_j + (x_i - x_j)\eta_j, \tilde{y}_j)$. The expected label $\tilde{y}$ of a point $x(t)$ on the line segment is

$$\mathbb{E}[\tilde{y} \mid x(t), \mathcal{E}_{lin}]$$

$$\propto \mathbb{E}[\tilde{y} \mid x_i + (x_j - x_i)\eta_i = x(t)] \cdot p_\eta(t) + \mathbb{E}[y \mid x_j + (x_i - x_j)\eta_j = x(t)] \cdot p_\eta(1-t)$$

$$= [r_i \cdots s(\|(x_j - x_i)\eta_i\|) \cdots r_i \cdots r_i] \cdot p_\eta(t) + [r_j \cdots r_j \cdots s(\|(x_j - x_i)\eta_j\|) \cdots r_j] \, p_\eta(1-t)$$

$$\left(r_i := \frac{1 - s(\|(x_j - x_i)\eta_i\|)}{c-1}, r_j := \frac{1 - s(\|(x_i - x_j)\eta_j\|)}{c-1} \right)$$

$$= [r_i \cdots s(\|(x_j - x_i)t\|) \cdots r_i \cdots r_i] \cdot p_\eta(t) + [r_j \cdots r_j \cdots s(\|(x_j - x_i)(1-t)\|) \cdots r_j] \, p_\eta(1-t)$$

$(\because \eta_i = t$ and $\eta_j = 1 - t$ given events $x_i + (x_j - x_i)\eta_i = x(t)$ and $x_j + (x_i - x_j)\eta_j = x(t)$.)

For the sake of readibility, let $d_i(t) := \|(x_j - x_i)t\|$ and $d_j(t) := \|(x_i - x_j)(1-t)\|$.

$0 \le d_i(t), d_j(t) \le \|x_j - x_i\|$ and $d_i(t) + d_j(t) = \|x_j - x_i\|$. We will simply denote $d_i(t), d_j(t)$ as d_i, d_j, and $r_i(t), r_j(t)$ as r_i, r_j wherever not necessary.

The sum of the elements of the above vector is $p_\eta(t) + p_\eta(1-t)$. Thus,

$$\mathbb{E}[\tilde{y} \mid x(t), \mathcal{E}_{lin}]$$

$$= \left[\frac{r_i p_\eta(t) + r_j p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} \cdots \frac{s(d_i) p_\eta(t) + r_j p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} \right.$$
$$\left. \cdots \frac{r_i p_\eta(t) + s(d_j) p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} \cdots \frac{r_i p_\eta(t) + r_j p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} \right]$$

Let $f^*_{LS} : \mathbb{R}^n \to \Delta_c$ be any function satisfying

$$\forall\, t \in [0, 1], \quad f^*_{LS}(x(t))_{k \in [c]} = \begin{cases} \dfrac{s(d_i) p_\eta(t) + r_j p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} & \text{if } k = i, \\[2ex] \dfrac{r_i p_\eta(t) + s(d_j) p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} & \text{if } k = j, \\[2ex] \dfrac{r_i p_\eta(t) + r_j p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} & \text{otherwise.} \end{cases}$$

f^*_{LS} is an optimal function that minimizes surrogate loss along the segment $x(t), t \in [0, 1]$.

We can now calculate the boundary thickness of f^* and f^*_{LS}. To clearly distinguish $g_{ij}(x(t))$ for f^* and f^*_{LS}, we use notations $g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$.

$\Theta(f^*, \alpha, \beta, x_i, x_j) := \|x_i - x_j\| \int_0^1 I\{\alpha < g_{ij}(x(t))^* < \beta\} dt$

$\Theta(f^*_{LS}, \alpha, \beta, x_i, x_j) := \|x_i - x_j\| \int_0^1 I\{\alpha < g_{ij}(x(t))^*_{LS} < \beta\} dt$

$$g_{ij}(x(t))^* = \frac{p_\eta(t)}{p_\eta(t)+p_\eta(1-t)} - \frac{p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)} = \frac{p_\eta(t)-p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)}$$

$$\begin{aligned} g_{ij}(x(t))^*_{LS} &= \frac{s(d_i)p_\eta(t)+r_jp_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)} - \frac{r_ip_\eta(t)+s(d_j)p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)} \\ &= \frac{(s(d_i)-r_i)p_\eta(t)-(s(d_j)-r_j)p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)} \end{aligned}$$

$$s(d_i)-r_i = s(d_i) - \frac{1-s(d_i)}{c-1} = \frac{c\cdot s(d_i)-1}{c-1}$$

Since $1/c \leq s(d_i) \leq 1$, $0 \leq \dfrac{c\cdot s(d_i)-1}{c-1} \leq 1$, i.e. $0 \leq s(d_i)-r_i \leq 1$.

For the sake of simplicity, let $\gamma_i(d_i) := s(d_i) - r_i$. $r_i(d_i)$ is a monotonically decreasing function with respect to d_i ($\because s(\cdot)$ and $c\cdot s(\cdot)-1$ are monotonically decreasing functions.) Trivially, $0 \leq \gamma_j(d_j) \leq 1$ and $\gamma_j(\cdot)$ is a monotonically decreasing function.

Recap that $g_{ij}(x(t))^* = \dfrac{p_\eta(t)-p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)}$. Let us rewrite $g_{ij}(x(t))^*_{LS}$ as:

$$g_{ij}(x(t))^*_{LS} = \frac{\gamma_i(d_i)p_\eta(t)-\gamma_j(d_j)p_\eta(1-t)}{p_\eta(t)+p_\eta(1-t)}$$

One important thing to note here is that both $g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ are monotonically decreasing functions with respect to t that are symmetric to the point $(1/2, 0)$.

Lemma 1. *$g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ are monotonically decreasing functions with respect to $t \in [0,1]$ that are symmetric to the point $(1/2,0)$.*

Another thing to note is that if we show $0 \leq g_{ij}(x(t))^*_{LS} \leq g_{ij}(x(t))^*$ for $t \in [0,1/2]$ and $0 \geq g_{ij}(x(t))^*_{LS} \geq g_{ij}(x(t))^*$ for $t \in [1/2,1]$, at the same time we are showing $\forall -1 \leq \alpha \leq 0 \leq \beta \leq 1, \Theta(f^*_{LS}, \alpha, \beta, x_i, x_j) \geq \Theta(f^*, \alpha, \beta, x_i, x_j)$ considering Lemma 1 (Figure 3a.)

Lemma 2. *$0 \leq g_{ij}(x(t))^*_{LS} \leq g_{ij}(x(t))^*$ for $t \in [0,1/2]$ and $0 \geq g_{ij}(x(t))^*_{LS} \geq g_{ij}(x(t))^*$ for $t \in [1/2,1]$.*

With Lemma 1 and Lemma 2, we can now finally derive that the boundary thickness of f^*_{LS} is greater than or equal to f^*.

Let $0 \leq t^{LS}_\beta \leq 1/2 \leq t^{LS}_\alpha \leq 1$ be the values of t such that

$$(t^{LS}_\alpha, t^{LS}_\beta) = \left(\inf_t g_{ij}(x(t))^*_{LS} = \alpha, \sup_t g_{ij}(x(t))^*_{LS} = \beta\right).$$

Since $g_{ij}(x(t))^*_{LS}$ is monotonically decreasing, the boundary thickness of f^*_{LS} is $\|x_i - x_j\| \int_0^1 I\{\alpha < g_{ij}(x(t))^*_{LS} < \beta\}dt = \|x_i - x_j\| \cdot (t^{LS}_\alpha - t^{LS}_\beta)$.

Similarly, let $0 \leq t_\beta \leq 1/2 \leq t_\alpha \leq 1$ be the values of t such that

$$(t_\alpha, t_\beta) = \left(\inf_t g_{ij}(x(t))^* = \alpha, \sup_t g_{ij}(x(t))^*_{LS} = \beta\right).$$

The boundary thickness of f^* is $\|x_i - x_j\| \cdot (t_\alpha - t_\beta)$.

We now show that $t^{LS}_\beta \leq t_\beta$ using Lemma 1 and Lemma 2, with the definition of t^{LS}_β and t_β. For the sake of simplicity, let $h^*(t)$ and $h^*_{LS}(t)$ denote $g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ respectively.

$$\begin{aligned}
&h^*_{LS}(t^{LS}_\beta) = \beta \leq h^*(t^{LS}_\beta) && \text{Def. of } t^{LS}_\beta\text{, Lemma 2} \\
&h^*(t_\beta) = \beta \geq h^*_{LS}(t_\beta) && \text{Def. of } t_\beta\text{, Lemma 2} \\
&h^*_{LS}(t^{LS}_\beta) \geq h^*_{LS}(t_\beta) && h^*_{LS}(t^{LS}_\beta) = \beta, \beta \geq h^*_{LS}(t_\beta) \\
&t^{LS}_\beta \leq t_\beta && \text{Lemma 1}
\end{aligned}$$

Likewise, we can trivially derive $t^{LS}_\alpha \geq t_\alpha$. Then, $(t^{LS}_\alpha - t^{LS}_\beta) \geq (t_\alpha - t_\beta)$, i.e. $\|x_i - x_j\|(t^{LS}_\alpha - t^{LS}_\beta) \geq \|x_i - x_j\|(t_\alpha - t_\beta)$. □

B.2 Proof on Lemma 1

Lemma 1. *$g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ are monotonically decreasing functions with respect to $t \in [0,1]$ that are symmetric to the point $(1/2, 0)$.*

Proof. We divide the proof into two parts. First we prove that $g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ are symmetric to the point $\left(\frac{1}{2}, 0\right)$, and then prove the functions are monotonically decreasing.

A function f being symmetric to a point $\left(\frac{1}{2}, 0\right)$ indicates that $f\left(\frac{1}{2}+x\right) = -f\left(\frac{1}{2}-x\right)$.

$$g_{ij}(x(t))^* = \frac{p_\eta(t) - p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$$

$$g_{ij}(x\left(\tfrac{1}{2}+t\right))^* = \frac{p_\eta(\frac{1}{2}+t) - p_\eta(\frac{1}{2}-t)}{p_\eta(\frac{1}{2}+t) + p_\eta(\frac{1}{2}-t)} = -\frac{p_\eta(\frac{1}{2}-t) - p_\eta(\frac{1}{2}+t)}{p_\eta(\frac{1}{2}-t) + p_\eta(\frac{1}{2}+t)} = -g_{ij}(x\left(\tfrac{1}{2}-t\right))$$

$$g_{ij}(x(t))^*_{LS} = \frac{(s(d_i(t)) - r_i(t))p_\eta(t) - (s(d_j(t)) - r_j(t))p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}, \text{ where}$$

$$r_i(t) = \frac{1 - s(d_i(t))}{c-1}, r_j(t) = \frac{1 - s(d_j(t))}{c-1}$$

$d_i(t) = \|(x_j - x_i)t\|$, and $d_j(t) = \|(x_i - x_j)(1-t)\| = \|(x_j - x_i)(1-t)\|$.

We will unpack $g_{ij}(x(t))^*_{LS}$.

$$g_{ij}(x(t))^*_{LS} = \frac{1}{p_\eta(t) + p_\eta(1-t)} \cdot \left(\frac{c \cdot s(d_i(t)) - 1}{c-1} \cdot p_\eta(t) - \frac{c \cdot s(d_j(t)) - 1}{c-1} \cdot p_\eta(1-t)\right)$$

$$= \frac{1}{p_\eta(t) + p_\eta(1-t)} \times$$

$$\left(\frac{c \cdot s(\|(x_j - x_i)t\|) - 1}{c-1} \cdot p_\eta(t) - \frac{c \cdot s(\|(x_j - x_i)(1-t)\|) - 1}{c-1} \cdot p_\eta(1-t)\right)$$

$$g_{ij}(x(\tfrac{1}{2}+t))^*_{LS}$$

$$= \frac{1}{p_\eta(\frac{1}{2}+t) + p_\eta(\frac{1}{2}-t)} \times$$

$$\left(\frac{c \cdot s(\|(x_j - x_i)(\frac{1}{2}+t)\|) - 1}{c-1} \cdot p_\eta(\tfrac{1}{2}+t) - \frac{c \cdot s(\|(x_j - x_i)(\frac{1}{2}-t)\|) - 1}{c-1} \cdot p_\eta(\tfrac{1}{2}-t)\right)$$

$$= -g_{ij}(x(\tfrac{1}{2}-t))^*_{LS}$$

Now, we prove $g_{ij}(x(t))^*$ and $g_{ij}(x(t))^*_{LS}$ are monotonically decreasing.

$$g_{ij}(x(t))^* = \frac{p_\eta(t) - p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$$

We show that $g_{ij}(x(t))^*$ is monotonically decreasing for all $t \in [0, 1/2]$. For the sake of simplicity, we use $p(\cdot)$ to denote $p_\eta(\cdot)$ here. $\forall\ 0 \le a \le b \le 1/2$,

$$g_{ij}(x(a))^* - g_{ij}(x(b))^*$$

$$= \frac{p(a) - p(1-a)}{p(a) + p(1-a)} - \frac{p(b) - p(1-b)}{p(b) + p(1-b)}$$

$$= \frac{2p(a)p(1-b) - 2p(1-a)p(b)}{(p(a) + p(1-a))(p(b) + p(1-b))}$$

$$\text{sign}(g_{ij}(x(a))^* - g_{ij}(x(b))^*) = \text{sign}(p(a)p(1-b) - p(1-a)p(b)) \ \ (\because\ p(\cdot) : [0,1] \to \mathbb{R}_{>0})$$

$$p(a)p(1-b) - p(1-a)p(b)$$

$$= p(1-a)p(b) \cdot \left(\frac{p(a)p(1-b)}{p(1-a)p(b)} - 1 \right)$$
$$= p(1-a)p(b) \cdot \left(\frac{p(a)}{p(1-a)} \div \frac{p(b)}{p(1-b)} - 1 \right)$$

Since p is monotonically decreasing, $p(a) \geq p(b) \geq p(1-b) \geq p(1-a) > 0$.

$$\frac{p(a)}{p(1-a)} \geq \frac{p(b)}{p(1-a)} \geq \frac{p(b)}{p(1-b)} \rightarrow \frac{p(a)}{p(1-a)} - \frac{p(b)}{p(1-b)} \geq 0.$$

$$\frac{p(a)}{p(1-a)} \div \frac{p(b)}{p(1-b)} \geq 1 \left(\because \frac{p(t)}{p(1-t)} > 0 \quad \forall\, t \in [0,1]. \right)$$

Thus, $p(a)p(1-b) - p(1-a)p(b) \geq 0$, which leads to $g_{ij}(x(a)) - g_{ij}(x(b)) \geq 0$ ($\forall 0 \leq a \leq b \leq 1/2$.)

Since $g_{ij}(x(t))^*$ is symmetric to the point $\left(\frac{1}{2}, 0\right)$, $g_{ij}(x(t))^*$ is also monotonically decreasing in $t \in \left[\frac{1}{2}, 1\right]$.

Likewise, let $\gamma_i(d_i(t)) := \dfrac{c \cdot s(d_i(t)) - 1}{c - 1}$. $\forall\ t \in [0,1],\ 0 \leq \gamma_i(d_i(t)) \leq 1$ and $\gamma_i(d_i(t))$ is a monotonically decreasing function.

$$g_{ij}(x(t))^*_{LS} = \frac{\gamma_i(d_i(t))p_\eta(t) - \gamma_i(1-d_i(t))p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$$

Analogously, we can derive that $g_{ij}(x(t))^*_{LS}$ is monotonically decreasing function with trivial calculations. □

B.3 Proof on Lemma 2

Lemma 2. $0 \leq g_{ij}(x(t))^*_{LS} \leq g_{ij}(x(t))^*$ *for* $t \in [0, 1/2]$ *and* $0 \geq g_{ij}(x(t))^*_{LS} \geq g_{ij}(x(t))^*$ *for* $t \in [1/2, 1]$.

Proof. $g_{ij}(x(t))^* = \dfrac{p_\eta(t) - p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)}$ and

$$g_{ij}(x(t))^*_{LS} = \frac{1}{p_\eta(t) + p_\eta(1-t)} \cdot \left(\frac{c \cdot s(d_i(t)) - 1}{c-1} \cdot p_\eta(t) - \frac{c \cdot s(d - d_i(t)) - 1}{c-1} \cdot p_\eta(1-t) \right)$$

$$g_{ij}(x(t))^* - g_{ij}(x(t))^*_{LS}$$
$$= \frac{1}{p_\eta(t) + p_\eta(1-t)} \cdot \left(\frac{c - c \cdot s(d_i(t))}{c-1} \cdot p_\eta(t) - \frac{c - c \cdot s(d - d_i(t))}{c-1} \cdot p_\eta(1-t) \right)$$

$$p_\eta(t) \propto p_\epsilon((x_j - x_i)t) = \frac{p_\gamma(\|(x_j - x_i)t\|)}{S_n(\|(x_j - x_i)t)\|)} = \frac{p_\gamma(\|\epsilon_i\|)}{S_n(\|\epsilon_i\|)}$$

$$p_\eta(1-t) \propto p_\epsilon((x_i - x_j)(1-t)) = \frac{p_\gamma(\|(x_j - x_i)(1-t)\|)}{S_n(\|(x_j - x_i)(1-t))\|)} = \frac{p_\gamma(\|\epsilon_j\|)}{S_n(\|\epsilon_j\|)}$$

Case $t \in [0, 1/2)$**:**

$$\frac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)} = \frac{\dfrac{p_\gamma(\|\epsilon_i\|)}{S_n(\|\epsilon_i\|)}}{\dfrac{p_\gamma(\|\epsilon_i\|)}{S_n(\|\epsilon_i\|)} + \dfrac{p_\gamma(\|\epsilon_j\|)}{S_n(\|\epsilon_j\|)}} = \frac{1}{1 + \dfrac{p_\gamma(\|\epsilon_j\|)}{p_\gamma(\|\epsilon_i\|)} \cdot \dfrac{S_n(\|\epsilon_i\|)}{S_n(\|\epsilon_j\|)}} \geq \frac{1}{1 + \left(\dfrac{\|\epsilon_i\|}{\|\epsilon_j\|}\right)^{n-1}}$$

The last inequality comes from fact that $p_\gamma(\cdot)$ is a monotonically decreasing function and $\|\epsilon_i\| = \|(x_j - x_i)t\| < d/2 < \|(x_i - x_j)(1-t)\| = \|\epsilon_j\|$.

As n grows, $(\|\epsilon_i\|/\|\epsilon_j\|)^{n-1}$ converges to 0 for any $t \in [0, 1/2)$. In other terms, the high-dimensionality of the input space essentially makes $(|\epsilon_i|/|\epsilon_j|)^{n-1}$ to be practically zero. As an example, in the CIFAR-10/100 benchmarks where the dimensionality n equals $32 \cdot 32 \cdot 3 = 3071$, $t \in [0, 1/2 - 10^{-3}]$ gives $(|\epsilon_i|/|\epsilon_j|)^{n-1} < 4.6 \times 10^{-6}$. Essentially, we can regard $(|\epsilon_i|/|\epsilon_j|)^{n-1}$ as virtually zero in real-world scenarios, unless we encounter the unlikely cases where t is extraordinarily near to $1/2$. Formally,

$$1 \geq \frac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)} \geq \frac{1}{1 + (\|\epsilon_i\|/\|\epsilon_j\|)^{n-1}} \approx 1 \longrightarrow \frac{p_\eta(t)}{p_\eta(t) + p_\eta(1-t)} \approx 1.$$

Accordingly, $\dfrac{p_\eta(1-t)}{p_\eta(t) + p_\eta(1-t)} \approx 0.$

$$g_{ij}(x(t))^* - g_{ij}(x(t))^*_{LS}$$
$$= \frac{1}{p_\eta(t) + p_\eta(1-t)} \cdot \left(\frac{c \cdot s(d_i(t)) - 1}{c-1} \cdot p_\eta(t) - \frac{c \cdot s(d - d_i(t)) - 1}{c-1} \cdot p_\eta(1-t) \right)$$
$$\approx \frac{c \cdot s(d_i(t)) - 1}{c-1} \geq 0 \qquad (\because 1/c \leq s(d_i(t)) \leq 1)$$

$$g_{ij}(x(0))^*_{LS} = \frac{p_\eta(0) - p_\eta(1)}{p_\eta(0) + p_\eta(1)} \geq 0.$$

Using Lemma 1 and $g_{ij}(x(0))^*_{LS} \geq 0$, we have $g_{ij}(x(t))^*_{LS} \geq 0$ for $t \in [0, 1/2)$.

Case $t \in (1/2, 1]$**:**

Using Lemma 1 and $0 \leq g_{ij}(x(t))^*_{LS} \leq g_{ij}(x(t))^*$ for $t \in [0, 1/2)$, $0 \geq g_{ij}(x(t))^*_{LS} \geq g_{ij}(x(t))^*$ for $t \in (1/2, 1]$.

Case $t = 1/2$**:**

Using Lemma 1, $g_{ij}(x(1/2))^*_{LS} = g_{ij}(x(1/2))^* = 0$. □

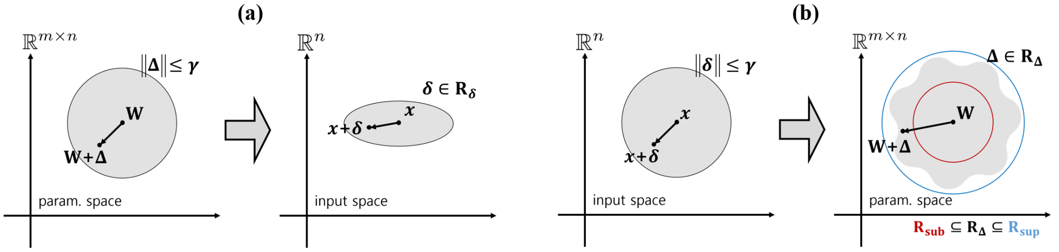

Figure 4: Illustrations of theorems on perturbation conversions. **(a)** The parameter perturbation region $\{\|\Delta\| \leq \gamma\}$ can be connected to an ellipsoidal perturbation region R_δ (Theorem 2.) **(b)** The input perturbation region $\{\|\delta\| \leq \gamma\}$ can be connected to parameter perturbation region R_Δ. R_Δ has the subset $R_{sub} := \{\|\Delta\| \leq (\lambda_{min}/\|x_{max}\|)^2\}$ (Theorem 3) and the superset $R_{sup} := \{\|\Delta\| \leq \rho^2\}$ (Theorem 4.)

C Converting Perturbations in Parameter Space to Input Space

Given weights $W \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$, input $x \in \mathbb{R}^n$, and parameter perturbation region $\|\Delta\| \leq \gamma$, we want to find the region R_δ so that $\forall\, \|\Delta\| \leq \gamma, \exists\, \delta \in R_\delta \;\; s.t. \;\; \sigma(W(x+\delta)+b) = \sigma((W+\Delta)x+b)$ and $\forall\, \delta \in R_\delta, \exists\, \|\Delta\| \leq \gamma \;\; s.t. \;\; \sigma(W(x+\delta)+b) = \sigma((W+\Delta)x+b)$. In other words, we want to find the region R_δ so that for every element e_1 in region $\{\Delta \in \mathbb{R}^{m\times n} \mid \|\Delta\| \leq \gamma\}$ there exists an element e_2 in region R_δ satisfying the equation and vice versa.

Since $\sigma(\cdot) : \mathbb{R}^m \to (0,1)^m$ is a bijective function, $\sigma(W(x+\delta)+b) = \sigma((W+\Delta)x+b) \iff W(x+\delta)+b = (W+\Delta)x+b$. This equality can be reduced to $W\delta = \Delta x$.

We will first examine the range of Δx in the output space, given $\|\Delta\| \leq \gamma$. Δx can be written in several ways:

$$\Delta x = \begin{bmatrix} | & | & & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= c_1x_1 + c_2x_2 + \cdots + c_nx_n = \begin{bmatrix} c_{11} \\ c_{21} \\ \cdots \\ c_{m1} \end{bmatrix} x_1 + \begin{bmatrix} c_{12} \\ c_{22} \\ \cdots \\ c_{m2} \end{bmatrix} x_2 + \cdots + \begin{bmatrix} c_{1n} \\ c_{2n} \\ \cdots \\ c_{mn} \end{bmatrix} x_n$$

, where c_i is the ith column vector and c_{ij} is an element in ith row, jth column of Δ.

Next, we will rewrite $\|\Delta\| \leq \gamma$ as the following constraints:

$$\|\Delta\| \leq \gamma$$
$$\iff \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2 \leq \gamma^2$$
$$\iff \sum_{j=1}^{n} \|c_j\|^2 \leq \gamma_j^2 \text{ subject to } \gamma_1^2 + \gamma_2^2 + \cdots + \gamma_n^2 = \gamma^2.$$

When we reexamine the above formulas in $\mathbb{R}^m$, finding the range of Δx can be regarded as finding the range of linear combination of column vectors in $\mathbb{R}^m$ such that each column vector c_i is restricted to $\|c_i\| \leq \gamma_i$.

Given two vectors v_1 and v_2 s.t. $\|v_1\| \leq \gamma_1$ and $\|v_2\| \leq \gamma_2$, $\|v_1 + v_2\| \leq \gamma_1 + \gamma_2$. Trivially, for any $\alpha \in \mathbb{R}$, $\|\alpha \cdot v_1\| \leq |\alpha|\gamma_1$. That is, the range of linear combination $\Delta x = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is also a ball, i.e. $\|\Delta x\| \leq \sum_{i=1}^n |x_i|\gamma_i$ subject to $\sum_{i=1}^n \gamma_i^2 = \gamma^2$.

Finding the range of $\|\Delta x\|$ is now equivalent to finding the maximum radius of $\sum_{i=1}^{n} |x_i|\gamma_i$ with the constraint $\sum_{i=1}^{n} \gamma_i^2 = \gamma^2$. Using Lagrange multipliers method, let $r := [\gamma_1, \gamma_2, \cdots, \gamma_n]$, $f(r) := \sum_{i=1}^{n} |x_i|\gamma_i$, $g(r) := \sum_{i=1}^{n} \gamma_i^2 - \gamma^2$, and $L(r, \lambda) := f(r) - \lambda(g(r))$.

$$\frac{\partial L}{\partial \gamma_i} = |x_i| - 2\lambda\gamma_i = 0 \iff \gamma_i = \frac{|x_i|}{2\lambda}$$

Substituting the above equality to $g(r) = 0$,

$$\sum_{i=1}^{n} \frac{x_i^2}{4\lambda^2} - \gamma^2 = 0 \iff \lambda = \frac{\sqrt{\sum {x_i}^2}}{2\gamma}$$

$$\gamma_i = \frac{|x_i|}{2\lambda} = \frac{|x_i|\gamma}{\sqrt{\sum {x_i}^2}}$$

$$f(r) = \sum_{i=1}^{n} \frac{{x_i}^2\gamma}{\sqrt{\sum {x_i}^2}} = \frac{\sum_{i=1}^{n} x_i^2}{\sqrt{\sum_{i=1}^{n} x_i^2}}\gamma = \|x\| \cdot \gamma$$

Therefore, $\|\Delta x\| \leq \|x\|\gamma$.

We now consider the LHS of equation $W\delta = \Delta x$. Let $W = U\Sigma V^\top$ be the SVD Decomposition of $W \in \mathbb{R}^{m\times n}$. Multiplying $U^\top$ to both sides of the equation, $\Sigma V^\top \delta = U^\top \Delta x$. The inequality induced by L_2 norm, i.e. ball, does not change when we multiply any orthogonal matrix. Thus, $\|U^\top \Delta x\| \leq \|x\|\gamma$.

Let $\delta' := V^\top \delta = [\delta'_1, \cdots, \delta'_n]^\top$.

$$\Sigma V^\top \delta = \Sigma \delta' = \begin{bmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_m & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \delta'_1 \\ \vdots \\ \delta'_m \\ \delta'_{m+1} \\ \vdots \\ \delta_n \end{bmatrix} = \begin{bmatrix} \sigma_1\delta'_1 \\ \vdots \\ \sigma_m\delta'_m \end{bmatrix}$$

Since $\|U^\top \Delta x\| \leq \|x\|\gamma$ and $\Sigma\delta' = U^\top \Delta x$, $\|\Sigma\delta'\| \leq \|x\|\gamma$, i.e.

$$\sigma_1^2\delta_1'^2 + \cdots + \sigma_m^2\delta_m'^2 + 0 \cdot (\sigma_{m+1}^2\delta_{m+1}'^2 + \cdots + \sigma_n^2\delta_n'^2) \leq \|x\|^2\gamma^2$$

However since $0 \cdot (\sigma_{m+1}^2\delta_{m+1}'^2 + \cdots + \sigma_n^2\delta_n'^2) = 0$ holds for any δ, i.e. the general solution to $Wa = Wb$ where $a \neq b$, we need not contain it in our perturbation region R_δ which is induced by $\|\Delta\| \leq \gamma$. Then, the above inequality represents a m-dim region bounded by a m-dim ellipsoid whose principal semi-axes have lengths $(\sigma_1\|x\|\gamma)^{-1}, \cdots, (\sigma_n\|x\|\gamma)^{-1}$ with respect to $\delta' \in \mathbb{R}^n$. Subsequently, the region of interest $R_\delta \in \mathbb{R}^n$ is an rotated m-dim ellipsoid whose principal semi-axes have lengths $(\sigma_1\|x\|\gamma)^{-1}, \cdots, (\sigma_n\|x\|\gamma)^{-1}$ with respect to $\delta \in \mathbb{R}^n$.

D Converting Perturbations in Input Space to Parameter Space

Given weights $W \in \mathbb{R}^{m \times n}$, input $x \in \mathbb{R}^n$, and parameter perturbation region $\|\delta\| \le \gamma$, we want to find the region R_Δ so that $\forall \ \|\delta\| \le \gamma, \exists \ \Delta \in R_\Delta \ s.t. \ W\delta = \Delta x$ and $\forall \ \Delta \in R_\Delta, \exists \ \|\delta\| \le \gamma \ s.t. \ W\delta = \Delta x$.

Using SVD decomposition, $W = U\Sigma V^\top$, where Σ is a diagonal matrix with entries $\sigma_1, \cdots, \sigma_n$.

$W\delta = U\Sigma V^\top \delta = U\Sigma\delta'$, where $\delta' := V^\top \delta$. Since rotating or reflecting does not change the region of a ball, $\|\delta\| \le \gamma$ gives $\|\delta'\| \le \gamma$, i.e. $\delta_1'^2 + \cdots \delta_n'^2 \le \gamma^2$.

Let $\delta'' := [\delta''_1, \cdots, \delta''_m] = \Sigma\delta' = [\sigma_1\delta'_1, \cdots, \sigma_m\delta'_m]$. $\forall i \in [m], \sigma_i^{-1}\delta''_i = \delta'_i$. Then,

$$\frac{\delta_1''^2}{\sigma_1^2} + \cdots + \frac{\delta_m''^2}{\sigma_m^2} \le \gamma^2 - \left(\delta_{m+1}'^2 + \cdots \delta_n'^2\right) \tag{1}$$

The maximum value of RHS in eq. 1 is γ^2, when $\left(\delta_{m+1}'^2 + \cdots \delta_n'^2\right) = 0$. This indicates that δ'' resides within an ellipsoid with with principle semi-axes of lengths $\lambda_i := \sigma_i\gamma, i \in [m]$. Thus, $U\delta'' = W\delta$ is a region bounded by an rotated ellipsoid.

Now, we will examine the region R_Δ such that Δx ($\Delta \in R_\Delta$) forms a rotated ellipsoid with principle semi-axes of lengths λ_i. Unlike the case of converting parameter space's perturbation region to input space's in Appendix B, R_Δ need not be in a form of ellipsoid. Instead, we provide a superset R_{sup} and a subset R_{sub} of R_Δ in the form of a ball such that $R_{sub} \subseteq R_\Delta \subseteq R_{sup}$.

Let W be deomposed into $U\Sigma V^\top$ using SVD decomposition. For now, we will consider the special case of W where $U = I$, i.e. the region of $W\delta$ is bounded by an ellipsoid alligned with standard basis. Afterwards, we will consider the general case of W, i.e. the region of $W\delta$ is bounded by a rotated ellipsoid.

Let d_{ij} denote the ith row, jth column element of $\Delta \in \mathbb{R}^{m \times n}$ and x_i the ith element of $x \in \mathbb{R}^n$. Since the range of Δx is an ellipsoid, Δx must satisfy the ellipsoid inequality

$$\frac{\left(x_1 d_{11} + x_2 d_{12} + \cdots + x_n d_{1n}\right)^2}{\lambda_1^2} + \cdots + \frac{\left(x_1 d_{m1} + x_2 d_{m2} + \cdots + x_n d_{1n}\right)^2}{\lambda_m^2} \le 1$$

Let r_i denote the ith row vector of Δ, and let X denote $xx^\top$. The above inequality can be rewritten as:

$$\frac{r_1^\top X r_1}{\lambda_1^2} + \frac{r_2^\top X r_2}{\lambda_2^2} + \cdots + \frac{r_m^\top X r_m}{\lambda_m^2} \le 1 \tag{2}$$

Since we are interested in finding the region of Δ in $\mathbb{R}^{m \times n}$ space, we may think of it as a vector $d = [r_1^\top, r_2^\top, \cdots, r_m^\top]$ in $\mathbb{R}^{(m \times n)}$ rather than as a matrix. Then, inequation 2 can be rewritten as:

$$d^\top X_\lambda d \le 1, \text{ where } X_\lambda := \begin{bmatrix} X/\lambda_1^2 & & & \\ & X/\lambda_2^2 & & \\ & & \cdots & \\ & & & X/\lambda_m^2 \end{bmatrix} \in \mathbb{R}^{(m \times n)^2}$$

One property of X_λ is that it is a rank m matrix with singular values $\|x\|^2/\lambda_1^2, \cdots, \|x\|^2/\lambda_m^2$, regarding that X/λ_i^2 is a rank 1 matrix with singular value $\|x\|^2/\lambda_i^2$. Another property is that X_λ is a positive-semidefinite matrix ($\because \forall i \in [m], \|x\|^2/\lambda_i^2 \ge 0$.)

When we think of a single input x, the area of d satisfying $d^\top X_\lambda d \le 1$ is not bounded. However, when we consider the constraint over multiple values of input datapoints $\{x_1, x_2, \cdots, x_N\} (N \gg n)$ that spans $\mathbb{R}^n$, the area becomes bounded. One justification of the multiple constraints is that when we consider x a uniform random variable over the input datapoints, the region of d that satisfies all the possible constraint is $\cup_{i=1}^N d^\top X_\lambda^{(i)} d \le 1$, where $X_\lambda^{(i)}$ denotes X_λ for $x = x_i$. Another justification is that when we reach a local plateau in training parameter W, there is little or no change in the value of W.

The following lemma and theorems provide a subset R_{sub} and superset R_{sup} of $R_{\triangle}$ in the form of balls in the parameter space.

Lemma 3. *Let R be the region of $x \in \mathbb{R}^n$ satisfying the inequality $x^\top Ax \leq 1$, where A is a non-zero positive semi-definite matrix having σ_{max} as the maximum nonzero singular value. Let R' be the region of $x \in \mathbb{R}^n$ satisfying the ineqaulity $x^\top x \leq \sigma_{max}^{-1}$. $R \subseteq R'$.*

Proof. We handle two cases where $rank(A) = m$ and $rank(A) < m$.

Case $rank(A) = m$**:**

Using SVD Decomposition, $A = U\Sigma U^\top$, where $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}$

$x^\top Ax = x^\top U\Sigma U^\top x = x'^\top \Sigma x' \leq 1$, where $x' := U^\top x$

Let x' be represented as $x' = [x'_1, \cdots, x'_n]$.

The constraint induced by R can be rewritten as:

$$x'^\top \Sigma x' = \sigma_1 x_1'^2 + \cdots + \sigma_n x_n'^2 \leq 1, \text{ where } \Sigma = U^\top AU$$

Let $x \in \mathbb{R}^n$ be some vector satisfyig $x^\top x \leq \sigma_{max}$. Since U is an orthogonal matrix and $x^\top x \leq \sigma_{max}^{-1}$ is an equidistant ball that is invariant under rotations and reflections, the constraint induced by R' can be rewritten as $x'^\top x' \leq \sigma_{max}^{-1}$, where $x' = U^\top x$.

To prove $x \in R'$ implies $x \in R$, we will show $x'^\top x' \leq \sigma_{max}^{-1}$ implies $x'^\top \Sigma x' \leq 1$.

$$x'^\top x' \leq \sigma_{max}^{-1} \iff \sigma_{max} x'^\top x' \leq 1$$

Let $\epsilon_i := \sigma_{max} - \sigma_i$. Then, $\forall i \in [n], \epsilon_i \geq 0$.

$$\begin{aligned} \sigma_{max} x'^\top x' - \sum_{i=1}^{n} \epsilon (x'_i)^2 &\leq 1 - \sum_{i=1}^{n} \epsilon_i (x'_i)^2 && \left(\because \sigma_{max} x'^\top x' \leq 1\right) \\ &\leq 1 && \left(\because \forall i \in [n], \epsilon_i (x'_i)^2 \geq 0\right) \end{aligned}$$

Case $rank(A) < m$**:**

Let $rank(A) = k < m$. A can be represented as $U\Sigma U^\top$ using SVD decomposition, where Σ is a diagonal matrix whose first k elements are non-zero singular values $\sigma_1, \cdots, \sigma_k$.

$$x^\top Ax = x^\top U\Sigma U^\top x = x'\Sigma x' \leq 1, \text{ where } \Sigma = U^\top AU \text{ and } x' := U^\top x$$

Let x' be represented as $[x'_1, \cdots, x'_n]$. The constrained induced by R can be rewritten as:

$$x'^\top \Sigma x' = \sigma_1 x_1'^2 + \cdots + \sigma_k x_k'^2 \leq 1$$

Let $x \in \mathbb{R}^n$ be any vector satisfying $x^\top x \leq \sigma_{max}^{-1}$. Since ball is equidistant, $x^\top x \leq \sigma_{max}^{-1} \iff x''^\top x' \leq \sigma_{max}^{-1}$, where $x' = U^\top x$.

To prove $x \in R'$ implies $x \in R$, we will show $x'^\top x' \leq \sigma_{max}^{-1}$ implies $x'\Sigma x' \leq 1$.

$$x'^\top x' \leq \sigma_{max}^{-1} \iff \sigma_{max} x'^\top x' \leq 1 \iff \sum_{i=1}^{n} \sigma_{max} (x'_i)^2 \leq 1$$

Let $\epsilon_i := \sigma_{max} - \sigma_i$. Then, $\forall i \in [n], \epsilon_i \geq 0$.

$$\begin{aligned}
\sum_{i=1}^{k} (\sigma_{max} - \epsilon_i) x_i'^2 &\leq \sigma_{max} x^\top x - \sum_{i=1}^{k} \epsilon_i (x_i')^2 && (\because \sum_{i=k+1}^{n} \sigma_{max} x_i^2 \geq 0) \\
&\leq 1 - \sum_{i=1}^{k} \epsilon_i (x_i')^2 && \left(\because \sigma_{max} x'^\top x' \leq 1\right) \\
&\leq 1 && \left(\because \forall i \in [k], \epsilon_i (x_i')^2 \geq 0\right)
\end{aligned}$$

Since $\sum\limits_{i=1}^{k} (\sigma_{max} - \epsilon_i) x_i'^2 = x'^\top \Sigma x', \ x'^\top \Sigma x' \leq 1$. □

Theorem 3. *Given $W \in \mathbb{R}^{m \times n}$, $D = \{x_1, \cdots, x_N\}(x_i \in \mathbb{R}^n / \{0\}$ for $i \in [N])$, and input perturbation region $\{\delta \in \mathbb{R}^n \mid \|\delta\| \leq \gamma\}$, let $x_{max} := \arg\max_{x_i} \|x_i\|$ and $\lambda_{min} := \min\{\lambda_1, \cdots, \lambda_m\}$. Then, $\{\Delta \in \mathbb{R}^{m \times n} \mid \|\Delta\| \leq (\|x_{max}\|^2 / \lambda_{min}^2)^{-1}\}$ is the subset of R_Δ.*

Proof. We will rewrite theorem 3 as the following statement:

Given a set of datapoints $D = \{x_1, x_2, \cdots, x_N\}(x_i \in \mathbb{R}^n / \{0\}, i \in [N])$, let R be the region of $d \in \mathbb{R}^{m \times n}$ satisfying the inequality $d^\top X_\lambda d \leq 1$ for all $x \in D$. Let R' be the region of $d \in \mathbb{R}^{m \times n}$ satisfying $d^\top d \leq (\|x_{max}\|^2 / \lambda_{min}^2)^{-1}$, where $x_{max} := \arg\max_{x_i} \|x_i\|$ and $\lambda_{min} := \min\{\lambda_1, \cdots, \lambda_m\}$. $R' \subseteq R$.

Remark that $X_\lambda^{(i)} = \begin{bmatrix} x_i^\top x_i / \lambda_1^2 & & & \\ & x_i^\top x_i / \lambda_2^2 & & \\ & & \cdots & \\ & & & x_i^\top x_i / \lambda_m^2 \end{bmatrix}$. $X_\lambda^{(i)}$ is a rank m matrix with singular values $\|x_i\|^2 / \lambda_1^2, \cdots, \|x_i\|^2 / \lambda_m^2$.

Let R_i denote the region of $d \in \mathbb{R}^n$ satisfying $d^\top X_\lambda^{(i)} d \leq 1$, and let R_i' denote the region $d^\top d \leq \left(\frac{\|x_i\|^2}{\lambda_{min}^2}\right)^{-1}$. $\frac{\|x_i\|^2}{\lambda_{min}^2}$ being the largest singular value of $X_\lambda^{(i)}$, $R_i' \subseteq R_i$ by Lemma 1. Since this holds for all $i \in [N]$, $\bigcup\limits_{i=1}^{N} R_i' \subseteq \bigcup\limits_{i=1}^{N} R_i$. $\bigcup\limits_{i=1}^{N} R_i = R$, and $\bigcup\limits_{i=1}^{N} R_i' = R'$ is a ball with smallest radius, i.e. $d^\top d \leq \left(\frac{\|x_{max}\|^2}{\lambda_{min}^2)^{-1}}\right)$. □

Theorem 4. *Given $W \in \mathbb{R}^{m \times n}$, $D = \{x_1, \cdots, x_N\}(x_i \in \mathbb{R}^n / \{0\}$ for $i \in [N])$, and input perturbation region $\{\delta \in \mathbb{R}^n \mid \|\delta\| \leq \gamma\}$, let $R_i := \{d \in \mathbb{R}^{m \times n} \mid d^\top X_\lambda^{(i)} d \leq 1\}$ and $\Gamma := \{R_i \mid i \in [N]\}$.Then, $\{\arg\min_{R_1, \cdots, R_n \in \Gamma} \max_{\rho \in \cup_{i \in [n]} R_i} \|\rho\|^2\}$ is the superset of R_Δ.*

Proof. Let $R_1^*, \cdots, R_n^*$ denote the elements of Γ satisfying $\arg\min_{R_1, \cdots, R_n \in \Gamma} \max_{\rho \in \cup R_i} \|\rho\|^2$.

$R = \bigcup\limits_{i=1}^{N} R_i \subseteq \bigcup\limits_{i=1}^{n} R_i^* \subseteq \max_{\rho \in \cup R_i} \|\rho\|^2$. □

We have so far addressed the case where $U = I$ for $W = U\Sigma V^\top$ in the equation $W\delta = \Delta x$. Now, let us consider the general case of full rank matrix W.

$\Delta \in \mathbb{R}^{m \times n}$ can be represented as either column vectors $[c_1, c_2, \cdots, c_n]$ or row vectors $[r_1, r_2, \cdots, r_m]^\top$. The equation $W\delta = \Delta x$ can be rewritten as:

$$\Sigma V^\top \delta = U^\top \Delta x = U^\top [c_1, c_2, \cdots, c_n] x = [U^\top c_1, U^\top c_2, \cdots, U^\top c_n] x$$

Let $\Delta' := U^\top \Delta = [c_1', c_2', \cdots, c_n'] = [r_1', r_2', \cdots, r_m']^\top$, and let d' be the flattened vector representation $[r_1'^\top, r_2'^\top, \cdots, r_m'^\top]$ of Δ'. Then, finding R_Δ is equivalent to finding the region of Δ' satisfying $d'^\top X_\lambda d' \leq 1$ and multiplying U to Δ'.

The relationship between Δ' and Δ can be expressed as:

$$U_{diag} \begin{bmatrix} c_1' \\ c_2' \\ \vdots \\ c_n' \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \text{ where } U_{diag} := \begin{bmatrix} U & & & \\ & U & & \\ & & \ddots & \\ & & & U \end{bmatrix} \in \mathbb{R}^{(m \times n)^2}$$

U_{diag} is an orthogonal matrix since U is an orthogonal matrix. Furthermore, any permutation π that permutes the row vectors of U_{diag} also results in another orthogonal matrix U^π_{diag}. Then for some π, $U^\pi_{diag}[r_1'^\top, r_2'^\top, \cdots, r_m'^\top]^\top = [r_1^\top, r_2^\top, \cdots, r_m^\top]^\top$, i.e. $U^\pi_{diag} d' = d$. Since the region of a ball is not affected by rotations or reflections, the superset and the subset obtained in Theorem 1 and 2 are not affected. In other words,

$$R_\Delta = \{d \in \mathbb{R}^{m \times n} \mid \forall i \in [n], d^\top X_\lambda^{(i)} d \leq 1\}$$

$$R_{sub} = \{d \in \mathbb{R}^{m \times n} \mid d^\top d \leq \left(\frac{\|x_{max}\|^2}{\lambda_{min}^2}\right)^{-1}\}$$

$$R_{sup} = \{d \in \mathbb{R}^{m \times n} \mid \operatorname*{arg\,min}_{R_1, \cdots, R_n \in \Gamma} \max_{\rho \in \cup R_i} \|\rho\|^2\}$$

satisfies $R_{sub} \subseteq R_\Delta \subseteq R_{sup}$.

Lastly, we provide the implications of our theorem:

Definition 3. (b-flat local minima) Given any real-valued loss function L and dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, a model parameter θ is said to have b-flat minima if the following conditions hold:

i) $\forall \|\epsilon\| \leq b, \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta), y)] = \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta+\epsilon), y)]$

ii) $\forall \|\epsilon\| > b, \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta), y)] < \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta+\epsilon), y)]$

Given any real-valued loss L and dataset D, let θ^* denote any optimal parameter that minimizes loss w.r.t. dataset, i.e. $\theta^* := \arg\min_\theta \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta), y)]$. Analogously, let θ^*_γ denote any optimal parameter such that $\forall \|\delta\| \leq \gamma, \theta^*_\gamma := \arg\min_\theta \mathbb{E}_{(x,y)\sim D}[L(f(x+\delta;\theta), y)]$.

The following holds for any linear classifier $f : x \mapsto \sigma(Wx + b)$:

Proposition 1. *θ^* can have 0-flat minima.*

Proof. There exists θ^* such that $\forall (x, y) \in D, \theta^* := \arg\min_\theta \mathbb{E}_{(x,y)\sim D}[L(f(x;\theta), y)]$ and $\forall \delta \in \mathbb{R}^m/\{0\}, L(f(x+\delta;\theta), y) > L(f(x;\theta), y)$. □

Corollary 1. *θ^*_γ has b-flat minima with $b \geq (\|x_{max}\|^2/\lambda^2_{min})^{-1}$.*

Proof. For θ^*_γ to be an optimal parameter, $\forall (x, y) \in D, \|\delta\| \leq \gamma, L(f(x+\delta;\theta^*_\gamma), y) = L(f(x;\theta^*_\gamma), y) = \min_\theta L(f(x;\theta), y)$. Using results from Theorem 3, $\forall (x, y) \in D, \|\Delta\| \leq (\|x_{max}\|^2/\lambda^2_{min})^{-1}, L(f(x;\theta^*_\gamma + \Delta), y) = L(f(x;\theta^*_\gamma), y) = \min_\theta L(f(x;\theta), y)$. □

The proposition and the collorary implies that exploiting perturbation-based algorithm will provide higher lower bound of b ($(\|x_{max}\|^2/\lambda^2_{min})^{-1} > 0$) for the b-flat minima of the optimal parameter.

E Additional Experimental Details

In this section, we provide an in-depth discussion on the experiments conducted in the main paper, as well as present additional experimental findings related to SPIDER including hyperparameter sensitivity analysis.

E.1 Experimental Details

In our experiments, the hardware resources employed differ based on the complexity of the tasks. All the tasks in the main paper related to CIFAR-10/100, and Tiny-ImageNet has been handled using 8 NVIDIA RTX A5000 GPUs. While a single A5000 GPU could have sufficed for the evaluation, the multi-GPU setup have been opted for to facilitate the extensive evaluations for each baseline methods. ImageNet training experiment has been carried out utilizing a single A100 GPU, owing to its superior computational capacity.

E.1.1 Main Table

Robustness Against General Corruptions

For CIFAR-10/100 experiments, we use WRN-40-2 architecture exploited in (Hendrycks et al., 2021b). For Tiny-ImageNet and ImageNet experiment, we use ResNet18 (He et al., 2015) as our backbone. SGD with momentum value of 0.9 has been used in all our experiments. Cosine learning rate decay scheduling (Loshchilov & Hutter, 2017) with initial learning rate of 0.1, 0.01, and 0.01 has been used respectively for CIFAR-10/100, Tiny-ImageNet, and ImageNet experiments to train a model until convergence. Models have been trained for 400, 100, and 90 epochs for CIFAR-10/100, Tiny-ImageNet, and ImageNet benchmarks respectively.

The search space for hyperparameters (τ, ξ) introduced by SPIDER instantiation are as follows. For τ, the range was set to $[0.0, 20.0]$ for CIFAR-10/100 and $[0.0, 30.0]$ for Tiny-ImageNet. ξ was tested within the range of $[0, 1 - c^{-1}]$, with c representing the number of classes in the given dataset. The rescaling and clipping algorithm by Rauber (Rauber & Bethge, 2020) was utilized to keep perturbed data points within a valid domain (i.e. $[0, 1]^m$).

Evaluation of Robustness to Common Data Corruptions

Benchmark Statistics: CIFAR-10/100-C, Tiny-ImageNet-C, and ImageNet (Deng et al., 2009) datasets contain 15 distinct corruption types: brightness changes, contrast alterations, defocus blur, elastic transformations, fog addition, frost addition, Gaussian blur, glass distortion, impulse noise, jpeg compression, motion blur, pixelation, shot noise, snow addition, and zoom blur with 5 different severity levels per each corruption. mCE calculates the average error of a model across all the distinct corruptions and severity levels. For ImageNet, we report mCE with normalization suggested as in Hendrycks & Dietterich (2019).

Evaluation Process: During training and validation, a model is trained and validated on the uncorrupted training and validation data. The validation data has been constructed using 20% of the training data. The model achieving the best validation accuracy is chosen and evaluated on a corrupted dataset, where the corruption types were not encountered either at the training or the validation stage. To reduce the discrepancy between the clean, non-augmented dataset and the corrupted dataset, we augment the clean validation data of CIFAR, Tiny-ImageNet, and ImageNet datasets using augmentations from (Mintun et al., 2021) during the validation phase. These augmentations are distinct from the corruptions used in the common corruption benchmarks (CIFAR/100-C, Tiny-ImageNet-C, and ImageNet-C). In short, we train a model using SPIDER on clean data, select the models with the highest accuracy on the augmented validation data (using functions from (Mintun et al., 2021)), and then assess these models' robustness on common corruption benchmarks.

Evaluation of Robustness to Adversarial Attacks

To assess the model's robustness against adversarial attacks, we use untargeted PGD attacks based on L_2 and L_∞ norms. We have chosen to use the absolute value α as the coefficient of gradient ascent for clarity, instead of using a relative step size with respect to ϵ. For L_2 attacks, PGD-20 attack with $(\epsilon, \alpha) = (0.5, 1/800)$ has been used. For L_∞ attacks, we have used PGD-7 attack with $(\epsilon, \alpha) = (8/255, 2/255)$ for CIFAR-10/100, PGD-3 attack with $(\epsilon, \alpha) = (3/255, 1/255)$ for the Tiny-ImageNet experiment, and PGD-2 L_∞ attack with $(\epsilon, \alpha) = (1/255, 1/510)$ for ImageNet experiment. Essentially, more intense attacks have been applied to simpler datasets, while milder attacks have been used for more complex datasets, with the CIFAR experiments' attack configurations borrowed from the (Yang et al., 2020). We then evaluate the adversarial robustness of models trained following the common corruption evaluation protocol detailed above. The key interest here is not to show that SPIDER is setting new records for robustness, but to demonstrate that SPIDER enhances both common corruption robustness and adversarial robustness compared to previous augmentation methods that had negligible impact on adversarial robustness.

E.1.2 Boundary Thickness

Our approach adheres closely to the original paper that introduced the boundary thickness metric (Yang et al., 2020). For each data point x_i in the dataset, labeled with one-hot encoded label i, we generate a corresponding adversarial instance x_j. This is achieved by conducting an attack on x_i targeting a randomly selected class j that differs from i. We use an L_2 PGD-20 untargeted attack with parameters $\epsilon = 5.0$ and $\alpha = 1.0$ to produce these adversarial instances. The integral $\int_0^1 I\{\alpha < g_{ij}x(t) < \beta\}dt$ is approximated by dividing the segment into 128 data points and determining the fraction of points that fall within the interval from α to β. For the purpose of measuring the mean and standard deviations of the boundary thickness, we generated 1600 data points, constructed from 50 batches of 32 images each, along with their adversarial counterparts. The data in Table 1 was calculated using baseline methods and SPIDER in combination with these baselines, using the weights obtained from the main experiment. Table 2 was calculated training the models that had identical configurations as the previous models, except with the smoothing function removed.

E.1.3 Flatness

Flatness is evaluated by sampling parameter fluctuations of growing radii and calculating the average loss on the model with the adjusted parameter. To elaborate, for every radius value, three independent and identical models, trained using either standalone baseline augmentation methods or a combination of SPIDER with the baselines, are utilized for flatness computation. For every model, 50 independent parameter perturbations are sampled and implemented (yielding a total of 150 disturbed weight samples) to determine the average loss related to the respective radius.

E.1.4 Ablation Table

The Baseline and SPIDER columns have statistics from the main table. The hyperparameter value (τ) of the removal of smoothing function (No LS) have been found following the same hyperparameter search space used in main table (Appendix E.1.1.) For standard label smoothing (STD LS), the degree of smoothing has been set as hyperparameter value and optimized using TPE sampler from Optuna library.

E.2 Additional Experiment

E.2.1 Hyperparameter Sensitivity

Analysis of Figure 5: The depicted charts illustrate the effect of perturbation sensitivity on the performance of the model trained with SPIDER augmentation solely. We gauged this sensitivity by keeping the shape of the exponential smoothing function constant - specifically, we set $s(\tau) = 0.5$ for the CIFAR-10/100 experiments, and observed the performance as the radius τ grew. As expected, an overly large perturbation radius risks pushing datapoints into the submanifolds of different labels, which degrades clean accuracy. Conversely, an excessively small perturbation radius fails to provide sufficient robustness enhancement, as evidenced by the elevated mCE and augmented clean error values for smaller radii. As the radius increases, there is an initial decrease in mCE, indicating increased robustness against corruptions. However, the trend begins to reverse for larger radii, with an accompanying rise in clean and augmented clean error. This pattern underscores the necessity of an appropriate balance in perturbation size to maintain performance across both clean and corrupted conditions. Despite these variations, the mean corruption error and augmented clean error consistently

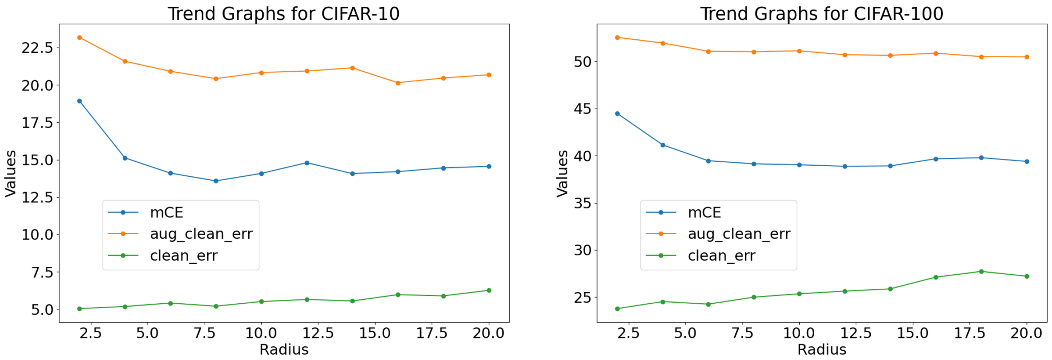

Figure 5: Sensitivity to increasing perturbations for CIFAR-10 and CIFAR-100. The plots depict the relationships between the radius of perturbation and the mean corruption error (mCE), augmented clean error (aug_clean_err), and clean error (clean_err) for the CIFAR-10 (left) and CIFAR-100 (right) datasets, with ξ = 0.4 and 0.49 respectively. Each point represents the error rate obtained with a different radius of parameter perturbation.

stay beneath the baseline's performance, thereby suggesting that SPIDER improves model robustness compared to the baseline approach.

Table 7: Comparison of training times and performance between baselines with the addition of SPIDER. The use of Optuna's automated hyperparameter search algorithm with pruning alleviates the added training cost.

Baselines	no aug.	AugMix	DeepAug.	PixMix
Pruned Training Duration(hour)	14.84	14.06	39.51	27.98
Predicted Full Training Duration (hour)	58.15	71.99	171.54	140.90
Speed Gain (x Faster)	3.92	5.12	4.34	5.04
Equivalent Non-Pruned Trials	7.65	5.86	6.91	5.96
Best Trial Index out of 30	2	3	2	5

Analysis of Table 7: The table compares the impact of incorporating SPIDER into various baseline augmentation methods in terms of training duration and performance. As the introduction of SPIDER brings additional hyperparameters for shaping the perturbation and smoothing functions, there is a potential for an increase in training time due to the associated hyperparameter search. To mitigate this, we have implemented Optuna's automated hyperparameter searching algorithm with an early stopping feature, also known as 'pruning'. The results show that the pruned training duration for each augmentation method (no augmentation, AugMix, DeepAugment, and PixMix) is significantly less than the predicted full training duration, indicating a substantial speed gain. The number of equivalent non-pruned trials ranges from approximately 5.86 to 7.65, implying that the use of pruning enables the same level of hyperparameter exploration to be achieved in a fraction of the time. The best trial index out of 30 shows that successful models can be identified relatively early in the process, further emphasizing the efficiency of the combined use of SPIDER with automated hyperparameter search and pruning. This approach, therefore, effectively alleviates the potentially increased training cost associated with the introduction of SPIDER's additional hyperparameters.

E.2.2 Varying Backbone Networks

SPIDER demonstrates substantial improvement in robustness against common corruptions and adversarial attacks across various neural network architectures. This supports previous research, which found that augmentation techniques enhancing robustness retain their effectiveness and influence on model resilience, regardless of differences in the underlying network structures (Hendrycks et al., 2021b;a).

Table 8: Evaluation of robustness over different backbone networks on CIFAR-100 benchmark.

Backbone	**Method**	**mCE** ↓		L_2 **(PGD)** ↓		L_∞ **(PGD)** ↓	
		original	+SPIDER	original	+SPIDER	original	+SPIDER
DenseNet	no aug.	57.67 $\pm$ 0.38	45.73 $\pm$ 0.56	99.44 $\pm$ 0.08	66.46 $\pm$ 0.38	100.00 $\pm$ 0.00	99.52 $\pm$ 0.06
	AugMix	41.79 $\pm$ 0.46	36.72 $\pm$ 0.15	96.94 $\pm$ 0.21	62.72 $\pm$ 0.43	99.98 $\pm$ 0.02	99.21 $\pm$ 0.09
	DeepAug.	44.92 $\pm$ 0.31	37.76 $\pm$ 0.29	97.37 $\pm$ 0.19	74.77 $\pm$ 0.26	89.05 $\pm$ 0.40	99.86 $\pm$ 0.01
	PixMix	35.71 $\pm$ 0.17	35.40 $\pm$ 0.07	99.97 $\pm$ 0.01	62.43 $\pm$ 0.19	99.99 $\pm$ 0.01	99.40 $\pm$ 0.05
WRN-40-1	no aug.	57.07 $\pm$ 0.48	43.56 $\pm$ 0.21	99.15 $\pm$ 0.06	77.74 $\pm$ 0.55	99.99 $\pm$ 0.01	99.61 $\pm$ 0.03
	AugMix	43.82 $\pm$ 0.27	35.92 $\pm$ 0.06	97.38 $\pm$ 0.09	86.91 $\pm$ 0.12	99.99 $\pm$ 0.01	99.26 $\pm$ 0.13
	DeepAug.	44.38 $\pm$ 0.33	36.36 $\pm$ 0.13	96.10 $\pm$ 0.21	79.34 $\pm$ 0.52	99.98 $\pm$ 0.01	99.60 $\pm$ 0.06
	PixMix	37.76 $\pm$ 0.29	37.48 $\pm$ 0.06	92.69 $\pm$ 0.35	87.86 $\pm$ 0.39	99.95 $\pm$ 0.01	99.90 $\pm$ 0.03
AllConvNet	no aug.	56.71 $\pm$ 0.11	43.35 $\pm$ 0.22	92.40 $\pm$ 0.15	66.50 $\pm$ 0.33	99.98 $\pm$ 0.01	99.50 $\pm$ 0.08
	AugMix	42.52 $\pm$ 0.43	35.92 $\pm$ 0.06	85.49 $\pm$ 0.42	62.83 $\pm$ 0.35	99.93 $\pm$ 0.02	99.84 $\pm$ 0.01
	DeepAug.	42.39 $\pm$ 0.18	36.39 $\pm$ 0.35	88.66 $\pm$ 0.31	74.84 $\pm$ 0.24	99.87 $\pm$ 0.02	99.86 $\pm$ 0.03
	PixMix	35.77 $\pm$ 0.01	33.84 $\pm$ 0.16	73.69 $\pm$ 0.28	62.37 $\pm$ 0.24	99.74 $\pm$ 0.01	99.39 $\pm$ 0.04

E.2.3 Varying Regularization Methods

Table 9: Evaluation of robustness over different regularizations on CIFAR-100 benchmark with DenseNet backbone.

Regularization	**Values**	**mCE** ↓	L_2 **(PGD)**	L_∞ **(PGD)** ↓	**Clean Acc.** ↑
Learning Rate	3e-03	53.38 $\pm$ 0.53	99.64 $\pm$ 0.10	99.99 $\pm$ 0.02	60.84 $\pm$ 0.31
	1e-02	46.99 $\pm$ 0.20	99.41 $\pm$ 0.10	99.99 $\pm$ 0.01	70.59 $\pm$ 1.82
	3e-02	43.81 $\pm$ 0.71	99.51 $\pm$ 0.10	99.99 $\pm$ 0.01	72.55 $\pm$ 0.52
Weight Decay(L_2)	0.0	44.98 $\pm$ 0.24	89.60 $\pm$ 0.40	99.83 $\pm$ 0.02	70.07 $\pm$ 0.11
	1e-4	43.49 $\pm$ 0.20	98.96 $\pm$ 0.08	99.99 $\pm$ 0.01	72.85 $\pm$ 0.35
L_1	5e-7	42.51 $\pm$ 0.12	99.48 $\pm$ 0.09	99.99 $\pm$ 0.01	74.35 $\pm$ 0.10
	2e-6	43.35 $\pm$ 0.07	99.99 $\pm$ 0.01	99.57 $\pm$ 0.13	73.91 $\pm$ 0.15
	5e-6	43.23 $\pm$ 0.30	99.63 $\pm$ 0.07	99.99 $\pm$ 0.01	73.61 $\pm$ 0.22
SPIDER (w/ best config)	-	36.90	81.80	99.95	84.65

Effective regularization techniques significantly enhance the robustness of models. The application of SPIDER, in conjunction with optimal regularization methods, notably amplifies this robustness across all measured metrics.

Additionally, it is worth noting that SPIDER, when optimally configured, not only strengthens the model's robustness against adversarial attacks but also tends to improve overall model accuracy. This dual benefit underscores SPIDER's effectiveness in balancing robustness with high performance in various challenging scenarios.