

APPENDIX

A CALCULATION OF GENUINE ADVERSARIAL ACCURACY ON ONE-DIMENSIONAL TOY EXAMPLE

Here, we explain the calculation of genuine adversarial accuracies for $f_1(x)$, $f_2(x)$ and $f_3(x)$ (see Figure 6). First, we calculate required sets and regions. The topological closure of \mathcal{X} is $\bar{\mathcal{X}} = [-2, -1] \cup [1, 2]$. Voronoi boundary $VB(\mathcal{X}) = \{0\}$. When $0 < \epsilon < 1$, previously allowed perturbation region $\mathcal{X}_\epsilon = (-2 - \epsilon, -1 + \epsilon) \cup (1 - \epsilon, 2 + \epsilon)$ and $S_{exact}(\epsilon) = \{-2, -1, 1, 2\}$. When $\epsilon \geq 1$, previously allowed perturbation region $\mathcal{X}_\epsilon = (-2 - \epsilon, 0) \cup (0, 2 + \epsilon)$ and $S_{exact}(\epsilon) = \{-2, 2\}$. For calculation of genuine adversarial accuracies, we will consider four points $-2 - \epsilon, -1 + \epsilon, 1 - \epsilon$ and $2 + \epsilon$ when $0 < \epsilon < 1$, and two points $-2 - \epsilon$ and $2 + \epsilon$ when $\epsilon \geq 1$. (If we use the definition of $R_{gen;exact}(\epsilon)$ which will be introduced in Lemma 3 in Section B, $R_{gen;exact}(\epsilon) = \{-2 - \epsilon, -1 + \epsilon, 1 - \epsilon, 2 + \epsilon\}$ when $0 < \epsilon < 1$, and $R_{gen;exact}(\epsilon) = \{-2 - \epsilon, 2 + \epsilon\}$ when $\epsilon \geq 1$.) Note that if we did not use closure $\bar{\mathcal{X}}$ in the definition of $S_{exact}(\epsilon)$, $S_{exact}(\epsilon) = \{-2, 1\}$ and we will only consider points $-2 - \epsilon$ and $1 - \epsilon$ when $0 < \epsilon < 1$. Likewise, when $\epsilon \geq 1$, $S_{exact}(\epsilon) = \{-2\}$ and we will only consider one point $-2 - \epsilon$. This will ignore many points and can not measure the proper robustness of classifiers.

In the change of genuine adversarial accuracy for $f_1(x)$ (shown in Figure 6. $f_1(x)$ is shown in Figure 3.), when $0 < \epsilon < 1$, points $-2 - \epsilon, -1 + \epsilon$ and a point $2 + \epsilon$ will be non-adversarial perturbed samples and $1 - \epsilon$ will be adversarial example (Biggio et al., 2013), and thus $a_{gen;exact}(\epsilon) = \frac{3}{4} = 0.75$. When $\epsilon \geq 1$, points $-2 - \epsilon$ and $2 + \epsilon$ will be non-adversarial perturbed samples, and thus its genuine adversarial accuracy is 1.

When considering the change of genuine adversarial accuracy for $f_2(x)$ (shown in Figure 6. $f_2(x)$ is shown in Figure 3.), for $0 < \epsilon < 1$, points $-2 - \epsilon, -1 + \epsilon, 1 - \epsilon$ and $2 + \epsilon$ will be non-adversarial perturbed samples, and thus $a_{gen;exact}(\epsilon) = 1$. When $1 \leq \epsilon \leq 2$, points $-2 - \epsilon$ and $2 + \epsilon$ will be non-adversarial perturbed samples, and thus $a_{gen;exact}(\epsilon) = 1$. However, when $\epsilon > 2$, only one point $2 + \epsilon$ will be non-adversarial perturbed samples, and the other point $-2 - \epsilon$ will be adversarial example (Biggio et al., 2013), and thus $a_{gen;exact}(\epsilon) = \frac{1}{2} = 0.5$.

Through a similar process, one can understand the change of genuine adversarial accuracy for $f_3(x)$.

B LEMMAS USED IN THE PROOFS OF THEOREMS

Lemma 1. Let $A_\epsilon = \{x'' \in \mathbb{R}^d | \exists x_{clean} \in \bar{\mathcal{X}} : \|x'' - x_{clean}\| < \epsilon\}$. Then, the following holds.

$$\mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c = A_\epsilon^c \cap VB(\mathcal{X})^c \quad (1)$$

Proof.

$\mathcal{X}_\epsilon = A_\epsilon \cap VB(\mathcal{X})^c$ (\because The definition of \mathcal{X}_ϵ)

$$\begin{aligned} \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c &= (A_\epsilon^c \cup VB(\mathcal{X})) \cap VB(\mathcal{X})^c \quad (\because \text{De Morgan's law}) \\ &= (A_\epsilon^c \cap VB(\mathcal{X})^c) \cup (VB(\mathcal{X}) \cap VB(\mathcal{X})^c) \quad (\because \text{Distributive law}) \\ &= (A_\epsilon^c \cap VB(\mathcal{X})^c) \cup \emptyset = A_\epsilon^c \cap VB(\mathcal{X})^c \end{aligned}$$

□

Lemma 2. If $\epsilon > 0$, then the following holds.

$$x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c \iff \forall x_{clean} \in \bar{\mathcal{X}} : \|x' - x_{clean}\| \geq \epsilon \text{ and } x' \in VB(\mathcal{X})^c \quad (2)$$

Proof.

$$\begin{aligned} x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c &\iff x' \in A_\epsilon^c \cap VB(\mathcal{X})^c \quad (\because \text{Equation (1) of Lemma 1}) \\ &\iff x' \notin \{x'' \in \mathbb{R}^d | \exists x_{clean} \in \bar{\mathcal{X}} : \|x'' - x_{clean}\| < \epsilon\} \text{ and } x' \in VB(\mathcal{X})^c \\ &\iff x' \in \{x'' \in \mathbb{R}^d | \nexists x_{clean} \in \bar{\mathcal{X}} : \|x'' - x_{clean}\| < \epsilon\} \text{ and } x' \in VB(\mathcal{X})^c \\ &\iff x' \in \{x'' \in \mathbb{R}^d | \forall x_{clean} \in \bar{\mathcal{X}} : \|x'' - x_{clean}\| \geq \epsilon\} \text{ and } x' \in VB(\mathcal{X})^c \\ &\iff \forall x_{clean} \in \bar{\mathcal{X}} : \|x' - x_{clean}\| \geq \epsilon \text{ and } x' \in VB(\mathcal{X})^c \end{aligned}$$

□

Lemma 3. When $\epsilon > 0$, by changing ϵ , x' that satisfies $x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c$ and $\|x' - x\| = \epsilon$ can fill up $\mathbb{R}^d - \bar{\mathcal{X}}$ except for $VB(\mathcal{X})$. In other words, $(\mathbb{R}^d - \bar{\mathcal{X}}) - VB(\mathcal{X}) \subset \bigcup_{\epsilon > 0} R_{gen;exact}(\epsilon)$ where

$R_{gen;exact}(\epsilon)$ is the region of points that will be used for calculating genuine adversarial accuracy for ϵ . It is defined as follows.

$$R_{gen;exact}(\epsilon) = \begin{cases} \{x' \in \mathbb{R}^d \mid \|x' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \cap \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c, & \text{when } \epsilon > 0, \\ \mathcal{X}, & \text{when } \epsilon = 0. \end{cases}$$

Proof. Part 1

As \mathcal{X} is a nonempty set, $\exists x \in \bar{\mathcal{X}}$.

$$x' \in \mathbb{R}^d - \bar{\mathcal{X}} \implies (\|x' - x\| = c > 0) \vee (\|x' - x\| = 0) \quad (\because \text{Non-negativity axiom of a metric})$$

$$\implies (\|x' - x\| = c > 0) \vee (x' = x \in \bar{\mathcal{X}})$$

$$(\because \text{Identity of indiscernibles axiom of a metric})$$

$$\implies (\|x' - x\| = c > 0) \vee (x' \in \bar{\mathcal{X}} \cap (\mathbb{R}^d - \bar{\mathcal{X}}) = \emptyset) \implies \|x' - x\| = c > 0$$

$$\implies x' \in \{x'' \in \mathbb{R}^d \mid \|x'' - x_{clean}\| = c \text{ where } x_{clean} \in \bar{\mathcal{X}}\} \text{ for } c > 0$$

$$\implies x' \in \bigcup_{\epsilon > 0} \{x'' \in \mathbb{R}^d \mid \|x'' - x_{clean}\| = \epsilon \text{ where } x_{clean} \in \bar{\mathcal{X}}\}$$

$$\implies x', \exists \epsilon > 0 : \|x' - x_{clean}\| = \epsilon \text{ where } x_{clean} \in \bar{\mathcal{X}}$$

$$\text{Let } \epsilon_{min;x'} = \min_{x_{clean} \in \bar{\mathcal{X}}} \|x' - x_{clean}\| > 0.$$

$$\implies x' \in \{x'' \in \mathbb{R}^d \mid \|x'' - x_{clean}\| = \epsilon_{min;x'} \text{ where } x_{clean} \in \bar{\mathcal{X}}\} \cap A_{\epsilon_{min;x'}}^c$$

$$(\because x' \notin A_{\epsilon_{min;x'}} \text{ because of the definition of } \epsilon_{min;x'})$$

$$\implies x' \in \bigcup_{\epsilon > 0} (\{x'' \in \mathbb{R}^d \mid \|x'' - x_{clean}\| = \epsilon \text{ where } x_{clean} \in \bar{\mathcal{X}}\} \cap A_\epsilon^c)$$

We proved the following relation.

$$\mathbb{R}^d - \bar{\mathcal{X}} \subset \bigcup_{\epsilon > 0} (\{x'' \in \mathbb{R}^d \mid \|x'' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \cap A_\epsilon^c) \quad (3)$$

Part 2

We finalize the proof of the Lemma 3.

$$\begin{aligned} \bigcup_{\epsilon > 0} R_{gen;exact}(\epsilon) &= \bigcup_{\epsilon > 0} \{x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c \mid \|x' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \\ &= \bigcup_{\epsilon > 0} \{x' \in A_\epsilon^c \cap VB(\mathcal{X})^c \mid \|x' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \\ &\quad (\because \text{Equation (1) of Lemma 1}) \\ &= \bigcup_{\epsilon > 0} (\{x' \in \mathbb{R}^d \mid \|x' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \cap A_\epsilon^c \cap VB(\mathcal{X})^c) \\ &= \left\{ \bigcup_{\epsilon > 0} (\{x' \in \mathbb{R}^d \mid \|x' - x\| = \epsilon \text{ where } x \in \bar{\mathcal{X}}\} \cap A_\epsilon^c) \right\} \cap VB(\mathcal{X})^c \\ &\quad (\because \text{Distributive law}) \\ &\supset (\mathbb{R}^d - \bar{\mathcal{X}}) - VB(\mathcal{X}) \quad (\because \text{Relation (3)}) \end{aligned}$$

□

Combing the fact that $R_{gen;exact}(0) = \mathcal{X}$ and Lemma 3 results in $\mathbb{R}^d - (\bar{\mathcal{X}} - \mathcal{X}) - VB(\mathcal{X}) \subset \bigcup_{\epsilon \geq 0} R_{gen;exact}(\epsilon)$. It indicates that even though genuine adversarial accuracy does not allow overlaps

by Theorem 1, in practice, genuine adversarial accuracy uses almost all points in \mathbb{R}^d by changing ϵ (Note that, in l_2 norm, $\bar{\mathcal{X}} - \mathcal{X}$ and $VB(\mathcal{X})$ are regions with measure zero in practice.).

C PROOF OF THEOREM 1 (NO OVERLAP IN GENUINE ADVERSARIAL ACCURACY)

Proof. Part 1

First, we prove that the regions of points will be used for calculation of genuine adversarial accuracy for different ϵ values have no intersection. We need to prove the following when we use the definition of $R_{gen;exact}(\epsilon)$ introduced in Lemma 3.

$$\epsilon_1 \neq \epsilon_2 \implies R_{gen;exact}(\epsilon_1) \cap R_{gen;exact}(\epsilon_2) = \emptyset \quad (4)$$

Let $x' \in R_{gen;exact}(\epsilon_1), R_{gen;exact}(\epsilon_2)$ for $\epsilon_1 \neq \epsilon_2$.

First, we consider when $\epsilon_1, \epsilon_2 > 0$.

Then, $\exists x_1 \in \bar{\mathcal{X}} : \|x' - x_1\| = \epsilon_1$ and $\exists x_2 \in \bar{\mathcal{X}} : \|x' - x_2\| = \epsilon_2$.

If $x_1 = x_2$, then $\|x' - x_1\| = \epsilon_1 \neq \epsilon_2 = \|x' - x_1\|$. It is a contradiction.

We now consider the case when $x_1 \neq x_2$. Without loss of generality, we can assume $\epsilon_1 < \epsilon_2$.

As $x' \in R_{gen;exact}(\epsilon_2)$, $x' \in \mathcal{X}_{\epsilon_2}^c \cap VB(\mathcal{X})^c$.

$\forall x_{clean} \in \bar{\mathcal{X}} : \|x' - x_{clean}\| \geq \epsilon_2$ and $x' \in VB(\mathcal{X})^c$ (\because Equivalence relation (2) of Lemma 2 in Section B)

As $x_1 \in \bar{\mathcal{X}}$, $\|x' - x_1\| = \epsilon_1 \geq \epsilon_2$ and it is a contradiction. Hence, there is no x' that satisfies $x' \in R_{gen;exact}(\epsilon_1)$ and $x' \in R_{gen;exact}(\epsilon_2)$.

Prove for $(\epsilon_1 = 0, \epsilon_2 > 0) \vee (\epsilon_1 > 0, \epsilon_2 = 0)$ can be done similarly, and we finished the prove for the statement (4).

Part 2

Now, we prove that the regions of points will be used for different $x \in S_{exact}(\epsilon)$ have no intersection.

$$\text{Let } R_{gen;exact}(\epsilon; x) = \begin{cases} \{x' \in \mathbb{R} \mid \|x' - x\| = \epsilon\} \cap \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c, & \text{when } \epsilon > 0, \\ \{x\}, & \text{when } \epsilon = 0 \end{cases}$$

We need to prove the following.

$$x_1 \neq x_2 \implies R_{gen;exact}(\epsilon; x_1) \cap R_{gen;exact}(\epsilon; x_2) = \emptyset \quad (5)$$

This is obvious when $\epsilon = 0$ as $R_{gen;exact}(0; x_1) = \{x_1\}$ and $R_{gen;exact}(0; x_2) = \{x_2\}$.

We consider when $\epsilon > 0$.

Let $x' \in R_{gen;exact}(\epsilon; x_1), R_{gen;exact}(\epsilon; x_2)$ for $x_1 \neq x_2$. Then, $\|x' - x_1\| = \epsilon = \|x' - x_2\|$ and $x' \in VB(\mathcal{X})$. However, as $x' \in R_{gen;exact}(\epsilon; x_1)$, $x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c$. It is a contradiction that $x' \in VB(\mathcal{X})$ and $x' \in VB(\mathcal{X})^c$, and $R_{gen;exact}(\epsilon; x_1) \cap R_{gen;exact}(\epsilon; x_2)$ needs to be \emptyset . We finished the prove for the statement (5).

Because of statements 4 and 5, there will be no overlap when we choose different ϵ or different $x \in S_{exact}(\epsilon)$, and thus we proved Theorem 1. \square

D PROOF OF THEOREM 2 (1-NN CLASSIFIER IS THE CLASSIFIER THAT MAXIMIZES GENUINE ADVERSARIAL ACCURACY)

Proof. Part 1

First, we prove that a 1-NN classifier maximizes genuine adversarial accuracy. We denote the 1-NN classifier as f_{1-NN} .

When $\epsilon = 0$, $a_{gen;exact}(0) = \mathbb{E}_{x \in \mathcal{X}} [\mathbb{1}(f_{1-NN}(x) = c_x)] = \mathbb{E}_{x \in \mathcal{X}} [1] = 1$.

When $\epsilon > 0$, let $x \in S_{exact}(\epsilon)$.

Then, $\exists x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c : \|x' - x\| = \epsilon$.

$\forall x_{clean} \in \bar{\mathcal{X}} : \|x' - x_{clean}\| \geq \epsilon$ and $x' \in VB(\mathcal{X})^c$ (\because Equivalence relation (2) in Section B)

Because of that and $\|x' - x\| = \epsilon$, x and x' are nearest neighbors. Thus, $f_{1-NN}(x') = c_x$. As x^* is a special case for x' , $f_{1-NN}(x^*) = c_x$ and $\mathbb{1}(f_{1-NN}(x^*) = c_x) = 1$.

Hence, $a_{gen;exact}(\epsilon) = 1$ and f_{1-NN} maximizes genuine adversarial accuracy.

Part 2

Now, we prove that if f^* maximizes genuine adversarial accuracy, then f^* becomes a 1-NN classifier (almost everywhere) except for $\bar{\mathcal{X}} - \mathcal{X}$ and Voronoi boundary $VB(\mathcal{X})$. As we know that a 1-NN classifier maximizes genuine adversarial accuracy from part 1, we only need to show that f^* is

almost everywhere unique (except for $\bar{\mathcal{X}} - \mathcal{X}$ and Voronoi boundary $VB(\mathcal{X})$).

Let f^{*1} be a function that maximizes genuine adversarial accuracy.

When $\epsilon = 0$, $\mathbb{E}_{x \in \mathcal{X}} [\mathbb{1}(f_{1-NN}(x) = c_x)] = 1 = \mathbb{E}_{x \in \mathcal{X}} [\mathbb{1}(f^{*1}(x) = c_x)]$. We get $\mathbb{1}(f^{*1}(x) = c_x) = 1$ almost everywhere for $x \in \mathcal{X}$. It is equivalent to $f^{*1}(x) = c_x = f_{1-NN}(x)$ almost everywhere for $x \in \mathcal{X}$.

When $\epsilon > 0$, let $x \in S_{exact}(\epsilon)$.

With similar process (when $\epsilon = 0$), we get $f^{*1}(x^*) = c_x$ and $f_{1-NN}(x^{**}) = c_x$ almost everywhere for $x \in S_{exact}(\epsilon)$, x^* and x^{**} .

As x^* and x^{**} are worst case adversarially perturbed samples, i.e., samples that output mostly different from c_x , $f^{*1}(x') = c_x$ and $f_{1-NN}(x'') = c_x$ almost everywhere where $x', x'' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c$, $\|x' - x\| = \epsilon = \|x'' - x\|$.

We can consider when $x' = x''$ and we get $f^{*1}(x') = c_x = f_{1-NN}(x')$ almost everywhere where $x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c$, $\|x' - x\| = \epsilon$.

By changing ϵ , x' that satisfies $x' \in \mathcal{X}_\epsilon^c \cap VB(\mathcal{X})^c$ and $\|x' - x\| = \epsilon$ can fill up $\mathbb{R}^d - \bar{\mathcal{X}}$ except for $VB(\mathcal{X})$ (: Lemma 3 in Section B.). Hence, f^{*1} is almost everywhere same with f_{1-NN} except for $\bar{\mathcal{X}} - \mathcal{X}$ and Voronoi boundary $VB(\mathcal{X})$. \square

E GENUINE ADVERSARIAL ACCURACY BY MAXIMUM PERTURBATION NORM

Even though the advantages of genuine adversarial accuracy by exact perturbation norm, it can be hard to calculate it in practice. That is due to the complex calculation in the projected gradient descent (PGD) method (Madry et al., 2017) when non-path-connected regions are used. This problem can be solved when we use genuine adversarial accuracy by maximum perturbation norm because, for each $x \in \mathcal{X}$, it uses the intersection of $Ball(x, \epsilon)$ and $Vor(x) - VB(\mathcal{X})$ where $Ball(x, \epsilon) = \{x' \in \mathbb{R}^d | \|x - x'\| \leq \epsilon\}$ and Voronoi cell $Vor(x) = \{x' \in \mathbb{R}^d | \|x - x'\| \leq \|x_{clean} - x'\|, \forall x_{clean} \in \mathcal{X} - \{x\}\}$. The intersection is convex as both $Ball(x, \epsilon)$ and $Vor(x) - VB(\mathcal{X})$ are convex. Hence, it is path-connected. When applying projections for the PGD method, for each iteration, one needs to apply projection on $Ball(x, \epsilon)$ first, then apply projection using $Vor(x) - VB(\mathcal{X})$.

Definition 4 (Genuine adversarial accuracy by maximum perturbation norm). We define genuine adversarial accuracy that uses the maximum perturbation norm. Note that $\mathbb{1}()$ is an indicator function that has value 1 if the condition in the bracket holds and value 0 if the condition in the bracket does not hold. Voronoi boundary $VB(\mathcal{X})$ is defined as $\{x' \in \mathbb{R}^d | \exists x_1, x_2 \in \mathcal{X} : x_1 \neq x_2, \|x' - x_1\| = \|x' - x_2\|\}$. Then, genuine adversarial accuracy (by maximum perturbation norm) $a_{gen; max}(\epsilon)$ is defined as follows.

$$\begin{aligned} \bullet \quad a_{gen; max}(\epsilon) &= \mathbb{E}_{x \in \mathcal{X}} [\mathbb{1}(f(x^*) = c_x)] \\ \text{where } x^* &= \arg \max_{x' \in Vor(x) - VB(\mathcal{X}) : \|x' - x\| \leq \epsilon} L(\theta, x', c_x). \end{aligned}$$

Genuine adversarial accuracy by maximum perturbation norm does not satisfy Theorem 1 (It satisfies similar property with part 2 property in the proof in Section D.). But, it satisfies Theorem 2 (Proof is omitted, but can be done similarly with the proof of genuine adversarial accuracy by exact perturbation norm.). As it still satisfies Theorem 2, when measuring adversarial robustness of classifiers, it can replace the genuine adversarial accuracy by exact perturbation norm. Figure 7 shows changes of genuine adversarial accuracy by maximum perturbation norm for the three classifiers defined in Section 1.2.1. Decreasing genuine adversarial accuracy by maximum perturbation norm ϵ indicates that genuine adversarial accuracy by exact perturbation norm ϵ will be smaller than 1 (In general, the converse of this statement does not hold).

F GRADUAL NEAREST NEIGHBOR CLASSIFIERS

In this section, we introduce gradual nearest neighbor (gradual 1-NN) classifiers which can take non-discrete values in their prediction values. These classifiers can be used for adversarial training with Voronoi constraints (Khouri & Hadfield-Menell, 2019) as gradual 1-NN classifiers have the same decision boundaries with standard (discrete) single nearest neighbor (1-NN) classifiers. Using

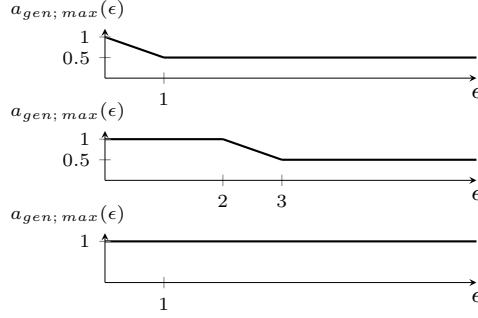


Figure 7: Change of genuine adversarial accuracy for $f_1(x)$, $f_2(x)$ and $f_3(x)$ by maximum perturbation norm ϵ from top to bottom. Details for calculation is omitted.

soft labels based on gradual 1-NN classifiers might help to mitigate the over-confidence of deep networks (Guo et al., 2017).

For each point $x' \in \mathbb{R}^d$, gradual 1-NN classifiers use the nearest distances to clean samples for each class. Let us denote the nearest distance to clean samples in class c as $d_{x';c}$. Then, for score of

class c , gradual 1-NN classifiers output $g_c(x') = \begin{cases} \frac{\frac{1}{d_{x';c}}}{\sum_{y \in \mathcal{Y}} \frac{1}{d_{x';y}}}, & \text{if } d_{x';c} \neq 0, \\ 1, & \text{if } d_{x';c} = 0 \end{cases}$. (It is not necessary to

use inverse value of $d_{x';c}$. One can use other decreasing non-negative functions whose right-hand limit at zero is infinity. For example, $\frac{1}{d_{x';c}^2}$ or $\frac{1}{\ln(1+d_{x';c})}$ can also be used instead.) Notice that $\sum_{y \in \mathcal{Y}} g_y(x') = 1$ for any $x' \in \mathbb{R}^d$, i.e., scores of gradual 1-NN classifiers are normalized, and the score of class c will approach maximum score 1 as $d_{x';c}$ approach zero.

The formula of gradual 1-NN classifiers is similar to that of neighbourhood components analysis (NCA) (Goldberger et al., 2005) and matching network (Vinyals et al., 2016). However, their formulas may not be proper relaxations of 1-NN classifiers as they do not output 1 when $d_{x';c} = 0$ for score of class c , and it is even possible that the score of class c is higher when $d_{x';c} \neq 0$ than $d_{x';c} = 0$. Also, the output for $d_{x';c} = 0$ may vary depending on the data points. (Like NCA and matching network, one can also use the formula of gradual 1-NN classifier for metric learning.)

The prediction values given by gradual 1-NN classifiers are equivalent with implicit scores of Nearest Neighbor Distance Ratio (NNDR)-based Open-Set Nearest Neighbor (OSNN) (Júnior et al., 2017) which was devised to handle open-set classification problems (Geng et al., 2020). One can consider the values of gradual 1-NN classifiers as a normalized form of OSNN for known classes. Gradual 1-NN classifier can solve open-set classification problems using a different method for estimating the probability of getting samples from unknown

classes. Let $g_{open-set}(x') = \begin{cases} (1 - g_{unknown}(x')) g_c(x'), & \text{if } x' \text{ is from a known class } c \in \mathcal{Y}, \\ g_{unknown}(x'), & \text{if } x' \text{ is from unknown classes} \end{cases}$ where $g_{unknown}(x') = -\alpha \sum_{y \in \mathcal{Y}} g_y(x') \log_{|\mathcal{Y}|} g_y(x')$ and α is a parameter that satisfies $0 \leq \alpha \leq 1$.

(In order for some samples to be classified as unknown classes, α needs to be larger than $\frac{1}{1+|\mathcal{Y}|}$.)

Then, $g_{unknown}(x')$ uses entropy to estimate the probability of getting samples from unknown classes. (If $g_{unknown}(x')$ is defined as $g_{unknown}(x') = \alpha \prod_{y \in \mathcal{Y}} g_y(x')^{\frac{1}{|\mathcal{Y}|}}$ instead, where α is a

parameter that satisfies $0 \leq \alpha \leq |\mathcal{Y}|$, then the cross entropy loss of $g_{open-set}(x')$ becomes similar to the Entropic Open-Set loss (Dhamija et al., 2018). Note that their loss is based on cross entropy instead of entropy. In order for some samples to be classified as unknown classes, α needs to be larger than $\frac{|\mathcal{Y}|}{1+|\mathcal{Y}|}$.)

G DETAILS OF THE MODELS AND FURTHER ANALYSIS RESULTS ON MNIST AND CIFAR-10 DATA

Pretrained non-adversarially trained model and PGD-AT model for MNIST data (LeCun et al., 2010) were from https://github.com/MadryLab/mnist_challenge. Pretrained non-adversarially trained model and PGD-AT model for CIFAR-10 data (Krizhevsky, 2009) were from (Engstrom et al., 2019). Pretrained TRADES models were from <https://github.com/yaodongyu/TRADES>. Non-adversarially trained models compared with TRADES models were obtained by non-adversarially training models with the default settings.

Let \mathcal{X}_{train} be the training input dataset. For each sample $x \in \mathcal{X}_{train}$, the smallest distance to $x_{clean} \in \mathcal{X}_{train}$ with different class is denoted as $d_{x,diff}$. Note that standard adversarial training (Goodfellow et al., 2014) with ϵ smaller than half of minimum $d_{x,diff}$ satisfy the condition of properly applied adversarial training. Adversarially trained models used in the analyses satisfy properly applied adversarial training based on Table 4.

While many number of prediction changes were changed to same predictions with 1-NN classifiers (see tables 7 to 13), there were many converse changes. This could be due to the limited capacities of convolutional neural networks that encourage translation invariance.

Table 4: Minimum of $d_{x,diff}$ values of MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky, 2009) training data.

Data	MNIST	CIFAR-10
l_2 norm	2.399	2.7501
l_∞ norm	0.7569 ($= \frac{193}{255}$)	0.2118 ($= \frac{54}{255}$)

Table 5: Proportions of agreements with 1-NN classifiers on whole predictions on test samples with non-adversarially trained models and adversarially trained models (Engstrom et al., 2019; Zhang et al., 2019) for MNIST (LeCun et al., 2010) training data.

Model (Distance metric, training ϵ)	PGD-AT (l_∞ norm, $\epsilon = 0.3$)	TRADES (l_∞ norm, $\epsilon = 0.3$)
Non-adversarially trained model	0.6144 (0.8431)	0.6143 (0.8422)
Adversarially trained model	0.6130 (0.8422)	0.6147 (0.8425)

Table 6: Proportions of agreements with 1-NN classifiers on whole predictions on test samples with non-adversarially trained models and adversarially trained models (Engstrom et al., 2019; Zhang et al., 2019) for CIFAR-10 (Krizhevsky, 2009) training data.

Model (Distance metric, training ϵ)	PGD-AT (l_2 norm, $\epsilon = 0.25$)	PGD-AT (l_2 norm, $\epsilon = 0.5$)	PGD-AT (l_2 norm, $\epsilon = 1.0$)	PGD-AT (l_∞ norm, $\epsilon = 0.03137$ $= \frac{8}{255}$)	TRADES (l_∞ norm, $\epsilon = 0.031$)
Non-adversarially trained model	0.3515	0.3515	0.3515	0.1738 (0.1787)	0.1717 (0.1764)
Adversarially trained model	0.3590	0.3632	0.3691	0.1771 (0.1821)	0.1851 (0.1907)

Table 7: Division of the change of different predictions on test samples from non-adversarially trained model to PGD-AT model (l_∞ norm, $\epsilon = 0.3$) (Engstrom et al., 2019) for MNIST (LeCun et al., 2010)

PGD-AT (l_∞ norm, $\epsilon = 0.3$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	10 (28)	28 (36)	75 (49)	113
Decreased error	5 (9)	2 (13)	29 (14)	36
Errors on both models	1 (3)	0 (0)	4 (2)	5
Total	16 (40)	30 (49)	108 (65)	154

Table 8: Division of the change of different predictions on test samples from non-adversarially trained model to TRADES model (l_∞ norm, $\epsilon = 0.3$) (Engstrom et al., 2019) for MNIST (LeCun et al., 2010)

TRADES (l_∞ norm, $\epsilon = 0.3$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	7 (9)	4 (7)	17 (12)	28
Decreased error	4 (6)	3 (6)	18 (13)	25
Errors on both models	0 (1)	0 (0)	1 (0)	1
Total	11 (16)	7 (13)	36 (25)	54

Table 9: Division of the change of different predictions on test samples from non-adversarially trained model to PGD-AT model (l_2 norm, $\epsilon = 0.25$) (Engstrom et al., 2019) for CIFAR-10 (Krizhevsky, 2009)

PGD-AT (l_2 norm, $\epsilon = 0.25$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	111	75	267	453
Decreased error	55	25	125	205
Errors on both models	20	11	37	68
Total	186	111	429	726

Table 10: Division of the change of different predictions on test samples from non-adversarially trained model to PGD-AT model (l_2 norm, $\epsilon = 0.5$) (Engstrom et al., 2019) for CIFAR-10 (Krizhevsky, 2009)

PGD-AT (l_2 norm, $\epsilon = 0.5$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	178	115	365	658
Decreased error	67	29	120	216
Errors on both models	26	10	53	89
Total	271	154	538	963

Table 11: Division of the change of different predictions on test samples from non-adversarially trained model to PGD-AT model (l_2 norm, $\epsilon = 1.0$) (Engstrom et al., 2019) for CIFAR-10 (Krizhevsky, 2009)

PGD-AT (l_2 norm, $\epsilon = 1.0$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	434	332	763	1529
Decreased error	59	21	86	166
Errors on both models	52	16	86	154
Total	545	369	935	1849

Table 12: Division of the change of different predictions on test samples from non-adversarially trained model to PGD-AT model (l_∞ norm, $\epsilon = \frac{8}{255}$) (Engstrom et al., 2019) for CIFAR-10 (Krizhevsky, 2009)

PGD-AT (l_∞ norm, $\epsilon = \frac{8}{255}$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	169 (171)	120 (123)	726 (721)	1015
Decreased error	20 (23)	21 (21)	152 (149)	193
Errors on both models	7 (7)	22 (23)	87 (86)	116
Total	196 (201)	163 (167)	965 (956)	1324

Table 13: Division of the change of different predictions on test samples from non-adversarially trained model to TRADES (l_∞ norm, $\epsilon = 0.031$) (Engstrom et al., 2019) for CIFAR-10 (Krizhevsky, 2009)

PGD-AT (l_∞ norm, $\epsilon = 0.031$)	Non 1-NN to 1-NN	1-NN to non 1-NN	Non 1-NN for both models	Total
Increased error	233 (239)	114 (117)	892 (883)	1239
Decreased error	29 (33)	18 (19)	146 (141)	193
Errors on both models	18 (21)	14 (14)	96 (93)	128
Total	280 (293)	146 (150)	1134 (1117)	1560

H SPECULATED OPTIMALLY ROBUST CLASSIFIERS WHEN DATA CONTAIN INPUT NOISE

In our analysis, we assumed the exclusive class assumption in the problem setting in order to simplify the analysis. This section describes how to get speculated optimally robust classifiers when certain input noises were added. Notice that such classifiers can also be used for adversarial training with Voronoi constraints (Khouri & Hadfield-Menell, 2019) by replacing exclusive labels with soft labels (as optimally robust classifiers would output probabilities). When input noises were added, we can represent that as $x = x_{\text{no-noise}} + n_x$ where $x_{\text{no-noise}}$ is an original point of the sample x before additive noise n_x was added.

Let us consider the case when noise n_x follows Gaussian distribution with zero mean and scalar covariance matrix $\sigma^2 I$ for a fixed $\sigma \geq 0$. As $x_{\text{no-noise}} = x - n_x$ and Gaussian distribution with zero mean is symmetric with respect to the zero, we know the distribution of the estimated position of $x_{\text{no-noise}}$. That is Gaussian distribution with mean x and covariance matrix $\sigma^2 I$. Based on the estimated position of $x_{\text{no-noise}}$, we can generate sets of estimated points of $x_{\text{no-noise}}$ for each x , and there will be a corresponding 1-NN classifier for each set. If we take the average (ensemble) on

these 1-NN classifiers, we get the speculated optimally robust classifier when Gaussian input noises were added.

Figure 8 shows a two-dimensional example with different metrics and varying σ (Figure 9 shows corresponding results combined with gradual 1-NN classifiers explained in Section F). Notice that we will get different decision boundaries depending on the metrics and noise even though we are using the same data. As σ gets large, the decision boundaries become more smooth, and they can allow misclassifications of some data samples.

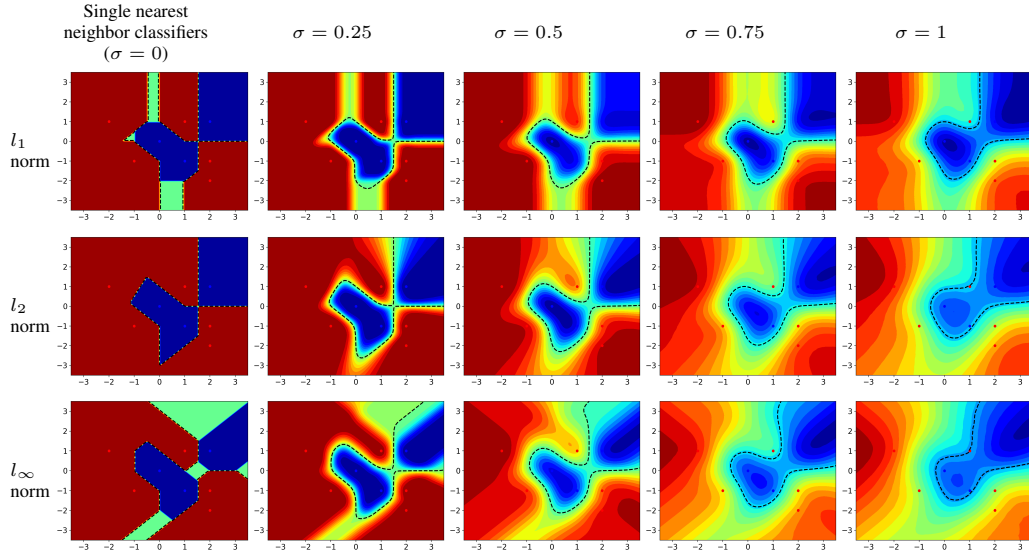


Figure 8: Contour plots of speculated optimally robust classifiers for a two-dimensional example when points $(1, 1)$, $(2, -1)$, $(2, -2)$, $(-1, -1)$ and $(-2, 1)$ were provided for **class A: red** and points $(2, 1)$, $(1, -1)$, and $(0, 0)$ were provided for **class B: blue**. Dashed black curves show decision boundaries for different cases. Figures on the left show single nearest neighbor classifiers for l_1 , l_2 , and l_∞ norms. Ensemble of 50000 1-NN classifiers were used for each ensemble classifier. For noise added cases of l_1 and l_∞ norms, we used normalized radial basis function (Moody & Darken, 1989) with Gaussian radial kernels as noise distributions. Uniform prior probability is assumed for all cases.

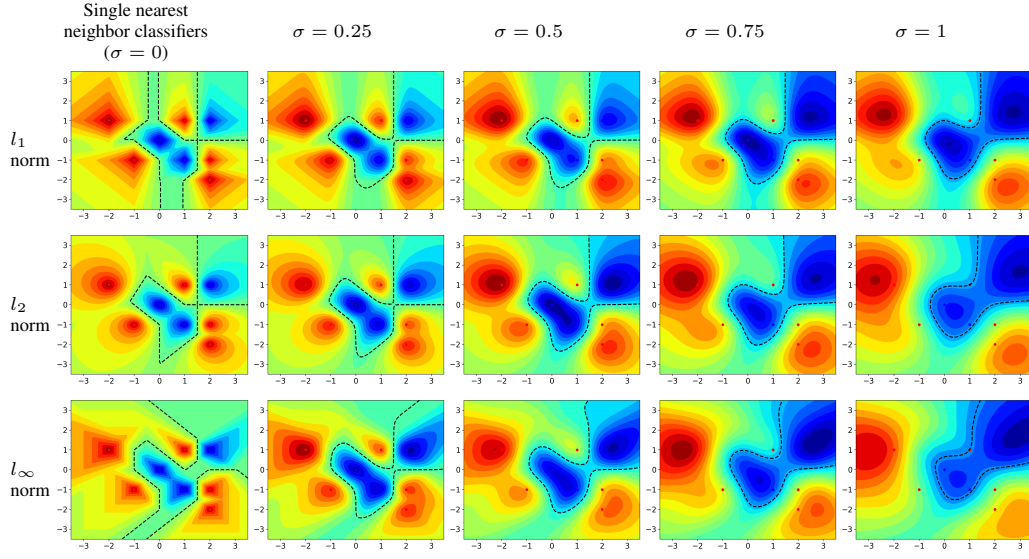


Figure 9: Contour plots of speculated optimally robust classifiers combined with gradual 1-NN classifiers (explained in Section F) for the same example in Figure 8. Dashed black curves show decision boundaries for different cases. Figures on the left show single nearest neighbor classifiers for l_1 , l_2 , and l_∞ norms. Ensemble of 50000 1-NN classifiers were used for each ensemble classifier. For noise added cases of l_1 and l_∞ norms, we used normalized radial basis function (Moody & Darken, 1989) with Gaussian radial kernels as noise distributions. Uniform prior probability is assumed for all cases.