Measuring Adversarial Robustness using a Voronoi-Epsilon Adversary

Hyeongji Kim*[1,2], Pekka Parviainen[1], and Ketil Malde[1,2]

[1]Department of Informatics, University of Bergen, Norway
[2]Institute of Marine Research, Bergen, Norway

Abstract

Previous studies on robustness have argued that there is a tradeoff between accuracy and adversarial accuracy. The tradeoff can be inevitable even when we neglect generalization. We argue that the tradeoff is inherent to the commonly used definition of adversarial accuracy, which uses an adversary that can construct adversarial points constrained by ϵ-balls around data points. As ϵ gets large, the adversary may use real data points from other classes as adversarial examples. We propose a Voronoi-epsilon adversary which is constrained both by Voronoi cells and by ϵ-balls. This adversary balances two notions of perturbation. As a result, adversarial accuracy based on this adversary avoids a tradeoff between accuracy and adversarial accuracy on training data even when ϵ is large. Finally, we show that a nearest neighbor classifier is the maximally robust classifier against the proposed adversary on the training data.

1 Introduction

By applying a carefully crafted, but imperceptible perturbation to input images, so-called adversarial examples can be constructed that cause classifiers to misclassify the perturbed inputs [Szegedy et al., 2014]. Defense methods like adversarial training [Madry et al., 2018] and certified defenses [Wong and Kolter, 2018] against adversarial examples have often resulted in decreased accuracies on clean samples [Tsipras et al., 2019]. Previous studies have argued that the tradeoff between accuracy and adversarial accuracy may be inevitable in classifiers [Tsipras et al., 2019, Dohmatob, 2019, Zhang et al., 2019].

1.1 Problem Settings

Problem setting. *Let $\mathcal{X} \subset \mathbb{R}^{\text{dim}}$ be a nonempty input space and $\mathcal{Y}$ be a set of possible classes. Data points $x \in \mathcal{X}$ and corresponding classes $c_x \in \mathcal{Y}$ are sampled from a joint distribution $\mathcal{D}$. The distribution $\mathcal{D}$ should satisfy the condition that c_x is unique for all x. The set of the data points is a finite, nonempty set X. A classifier f assigns a class label from $\mathcal{Y}$ for each point $x \in \mathcal{X}$. $l(y_1, y_2)$ is a classification loss function for $y_1, y_2 \in \mathcal{Y}$ and it satisfies the necessary condition:*

$$\forall y_1, y_2, y_3, y_4 \in \mathcal{Y},$$
$$l(y_1, y_2) \leq l(y_3, y_4) \Longrightarrow \mathbb{1}(y_1 = y_2) \geq \mathbb{1}(y_3 = y_4).$$

$L(x, y)$ is a classification loss based on the classifier f provided an input $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$. Mathematically, $L(x, y) := l(f(x), y)$.

To simplify the analysis, we do not consider generalization.

1.2 Adversarial Accuracy (AA)

For a classifier, (natural) accuracy a is the expectation of a correct classification of data sampled from the data distribution. Mathematically, it is defined as:

$$a = \mathbb{E}_{(x,c_x)\sim\mathcal{D}} \left[\mathbb{1}\left(f(x) = c_x\right)\right].$$

Adversarial accuracy is a commonly used measure of adversarial robustness of classifiers [Madry et al., 2018, Tsipras et al., 2019]. It is defined by an adversary region $R(x) \subset \mathcal{X}$, which is an allowed region of the perturbations for a data point x.

[*] Corresponding Author: hjk92g@gmail.com

https://doi.org/10.7557/18.6827

Definition 1 (**Adversarial accuracy**). *Given an adversary that is constrained to an adversary region $R(x)$, adversarial accuracy a is defined as:*

$$a = \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f(x^*) = c_x\right)\right],$$

$$\text{where } x^* = \underset{x' \in R(x)}{\arg\max}\, L(x', c_x).$$

The choice of $R(x)$ will determine the adversarial accuracy that we are measuring. Commonly considered adversary region is $\mathbb{B}(x, \epsilon)$, which is a ϵ-ball around a data point x based on a distance metric d [Biggio et al., 2013, Madry et al., 2018, Tsipras et al., 2019, Zhang et al., 2019].

Definition 2 (**Standard adversarial accuracy**). *When the adversary region is $\mathbb{B}(x, \epsilon)$, we refer to the adversarial accuracy a as standard adversarial accuracy (SAA) $a_{std}(\epsilon)$. For SAA, we denote $R(x)$ as $R_{std}(\epsilon; x)$.*

$$a_{std}(\epsilon) = \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f(x^*) = c_x\right)\right],$$

$$\text{where } x^* = \underset{x' \in R_{std}(\epsilon;x)}{\arg\max}\, L(x', c_x).$$

This adversary region $\mathbb{B}(x, \epsilon)$ is based on an implicit assumption that there might be an adequate single epsilon ϵ that perturbed samples do not change their classes. However, this assumption has some limitations. We explain that in the next section.

1.3 The Tradeoff Between Accuracy and Standard Adversarial Accuracy

The usage of ϵ-ball-based adversary can cause the tradeoff between accuracy and adversarial accuracy. When the two clean samples x_1 and x_2 with $d(x_1, x_2) \leq \epsilon$ have different classes, achieving local SAA higher than 0 on these two points implies misclassification. We illustrate this with a toy example.

1.3.1 Toy Example

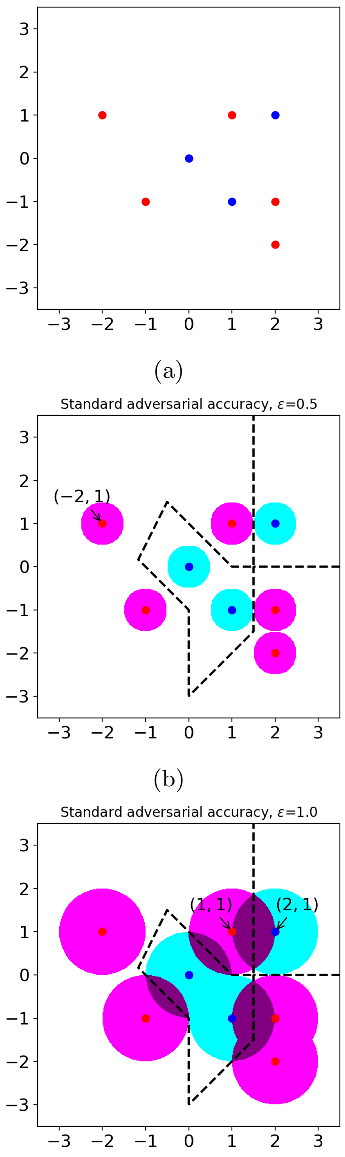
Let us consider an example visualized in Figure 1a. The input space is $\mathbb{R}^2$. There are only two classes A and B, i.e., $\mathcal{Y} = \{A, B\}$. We use the l_2 norm as a distance metric in this example.

Let us consider a situation when $\epsilon = 1.0$ (see Figure 1c). In this case, clean samples can also be considered as adversarial examples. For example, the point $(2, 1)$ can be considered as an adversarial example originating from the point $(1, 1)$. If one choose a robust model based on SAA, one might choose a model with excessive invariance. For example, one might choose a model that predicts points belong to $\mathbb{B}((1, 1), 1)$ (including the point $(2, 1)$) have class A. Or, one can choose a model that predicts points belong to $\mathbb{B}((2, 1), 1)$ (including the point $(1, 1)$) have class B. In either case, the accuracy of the chosen model is smaller than 1. This situation explains the tradeoff between accuracy and standard adversarial accuracy when large ϵ is used. It originates from the overlapping adversary regions from the samples with different classes.

To avoid the tradeoff between accuracy and adversarial accuracy, one can use small ϵ values. Actually, a previous study has argued that commonly used ϵ values are small enough to avoid the tradeoff [Yang et al., 2020b]. However, when small ϵ values are used, we can only analyze local robustness, and we need to ignore robustness beyond the chosen ϵ. For instance, let us consider our example when $\epsilon = 0.5$ (see Figure 1b). In this case, we ignore robustness on $\mathbb{B}((-2, 1), 1.0) - \mathbb{B}((-2, 1), 0.5)$. Models with local but without global robustness enable attackers to use large ϵ values to fool the models. Ghiasi et al. [2019] have experimentally shown that even models with certified local robustness can be attacked by attacks with large ϵ values. Note that their attack applies little semantic perturbations even though the perturbation norms measured by l_p norms are large.

These limitations motivate us to find an alternative way to measure robustness. **The contributions of this paper are as follows.**

- We propose Voronoi-epsilon adversarial accuracy (VAA) that avoids the tradeoff between accuracy and adversarial accuracy. This allows the adversary regions to scale to cover most of the input space without incurring a tradeoff. To our best knowledge, this is the first work to achieve this without an external classifier.
- We explain the connection between SAA and VAA. We define global Voronoi-epsilon robustness as a limit of the Voronoi-epsilon adversarial accuracy. We show that a nearest neighbor (1-NN) classifier maximizes global Voronoi-epsilon robustness.



(a)

(b)

(c)

Figure 1: (a): Plot of the two-dimensional toy example. Data points are colored based on their classes (class A: red and class B: blue). (b): Visualization of the adversary regions for SAA when $\epsilon = 0.5$. The regions are colored differently depending on their classes (class A: magenta and class B: cyan). The decision boundary of a single nearest neighbor classifier is shown as a dashed black curve. (c): Visualization of the adversary regions for SAA when $\epsilon = 1.0$. The overlapping adversary regions from the samples with different classes are colored in purple.

2 Voronoi-Epsilon Adversarial Accuracy (VAA)

Our approach restricts the allowed region of the perturbations to avoid the tradeoff originating from the definition of standard adversarial accuracy. This is achieved without limiting the magnitude of ϵ and without using an external model. We want to have the following property to avoid the tradeoff.

$$\forall x_i, x_j \in X,\ x_i \neq x_j \implies R(x_i) \cap R(x_j) = \varnothing \quad (1)$$

When Property (1) holds for the adversary region, we no longer have the tradeoff as $x_i \notin R(x_j)$ for $x_i \neq x_j$. In other words, a clean sample cannot be an adversarial example originating from another clean sample. We propose a new adversary called a Voronoi-epsilon adversary that combines the Voronoi-adversary introduced by Khoury and Hadfield-Menell [2019] with an ϵ-ball-based adversary. This adversary is constrained to an adversary region $Vor(x) \cap \mathbb{B}(x, \epsilon)$ where $Vor(x)$ is the (open) Voronoi cell around a data point $x \in X$. $Vor(x)$ consists of every point in $\mathcal{X}$ that is closer than any $x_{clean} \in X - \{x\}$. Mathematically, $Vor(x) = \{x' \in \mathcal{X} | d(x, x') < d(x_{clean}, x'), \forall x_{clean} \in X - \{x\}\}$. Then, Property (1) holds as $Vor(x_i) \cap Vor(x_j) = \varnothing$ for $x_i \neq x_j$.

Based on a Voronoi-epsilon adversary, we define Voronoi-epsilon adversarial accuracy (VAA).

Definition 3 (**Voronoi-epsilon adversarial accuracy**). *When a Voronoi-epsilon adversary is used for the adversary, we refer to the adversarial accuracy as Voronoi-epsilon adversarial accuracy (VAA) $a_{Vor}(\epsilon)$. For VAA, we denote $R(x)$ as $R_{Vor}(\epsilon; x)$, i.e., $R_{Vor}(\epsilon; x) = Vor(x) \cap \mathbb{B}(x, \epsilon)$.*

$$a_{Vor}(\epsilon) = \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f(x^*) = c_x\right)\right]$$ [1]

where $x^* = \underset{x' \in R_{Vor}(\epsilon;x)}{\arg\max}\ L(x', c_x)$.

Figure 2 shows the adversary regions for VAA with varying ϵ values. When $\epsilon = 0.5$, the regions are same with SAA except for the points $(1.5, 1), (1.5, -1)$ and $(2, -1.5)$. Even when ϵ is large ($\epsilon > 0.5$), there is no overlapping adversary

[1] Using the expectation here is a slight abuse of notation, since $a_{Vor}(\cdot)$ is defined on a finite set. We retain it for consistency with previous definitions, and understand it to mean the empirical average.

region, which was a source of the tradeoff in SAA. Therefore, when we choose a robust model based on VAA, we can get a model that is both accurate and robust. Figure 2c shows the single nearest neighbor (1-NN) classifier would maximize VAA. The adversary regions cover most of the points in $\mathbb{R}^2$ for large ϵ.

Observation 1. *Let d_{min} be the nearest distance of the data point pairs, i.e., $d_{min} = \min_{x_i,x_j \in X, x_i \neq x_j} d(x_i, x_j)$. Then, the following equivalence holds.*

$$a_{Vor}(\epsilon) = a_{std}(\epsilon), \tag{2}$$

when $\epsilon < \frac{1}{2} d_{min}$.

Observation 1 shows that VAA is equivalent to SAA for sufficiently small ϵ values. This indicates that VAA is an extension of SAA that avoids the tradeoff when ϵ is large. The proof of the observation is in Appendix A.1. We point out that equivalent findings were also mentioned in Yang et al. [2020a,b], Khoury and Hadfield-Menell [2019].

As explained in Section 1.3.1, studying the local robustness of classifiers has a limitation. Attackers can attack models with only local robustness by using large ϵ values. The absence of a tradeoff between accuracy and VAA enables us to increase ϵ values and to study global robustness. We define a measure for global robustness using VAA.

Definition 4 (Global Voronoi-epsilon robustness). *Global Voronoi-epsilon robustness a_{global} is defined as:*

$$a_{global} = \lim_{\epsilon \to \infty} a_{Vor}(\epsilon).$$

Global Voronoi-epsilon robustness considers the robustness of classifiers for most points in $\mathcal{X}$ (all points except for Voronoi boundary $VB(X)$, which is the complement set of the unions of Voronoi cells.). We derive the following theorem from global Voronoi-epsilon robustness.

Theorem 1. *A single nearest neighbor (1-NN) classifier maximizes global Voronoi-epsilon robustness a_{global} on training data. 1-NN classifier is a unique classifier that satisfies this except for Voronoi boundary $VB(X)$.*

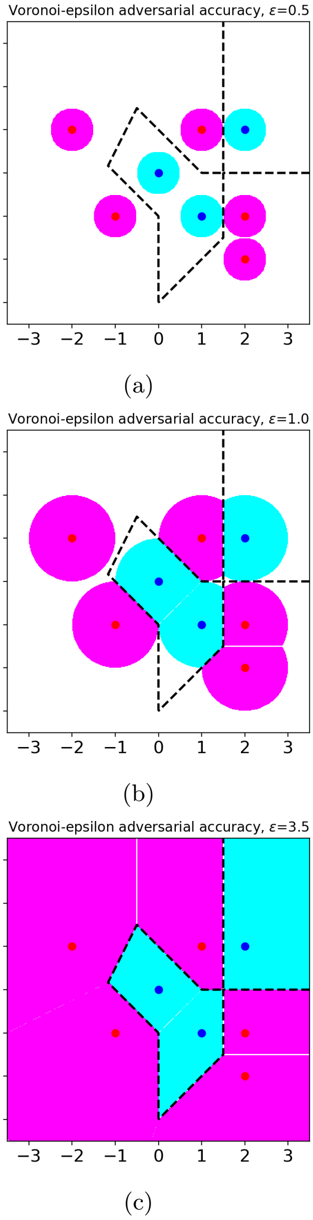


Figure 2: Visualization of the adversary regions for VAA with varying ϵ values. The data points and regions are colored as in Figure 1. (a): When $\epsilon = 0.5$. (b): When $\epsilon = 1.0$. (c): When $\epsilon = 3.5$.

When l_p norm with $1 < p < \infty$ is used as a distance metric, a data point will almost never lie on the Voronoi boundary $VB(X)$ in practical situations with only finite number of available training data points. Note that Theorem 1 only holds for exactly the same data under the exclusive class condition as mentioned in the problem settings 1.1. It does not take into account generalization. The proof of the theorem is in A.2.

3 Discussion

In this work, we address the tradeoff between accuracy and adversarial robustness by introducing the Voronoi-epsilon adversary. Another way to address this tradeoff is to use a Bayes optimal classifier [Suggala et al., 2019, Kim and Wang, 2020]. Since this is not available in practice, a reference model must be used as an approximation. In that case, the meaning of adversarial robustness is dependent on the choice of the reference model. VAA removes the need for a reference model by using the data point set X and the distance metric d to construct adversary. This is in contrast to Khoury and Hadfield-Menell [2019] who used Voronoi cell-based constraints (without ϵ-balls) for an adversarial training purpose, but not for measuring adversarial robustness.

By avoiding the tradeoff with VAA, we can extend the study of local robustness to global robustness. Also, Theorem 1 implies that VAA is a measure of agreement with the 1-NN classifier. For sufficiently small ϵ values, SAA is also a measure of agreement with the 1-NN classifier because SAA is equivalent to VAA as in Observation 1. This implies that many adversarial defenses [Goodfellow et al., 2015, Madry et al., 2018, Zhang et al., 2019, Wong and Kolter, 2018, Cohen et al., 2019] with small ϵ values unknowingly try to make locally the same predictions with a 1-NN classifier.

In our analysis, we do not take into account generalization, and robust models are known to often generalize poorly [Raghunathan et al., 2020]. Many defense models use softmax classifiers and the final classifications of softmax classifiers are done on the trained feature representations. The close relationship between adversarially robust models and the 1-NN classifier revealed by Observation 1 and Theorem 1 indicates that feature representations are affected by the distance relationship in the input space. It will be worth exploring if that can explain the reduced discriminative power [Wu et al., 2021] of robust models and their decreased generalization power.

Acknowledgments

We thank Dr. Nils Olav Handegard, Dr. Yi Liu, and Jungeum Kim for the helpful feedback. We also also thank Dr. Wieland Brendel for the helpful discussions.

References

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013. doi: https://doi.org/10.1007/978-3-642-40994-3_25.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. URL `https://proceedings.mlr.press/v97/cohen19c.html`.

E. Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pages 1646–1654. PMLR, 2019. URL `https://proceedings.mlr.press/v97/dohmatob19a.html`.

A. Ghiasi, A. Shafahi, and T. Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In *International Conference on Learning Representations*, 2019. URL `https://iclr.cc/virtual_2020/poster_HJxdTxHYvB.html`.

I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. doi: https://doi.org/10.48550/arXiv.1412.6572.

M. Khoury and D. Hadfield-Menell. Adversarial training with Voronoi constraints. *arXiv preprint* *arXiv:1905.01019*, 2019. doi: https://doi.org/10.48550/arXiv.1905.01019.

J. Kim and X. Wang. Sensible adversarial learning, 2020. URL https://openreview.net/forum?id=rJlf_RVKwr.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning*, pages 7909–7919. PMLR, 2020. URL https://proceedings.mlr.press/v119/raghunathan20a.html.

A. S. Suggala, A. Prasad, V. Nagarajan, and P. Ravikumar. Revisiting adversarial risk. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2331–2339. PMLR, 2019. URL http://proceedings.mlr.press/v89/suggala19a.html.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. 2014. doi: https://doi.org/10.48550/arXiv.1312.6199. 2nd International Conference on Learning Representations, ICLR 2014; Conference date: 14-04-2014 Through 16-04-2014.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018. URL http://proceedings.mlr.press/v80/wong18a.html?ref=https://githubhelp.com.

Z. Wu, H. Gao, S. Zhang, and Y. Gao. Understanding the robustness-accuracy tradeoff by rethinking robust fairness. 2021. URL https://openreview.net/forum?id=bl9zYxOVwa.

Y.-Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pages 941–951. PMLR, 2020a. URL http://proceedings.mlr.press/v108/yang20b.html.

Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/61d77652c97ef636343742fc3dcf3ba9-Abstract.html.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled tradeoff between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

A Appendix

A.1 Proof of Observation 1

To prove Observation 1, we introduce the following lemma.

Lemma 1. *When N is the number of data points, let $x_2, \cdots, x_N \in X - \{x\}$ be the sorted neighbors of a data point $x \in X$. Mathematically, $d(x, x_2) \leq d(x, x_3) \leq \cdots \leq d(x, x_N)$. Then, when $\epsilon < \frac{1}{2}d(x, x_2)$, the following equation holds.*

$$R_{Vor}(\epsilon; x) = \mathbb{B}(x, \epsilon). \tag{3}$$

Proof. **Lemma 1**
We only consider when $\epsilon < \frac{1}{2}d(x, x_2)$.
Let $x' \in \mathbb{B}(x, \epsilon)$. Then, $d(x, x') \leq \epsilon$.
$\frac{1}{2}d(x, x_2) \leq \frac{1}{2}d(x, x_{clean}), \forall x_{clean} \in X - \{x\}$.
Due to the triangle inequality, $\frac{1}{2}d(x, x_{clean}) \leq \frac{1}{2}d(x, x') + \frac{1}{2}d(x', x_{clean})$.
When we combine the above inequalities, $d(x, x') \leq \epsilon < \frac{1}{2}d(x, x_2) \leq \frac{1}{2}d(x, x_{clean}) \leq \frac{1}{2}d(x, x') + \frac{1}{2}d(x', x_{clean}), \forall x_{clean} \in X - \{x\}$.
Then, $\frac{1}{2}d(x, x') < \frac{1}{2}d(x', x_{clean}) = \frac{1}{2}d(x_{clean}, x'), \forall x_{clean} \in X - \{x\}$. Thus, $x' \in Vor(x)$.
Hence, $\mathbb{B}(x, \epsilon) \subset Vor(x)$ and $R_{Vor}(\epsilon; x) = \mathbb{B}(x, \epsilon) \cap Vor(x) = \mathbb{B}(x, \epsilon)$.
□

Now, we prove Observation 1.

Proof. **Observation 1**
$d_{min} \leq d(x, x_i), \forall x, x_i \in X, x \neq x_i$.
When $\epsilon < \frac{1}{2}d_{min}$, $\epsilon < \frac{1}{2}d_{min} \leq \frac{1}{2}d(x, x_i)$, $\forall x, x_i \in X, x \neq x_i$. Thus, $R_{Vor}(\epsilon; x) = \mathbb{B}(x, \epsilon)$, $\forall x \in X$ due to the equation (3) in Lemma 1.
Then, $a_{Vor}(\epsilon)$ is same with $a_{std}(\epsilon)$ as $R_{Vor}(\epsilon; x) = \mathbb{B}(x, \epsilon) = R_{std}(\epsilon; x), \forall x \in X$.
□

A.2 Proof of Theorem 1

To prove Theorem 1, we introduce the following lemma.

Lemma 2. *By changing ϵ and $x \in X$, x' that satisfies $x' \in R_{Vor}(\epsilon; x)$ can fill up $\mathcal{X}$ except for Voronoi boundary $VB(X)$. In other words, $VB(X)^{\mathsf{c}} = \mathcal{X} - VB(X) \subset \bigcup_{\epsilon \geq 0} \left(\bigcup_{x \in X} R_{Vor}(\epsilon; x) \right)$.*

Proof. **Lemma 2**
Let $x' \in VB(X)^{\mathsf{c}}$.
Note that mathematically, $VB(X) = \left(\bigcup_{x \in X} Vor(x) \right)^{\mathsf{c}}$.
Hence, $VB(X)^{\mathsf{c}} = \left(\left(\bigcup_{x \in X} Vor(x) \right)^{\mathsf{c}} \right)^{\mathsf{c}} = \bigcup_{x \in X} Vor(x)$.
$\exists x \in X$ such that $x' \in Vor(x)$.
Let $\epsilon^* = d(x, x')$. Then, $d(x, x') \leq \epsilon^*$ and $x' \in Vor(x)$.
$x' \in \mathbb{B}(x, \epsilon^*) \cap Vor(x) = R_{Vor}(\epsilon^*; x) \subset \bigcup_{\epsilon \geq 0} \left(\bigcup_{x \in X} R_{Vor}(\epsilon; x) \right)$.
We proved $VB(X)^{\mathsf{c}} \subset \bigcup_{\epsilon \geq 0} \left(\bigcup_{x \in X} R_{Vor}(\epsilon; x) \right)$. □

Now, we prove Theorem 1.

Proof. **Part 1**
First, we prove that a 1-NN classifier maximizes global Voronoi-epsilon robustness. We denote the 1-NN classifier as f_{1-NN} and calculate its global Voronoi-epsilon robustness.
For a data point $x \in X$, let $x' \in R_{Vor}(\epsilon; x) = \mathbb{B}(x, \epsilon) \cap Vor(x)$.
$x' \in Vor(x) \iff d(x, x') < d(x_{clean}, x'), \forall x_{clean} \in X - \{x\}$.
As $x' \in R_{Vor}(\epsilon; x) \subset Vor(x)$, x is unique nearest data point in X and thus $f_{1-NN}(x') = c_x$.
When $x^* = \underset{x' \in R_{Vor}(\epsilon; x)}{\arg\max} \; L(x', c_x)$,
$a_{Vor}(\epsilon) = \mathbb{E}_{(x, c_x) \sim \mathcal{D}} \left[\mathbb{1} \left(f_{1-NN}(x^*) = c_x \right) \right] = \mathbb{E}_{(x, c_x) \sim \mathcal{D}} [1] = 1$.
$a_{global} = \lim_{\epsilon \to \infty} a_{Vor}(\epsilon) = \lim_{\epsilon \to \infty} 1 = 1$. Thus, f_{1-NN} takes the maximum global Voronoi-epsilon robustness 1.

Part 2
Now, we prove that if f^* maximizes global Voronoi-epsilon robustness, then f^* becomes the 1-NN classifier except for Voronoi boundary $VB(X)$.
Let f^{*1} be a function that maximizes global Voronoi-epsilon robustness.
From the last part of the part 1, when we calculate global Voronoi-epsilon robustness of f^{*1}, it should satisfy the equation $a_{global} = 1$.
For a data point $x \in X$ and $\epsilon_1 < \epsilon_2$, $R_{Vor}(\epsilon_1; x) = \mathbb{B}(x, \epsilon_1) \cap Vor(x) \subset \mathbb{B}(x, \epsilon_2) \cap Vor(x) = R_{Vor}(\epsilon_2; x)$.
Thus, for a data point $x \in X$ and $\epsilon_1 < \epsilon_2$,

$L(x^{*1}, c_x) \leq L(x^{*2}, c_x)$ where $x^{*1} = \underset{x' \in R_{Vor}(\epsilon_1;x)}{\arg\max} L(x', c_x)$ and $x^{*2} = \underset{x' \in R_{Vor}(\epsilon_2;x)}{\arg\max} L(x', c_x)$. From the definition of L, $l(f^{*1}(x^{*1}), c_x) \leq l(f^{*1}(x^{*2}), c_x)$. From the necessary condition of classification loss l, we obtain the inequality $\mathbb{1}\left(f^{*1}(x^{*1}) = c_x\right) \geq \mathbb{1}\left(f^{*1}(x^{*2}) = c_x\right)$.

$a_{Vor}(\epsilon_1) = \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f^{*1}(x^{*1}) = c_x\right)\right] \geq \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f^{*1}(x^{*2}) = c_x\right)\right] = a_{Vor}(\epsilon_2)$ for $\epsilon_1 < \epsilon_2$. In other words, $a_{Vor}(\epsilon)$ is a monotonically decreasing (non-increasing) function.

$a_{Vor}(\epsilon) = 1, \forall\epsilon \geq 0$ ($\because$ If $a_{Vor}(\epsilon^*) < 1$ for an $\epsilon^* > 0$, then it is a contradictory to $a_{global} = 1$ as $a_{Vor}(\epsilon)$ is a monotonically decreasing function.)

$1 = a_{Vor}(\epsilon) = \mathbb{E}_{(x,c_x)\sim\mathcal{D}}\left[\mathbb{1}\left(f^{*1}(x^*) = c_x\right)\right]$ where $x^* = \underset{x' \in R_{Vor}(\epsilon;x)}{\arg\max} L(x', c_x)$.

As the calculation is based on the finite set X, $f^{*1}(x^*) = c_x$ ($\because$ $\mathbb{1}\left(f^{*1}(x^*) = c_x\right) = 1$) where $x^* = \underset{x' \in R_{Vor}(\epsilon;x)}{\arg\max} L(x', c_x)$.

As x^* are the worst case adversarially perturbed samples, i.e., samples that output mostly different from c_x, $f^{*1}(x') = c_x = f_{1-NN}(x')$ where $x' \in R_{Vor}(\epsilon; x)$.

By changing ϵ and $x \in X$, x' that satisfies $x' \in R_{Vor}(\epsilon; x)$ can fill up $\mathcal{X}$ except for $VB(X)$ ($\because$ Lemma 2). Hence, f^{*1} is equivalent to f_{1-NN} except for Voronoi boundary $VB(X)$.

□