A APPENDIX

In the following sections B and C, we derive the regions of interest R_Δ, R_δ in the form of balls and rotated ellipsoidal inequalities. Beforehand, we assume W is full rank. If W is not full rank, it means that our linear model fails to distinguish two or more distinct classes for every possible input. During and at the end of training, we will assume such event does not occur. The initial values of W are selected randomly, and random matrices have full rank almost surely. Consequently, we will assume W is a full rank matrix henceforth.

B 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:

$\|\Delta\| \leq \gamma$

$\iff \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2 \leq \gamma^2$

$\iff \sum_{j=1}^{n} \|c_j\|^2 \leq \gamma_j^2$ 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 $\delta \in \mathbb{R}^n$.

C 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 $\|\delta\| \leq \gamma$, we want to find the region R_Δ so that $\forall\ \|\delta\| \leq \gamma, \exists\ \Delta \in R_\Delta\ s.t.\ W\delta = \Delta x$ and $\forall\ \Delta \in R_\Delta, \exists\ \|\delta\| \leq \gamma\ \ s.t.\ W\delta = \Delta x$.

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, $\|\delta\| \leq \gamma$ gives $\|\delta'\| \leq \gamma$, i.e. $\delta_1'^2 + \cdots \delta_n'^2 \leq \gamma^2$.

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,

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

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

$$\frac{(x_1d_{11} + x_2d_{12} + \cdots + x_nd_{1n})^2}{\lambda_1^2} + \cdots + \frac{(x_1d_{m1} + x_2d_{m2} + \cdots + x_nd_{1n})^2}{\lambda_m^2} \leq 1$$

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

$$\frac{r_1^\top X r_1}{\lambda_1^2} + \frac{r_2^\top X r_2}{\lambda_2^2} + \cdots + \frac{r_m^\top X r_m}{\lambda_m^2} \leq 1 \tag{3}$$

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:

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

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 ($\because \forall i \in [m], \|x\|^2/\lambda_i^2 \geq 0$.)

When we think of a single input x, the area of d satisfying $d^\top X_\lambda d \leq 1$ 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 $\cup_{i=1}^N d^\top X_\lambda^{(i)} d \leq 1$, 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_Δ in the form of balls in the parameter space.

Lemma 1. *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, $\forall i \in [n], \epsilon_i \geq 0$.

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

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

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 $\lambda_{min} := \min\{\lambda_1, \cdots, \lambda_m\}$. $\{\Delta \in \mathbb{R}^{m\times n} \mid \|\Delta\| \leq (\|x_{max}\|^2/\lambda_{min}^2)^{-1}\} \subseteq R_{\Delta}$*

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 $i \in [N]$, $\bigcup_{i=1}^N R_i' \subseteq \bigcup_{i=1}^N R_i$. $\bigcup_{i=1}^N R_i = R$, and $\bigcup_{i=1}^N R_i' = R'$ 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 $\Gamma := \{R_i \mid i \in [N]\}$. $R_\Delta \subseteq \{\arg\min_{R_1,\cdots,R_n\in\Gamma} \max_{\rho\in\cup_{i\in[n]}R_i} \|\rho\|^2\}$.*

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

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

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_{diag}^π. Then for some π, $U_{diag}^\pi [r_1'^\top, r_2'^\top, \cdots, r_m'^\top]^\top = [r_1^\top, r_2^\top, \cdots, r_m^\top]^\top$, i.e. $U_{diag}^\pi d' = d$. 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,

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

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

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

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

D PROOF OF THEOREM 1

Theorem 1. *For any x, let $n' \in \mathbb{R}^+$ be the number such that for all $M \geq n'$, $\mathbb{E}[y_{\delta 2} \mid X_\delta = x] \geq 0.5$ holds, and let $n \in \mathbb{R}^+$ be the number such that for all $M \geq n$, $\mathbb{E}[y_{E2} \mid X_E = x] \geq 0.5$ holds. Then, $n \geq n'$.*

Proof. **Case 1.** Let $z_1 := \|x_1 - x\| - \epsilon, z_2 := \|x_2 - x\| - \epsilon$. For the sake of simplicity, we will use $\mathcal{E}$ to denote the event $\{X_\delta = x, 0 < z_1 < z_2 < \tau\}$, and $p(z)$ to denote the truncated normal pdf $N_{truc}(z; \epsilon, \sigma, \epsilon + \tau, \epsilon - \tau)$.

$$\mathbb{E}[y_{\delta 2} \mid \mathcal{E}] = 0.5 \Longleftrightarrow \frac{\mathbb{E}[y_{\delta 1}|\mathcal{E}]}{\mathbb{E}[y_{\delta 2}|\mathcal{E}]} = 1$$

$$\frac{\mathbb{E}[y_{\delta 1}|\mathcal{E}]}{\mathbb{E}[y_{\delta 2}|\mathcal{E}]} = \frac{p(z_1)}{n' p(z_2)} = 1 \qquad n' = \frac{p(z_1)}{p(z_2)}.$$

$$\mathbb{E}[y_{E2} \mid \mathcal{E}] = 0.5 \Longleftrightarrow \frac{\mathbb{E}[y_{E1}|\mathcal{E}]}{\mathbb{E}[y_{E2}|\mathcal{E}]} = 1$$

$$\frac{\mathbb{E}[y_{E1}|\mathcal{E}]}{\mathbb{E}[y_{E2}|\mathcal{E}]} = \frac{e^{-\lambda z_1} p(z_1) + n \cdot \left(\frac{1 - e^{-\lambda z_2}}{C - 1} p(z_2) \right)}{n \cdot e^{-\lambda z_2} p(z_2) + \frac{1 - e^{-\lambda z_1}}{C - 1} p(z_1)} = 1$$

$$e^{-\lambda z_1} p(z_1) + n \cdot \frac{1 - e^{-\lambda z_2}}{C - 1} p(z_2) = n \cdot e^{-\lambda z_2} p(z_2) + \frac{1 - e^{-\lambda z_1}}{C - 1} p(z_1)$$

$$\left(e^{-\lambda z_1} - \frac{1 - e^{-}\lambda z_1}{C - 1} \right) p(z_1) = n \cdot \left(e^{-\lambda z_2} - \frac{1 - e^{-}\lambda z_2}{C - 1} \right) p(z_2)$$

$$n = \frac{\left(e^{-\lambda z_1} - \frac{1 - e^{-\lambda z_1}}{C - 1} \right) p(z_1)}{\left(e^{-\lambda z_2} - \frac{1 - e^{-\lambda z_2}}{C - 1} \right) p(z_2)} = \frac{B}{A} \cdot n'$$

$$B - A = e^{-\lambda z_1} - e^{-\lambda z_2} + \frac{1}{C - 1}(e^{-\lambda z_1} - 1 - e^{-\lambda z_2} + 1) = (e^{-\lambda z_1} - e^{-\lambda z_2}) \left(1 + \frac{1}{C - 1} \right)$$

$e^{-\lambda z_1} > e^{-\lambda z_2}$ ($\because 0 < z_1 < z_2 < \tau$), $1 + \frac{1}{C - 1} > 0$ ($\because C > 1$)

$\therefore B - A > 0$

$e^{-\lambda z} - \frac{1 - e^{-\lambda z}}{C - 1} = \frac{1}{C - 1}(C e^{-\lambda z} - 1) > 0$ ($\because c > 1$ and $e^{-\lambda z} > \frac{1}{C - 1}$ for $0 < z < \tau$)

$\therefore B > A > 0$

$n = \frac{B}{A} n' > n'.$

Case 2. Let $z_1 := \|x_1 - x\|, z_2 := \|x_2 - x\| - \epsilon$. For the sake of simplicity, we will use $\mathcal{E}$ to denote the event $\{X_\delta = x, 0 < z_1 < \epsilon, 0 < z_2 < \tau\}$, and $p(z)$ to denote the truncated normal pdf $N_{truc}(z; \epsilon, \sigma, \epsilon + \tau, \epsilon - \tau)$.

$$\frac{\mathbb{E}[y_{\delta 1}|\mathcal{E}]}{\mathbb{E}[y_{\delta 2}|\mathcal{E}]} = \frac{\frac{1}{n'+1}p(z_1)}{\frac{n'}{n'+1}p(z_2+\epsilon)} = 1 \quad n' = \frac{p(z_1)}{p(z_2+\epsilon)}$$

$$\frac{\mathbb{E}[y_{E1}|\mathcal{E}]}{\mathbb{E}[y_{E2}|\mathcal{E}]} = \frac{\frac{1}{n+1}p(z_1) + \frac{n}{n+1}p(z_2+\epsilon)\left(\frac{1-e^{-\lambda x_2}}{C-1}\right)}{\frac{n}{n+1}p(z_2+\epsilon)e^{-\lambda z_2}} = 1$$

$$p(z_1) + n \cdot p(z_2+\epsilon)\frac{1-e^{-\lambda z_2}}{C-1} = n \cdot p(z_2+\epsilon)e^{-\lambda z_2}$$

$$p(z_1) = n \cdot p(z_2+\epsilon)\left(e^{-\lambda z_2} - \frac{1-e^{-\lambda z_2}}{C-1}\right)$$

$$n = \frac{1}{e^{-\lambda z_2} - \frac{1-e^{-\lambda z_2}}{C-1}} \cdot \frac{p(z_1)}{p(z_2+\epsilon)} > \frac{e^{-\lambda z_1} - \frac{1-e^{-\lambda z_1}}{C-1}}{e^{-\lambda z_2} - \frac{1-e^{-\lambda z_2}}{C-1}} = \frac{B}{A} \cdot n' > n'$$

Case 3. Let $z_1 := \|x_1 - x\|, z_2 := \|x_2 - x\|$. Suppose an event $0 < z_1, z_2 < \epsilon$ has occurred.

Since $s(z) = 1$ for $0 < z_1, z_2 < \epsilon$, $n = n'$. □

E EXPERIMENT DETAILS

We have used grid search to find the optimal hyperparameter configuration per each augmentation method. The optimal hyperparameter configuration for AugMix (Hendrycks et al., 2021b) ($\alpha = 1$) is included in the search space. For Tiny-ImageNet-C experiment, we have carried out experiments applying the default parameter setting of AugMix (Hendrycks et al., 2021b) and DeepAugment (Hendrycks et al., 2021a) with 200 epochs.

Table 2: The search space of baseline methods and the ensemble of the baseline methods. The best hyperparameter configurations are marked in bold.

Dataset	Augmentation	Hyperparameter	Max Epoch
MNIST-C	No augmentation	N/A	[**100**, 200, 400]
	AugMix	α=[**0.25**, 0.5, 0.75, 1, 1.5, 2]	[100, 200, **400**]
	DeepAugment	N/A	[100, 200, **400**]
	AugMix + DeepAugment	α=[**0.25**, 0.5, 0.75, 1, 1.5, 2]	[100, 200, **400**]
CIFAR-10-C	No augmentation	N/A	[100, 200, **400**]
	AugMix	α=[0.25, 0.5, 0.75, **1**, 1.5, 2]	[100, **200**, 400]
	DeepAugment	N/A	[100, **200**, 400]
	AugMix + DeepAugment	α=[0.25, 0.5, 0.75, 1, 1.5, 2]	[100, 200, **400**]
CIFAR-100-C	No augmentation	N/A	[100, 200, **400**]
	AugMix	α=[0.25, 0.5, 0.75, **1**, 1.5, 2]	[100, 200, **400**]
	DeepAugment	N/A	[100, 200, **400**]
	AugMix + DeepAugment	α=[0.25, 0.5, 0.75, **1**, 1.5, 2]	[**100**, 200, 400]

Table 3: The search space of baseline methods and the ensemble of the baseline methods combined with random noise with fixed L_2 distance. The best hyperparameter configurations are marked in bold.

Dataset	Aug.	Radius	Max Epoch
MNIST-C	-	[0.1, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, **7.0**, 8.0, 9.0, 10.0]	[**100**, 200, 400]
	A	[0.1, 0.5, 1.0, 2.0, **3.0**, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]	[100, **200**, 400]
	D	[0.1, 0.5, **1.0**, 2.0, 3.0, 4.0, 5.0, 10.0]	[100, 200, **400**]
	A+D	[0.1, 0.5, **1.0**, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]	[100, 200, **400**]
CIFAR-10-C	-	[0.1, 0.5, 1.0, 2.0, **3.0**, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]	[100, **200**, 400]
	A	[0.1, **0.5**, 0.6, 0.7, 0.8, 0.9, 1.0 5.0, 10.0]	[**100**, 200, 400]
	D	[0.1, 0.5, 0.6, 0.8, **1.0**, 1.2, 1.4, 1.6, 2.0, 5.0]	[100, **200**, 400]
	A+D	[0.02, 0.04, 0.06, 0.08, **0.10**, 0.20, 0.50, 1.0, 5.0]	[**100**, 200, 400]
CIFAR-100-C	-	[0.1, 0.3, 0.5, 1.0, 2.0, **3.0**, 4.0, 5.0, 6.0, 7.0]	[100, 200, **400**]
	A	[0.1, 0.5, 0.6, **0.7**, 0.8, 0.9, 1.0, 3.0, 5.0]	[100, **200**, 400]
	D	[0.1, 0.3, 0.5, 0.7, 1.0, **1.2**, 1.4, 1.6, 1.8, 2.0]	[100, **200**, 400]
	A+D	[**0.01**, 0.02, 0.03, 0.04, 0.05, 0.1, 0.5, 1.0]	[**100**, 200, 400]
Tiny-IN-C	-	[0.1, 0.5, 1.0, 5.0, **10.0**]	[**200**]
	A	[**0.1**, 0.5, 1.0, 5.0, 10.0]	[**200**]
	D	[**0.1**, 0.5, 1.0, 5.0, 10.0]	[**200**]
	A+D	[0.1, 0.5, **1.0**, 5.0, 10.0]	[**200**]

Besides, the authors of AugMix have proposed an additional Jenson-Shannon divergence(JSD) loss term defined between the original image x and two augmented images $x_{\text{augmix1}}, x_{\text{augmix2}}$. Given the original loss term $L(\hat{p}(y \mid x), y)$, AugMix suggests to minimize the additional JSD loss term to further increase model robustness:

$$L_{\text{aug}} := L(\hat{p}(y \mid x), y) + \lambda \, \text{JSD}(\hat{p}(y \mid x); \hat{p}(y \mid x_{\text{augmix1}}); \hat{p}(y \mid x_{\text{augmix2}}))$$

The value of λ has been decided empirically for CIFAR-10/100-C experiments in the original paper ($\lambda = 12.$) Unlike CIFAR-10/100-C experiments, we have found that the additional JSD loss term in fact damaged the model robustness in the MNIST-C benchmark. Furthermore, the additional JSD loss term makes training longer. Hence, we have minimized the additional JSD term only in the CIFAR-10/100-C experiments.

Inspired by the additional robustness gain induced by the JSD term, we have minimized the following objective loss function when we combined ESP with AugMix:

$$L_{\text{ensemble}} = L_{\text{augmix}} + \beta(L(\hat{p}(y \mid x_{\text{ESP1}}), y) + L(\hat{p}(y \mid x_{\text{ESP2}}), y)) + \gamma JSD(x_{\text{ESP1}}; x_{\text{ESP2}})$$

We have chosen the values of β and γ to be $1/8$ and 1, of which the values have been determined experimentally analogy to AugMix.

Table 4: The search space of ESP and the ensemble of the baseline methods combined with ESP. The best hyperparameter configurations are marked in bold.

Dataset	**Aug.**	(ϵ, τ)	ξ	**Epoch** (10^2)
MNIST-C	-	[(4.0, 3.8), **(5.0, 4.8)**, (6.0, 5.8), (7.0, 6.8), (8.0, 7.8)] $\times 10^0$	[0.42, **0.59**]	[**1**, 2, 4]
	A	[(1.0, 0.8), (2.0, 1.8), **(3.0, 2.8)**, (4.0, 3.8), (5.0, 4.8)] $\times 10^0$	[**0.42**, 0.59]	[1, 2, **4**]
	D	[**(0.6, 0.5)**, (0.7, 0.6), (0.8, 0.7), (0.9, 0.8), (1.0, 0.9)] $\times 10^0$	[0.42, **0.59**]	[1, 2, **4**]
	A+D	[**(0.5, 0.4)**, (0.6, 0.5), (0.7, 0.6), (0.8, 0.7), (0.9, 0.8), (1.0, 0.9)] $\times 10^0$	[**0.42**, 0.59]	[**1**, 2, 4]
CIFAR-10-C	-	[(1.0, 0.8), (2.0, 1.8), **(3.0, 2.8)**, (4.0, 3.8), (5.0, 4.8)] $\times 10^0$	[0.42, **0.59**]	[1, **2**, 4]
	A	[(3.0, 2.8), (4.0, 3.8), (5.0, 4.8), **(6.0, 5.8)**, (7.0, 6.8)] $\times 10^0$	[**0.42**, 0.59]	[1, 2, **4**]
	D	[(0.6, 0.5), (0.8, 0.7), (1.0, 0.8), (1.2, 1.0), **(1.4, 1.2)**] $\times 10^0$	[0.42, **0.59**]	[1, **2**, 4]
	A+D	[(0.2, 0.1), (0.4, 0.3), (0.6, 0.5), **(0.8, 0.7)**, (1.0, 0.9)] $\times 10^{-1}$	[**0.42**, 0.59]	[1, **2**, 4]
CIFAR-100-C	-	[(1.0, 0.8), (2.0, 1.8), **(3.0, 2.8)**, (4.0, 3.8), (5.0, 4.8)] $\times 10^0$	[0.16, 0.33, **0.49**]	[1, 2, **4**]
	A	[(0.5, 0.4), (0.6, 0.5), **(0.7, 0.6)**, (0.8, 0.7), (1.0, 0.8)] $\times 10^0$	[0.16, **0.33**, 0.49]	[1, 2, **4**]
	D	[(0.8, 0.6), (1.0, 0.8), (1.2, 1.0), **(1.4, 1.2)**, (1.6, 1.4)] $\times 10^0$	[0.16, 0.33, **0.49**]	[1, **2**, 4]
	A+D	[(0.6, 0.3), (0.8, 0.4), (1.0, 0.5), (1.2, 0.6), **(1.4, 0.7)**] $\times 10^{-2}$	[**0.16**, 0.33, 0.49]	[**1**, 2, 4]
Tiny-IN-C	-	[(0.1, 0.08), (0.5, 0.4), **(1.0, 0.8)**, (5.0, 4.0), (10.0, 8.0)] $\times 10^0$	[**0.50**]	[**2**]
	A	[(0.1, 0.08), **(0.5, 0.4)**, (1.0, 0.8), (5.0, 4.0), (10.0, 8.0)] $\times 10^0$	[**0.50**]	[**2**]
	D	[**(0.1, 0.08)**, (0.5, 0.4), (1.0, 0.8), (5.0, 4.0), (10.0, 8.0)] $\times 10^0$	[**0.50**]	[**2**]
	A+D	[**(0.1, 0.08)**, (0.5, 0.4), (1.0, 0.8), (5.0, 4.0), (10.0, 8.0)] $\times 10^0$	[**0.50**]	[**2**]

Lastly, we have measured the running time of each augmentation method per epoch in seconds. DeepAugment augments the original image using pretrained image-to-image models by either randomly perturbing the models' weights or changing the actication functions of the models. As the perturbation operation itself is time-consuming due to its dependency on image-to-image models, the authors have trained a target model with stored perturbation results along with the original dataset. In our experiment, perturbing a single MNIST dataset has required more than an hour. Consequently, we follow the training scheme proposed in the original paper. Despite the fact that we have stored and loaded the perturbed images, the training time of ESP was shorter than the DeepAugment.

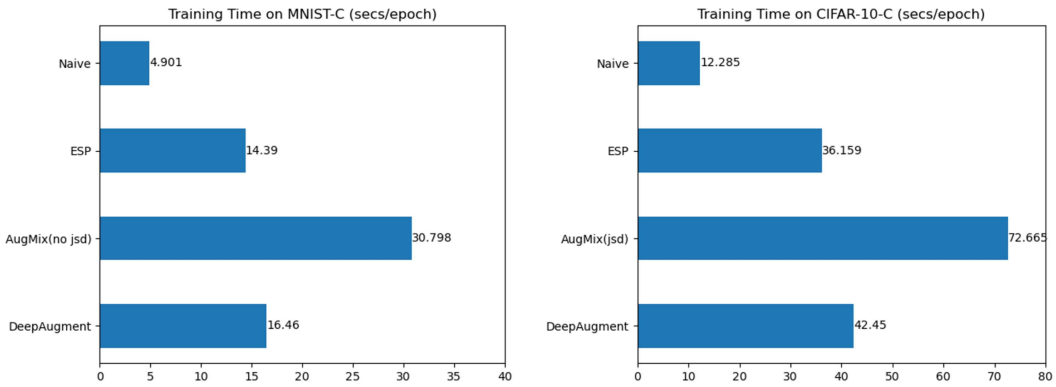

Figure 5: Training time comparison of augmentation methods on MNIST-C and CIFAR-10-C benchmarks. ESP exhibits the fastest running time compared to the other baseline methods. The training time is averaged over 5 epochs.

Table 5: Clean error over MNIST-C, CIFAR-10/100-C, and Tiny-ImageNet-C benchmarks. The reported values are the average clean test error with three individual runs for each method. Best results are marked in bold.

Augmentation	**MNIST-C**	**CIFAR-10-C**	**CIFAR-100-C**	**Tiny-IN-C**
Naive	1.43 ± 0.48	4.62 ± 0.07	**22.79** ± 0.14	40.69 ± 0.05
Naive + L_2	0.65 ± 0.18	7.32 ± 0.12	30.54 ± 0.27	40.52 ± 0.45
Naive + ESP	**0.48** ± 0.02	8.02 ± 0.20	30.56 ± 0.47	43.36 ± 0.54
AugMix	0.77 ± 0.03	4.45 ± 0.05	22.92 ± 0.20	39.62 ± 0.27
AugMix + L_2	0.77 ± 0.07	4.53 ± 0.10	23.78 ± 0.06	39.97 ± 0.23
AugMix + ESP	0.87 ± 0.01	**4.34** ± 0.10	23.03 ± 0.21	41.25 ± 0.30
DeepAugment	0.98 ± 0.06	5.10 ± 0.21	24.07 ± 0.15	39.96 ± 0.03
DeepAugment + L_2	1.05 ± 0.04	5.86 ± 0.11	27.68 ± 0.15	39.93 ± 0.29
DeepAugment + ESP	1.07 ± 0.03	6.14 ± 0.05	26.30 ± 0.09	37.35 ± 0.07
AugMix + DeepAug	1.12 ± 0.07	4.77 ± 0.12	24.11 ± 0.22	39.21 ± 0.62
AugMix + DeepAug + L_2	1.14 ± 0.05	4.86 ± 0.16	24.06 ± 0.14	39.30 ± 0.16
AugMix + DeepAug + ESP	1.12 ± 0.10	4.79 ± 0.04	24.88 ± 0.22	**37.91** ± 0.05

F Additional Experiment

In addition to MNIST-C, CIFAR-10/100-C, and Tiny-ImageNet-C benchmark, we provide partial experiment on ImageNet-C benchmark to show that ESP is effective way to enhance model robustness in ImageNet-C benchmark as well. ResNet18 has been employed with the same cosine annealing scheduling as in 4.2.

Table 6: ImageNet-C experiment results.

Augmentation	Naive	ESP
Avg. Corruption Error	70.24	60.28