A Theory of Transfer-Based Black-Box Attacks: Explanation and Implications (Supplementary Material)

Anonymous Author(s)
Affiliation
Address
`email`

This article serves as the supplementary material to the central part of our paper. Appendix A includes some further discussions. Complete proofs of the theorems and propositions in Sections 4 and 5 can be found in Appendix B. A multi-class analysis of the manifold attack model is given in Appendix C.

A Further Discussions

A.1 What makes a good explanatory model?

As its title suggests, our paper's primary effort is to explain the properties of TBAs by the manifold attack model. During the writing of this paper, the following question is discussed repeatedly:

What makes a good explanatory model and how to evaluate an explanatory model?

This subsection provides our answer to this question. First of all, we believe that a good explanatory model should be:

- **(Criterion 1)** consistent with existing empirical results,
- **(Criterion 2)** based on reasonable assumptions, and
- **(Criterion 3)** theoretically tractable.

Throughout this paper, we make many efforts to validate our model. Specifically, we try to check whether our model fulfills criteria 1-3. Clearly, our model is theoretically tractable. We theoretically analyze TBAs and provide many explanatory results in Sections 4 and 5.

In the rest of this subsection, we briefly discuss criteria 1 and 2. For the first criterion, we discuss the intriguing properties of TBAs (i.e., the empirical results observed by previous works) in Sections 1 and 2. Two of the most widely-known properties of TBAs are: 1) TBAs can craft transferable adversarial examples even when the source model is inaccurate [1], and 2) the success rates of TBAs are constantly lower than other methods of black-box adversarial attacks [2–4]. Section 4 demonstrates that our model is consistent with the existing empirical results and provides reasonable explanations for these properties.

As for criterion 2, our model assumes that the natural data lies on a low-dimensional manifold. This assumption is commonly seen in previous works [5–8]. We also assume that the classifiers (i.e., the source and target models in TBAs) can be decomposed into the product of a semantic classifier f_b (Definition 4.1) and a concentration multiplier ϕ (Definition 4.2). This assumption is based on the empirical observation that *ML models can capture semantic and geometrical information of the natural data* [9, 10]. Here, our concerns are two folds: 1) what are the semantic and geometrical information, and 2) how does an ML model capture such information?

Semantic information We first focus on semantic information. The following remark explains what is the semantic information of a dataset by an example.

Remark A.1 (The semantic information of CIFAR-10). Generally speaking, "semantic" refers to the relationship between natural data and their true label, which should be consistent with human recognition. For example, the semantic information contained in the CIFAR-10 dataset is the true labels (e.g., airplane, automobile, and bird) and their corresponding natural images (e.g., images of airliners, SUVs, and chickens). In this example, an image cannot simultaneously include an airplane and an automobile since "the classes are completely mutually exclusive" in the CIFAR-10 dataset, cf. the official website of CIFAR-10. That is, the semantic information provided by CIFAR-10 is separated. ▲

In our paper, we formalize the semantic information of natural data by separated sets $A^1, A^2, \cdots, A^k \subset \mathcal{M}$ (for a k-class classification task), see Section 3.2 for the definitions. As is discussed in Remark A.1, these sets reflect the relationship between true labels and their corresponding natural data, and more importantly, these sets should be separated. In this paper, we define separated sets in Definition 3.2 and assume that $A^1, A^2, \cdots, A^k \subset \mathcal{M}$ are separated. It is worth noting that the definition of "semantic information" in our paper is motivated by that of the "concept" in classical learning theory [11, 12]. In these works, learning a concept is equivalent to approximating the decision boundary of ML models to the concept sets (i.e., subsets in the sample space).

Our model captures the semantic information in a similar way as [11]. We let $A_f^1, A_f^2, \cdots, A_f^k \subset \mathcal{M}$ be the semantic information learned by f. Note that we do not assume these sets to be regions or to have any compactness or connectedness restriction. Instead, we only assume that these sets are separated (as the semantic information of natural data). The "similarity" between A_f^i and A^i ($1 \leq i \leq k$) reflects how well the ML model f has learned the semantic information of the training data.

Geometrical information As for the geometrical information, we are motivated by the methods in OOD detection [13–16]. In these works, the scores of the OOD samples are lower than in-distribution samples. In our setting, by the low-dimensional manifold assumption, we know that the off-manifold data are also outside of the data distribution. Thus, by approximating the shape of the manifold, the concentration multiplier ϕ should assign lower scores to those off-manifold samples, see Definition 4.2 for a formal definition. In summary, our paper assumes that natural data is drawn from a low-dimensional manifold and the source and target models capture the semantic and geometrical information in the way we have discussed above. Our assumption is intuitive, reasonable, and milder than previous works that theoretically analyze TBAs. Our model fulfills criterion 2.

Last but not least, the following remark explains why our paper does not present any experiments.

Remark A.2 (Experiments are unnecessary for validating our model). As mentioned in Section 2.1, most of the recent studies on TBAs focus on empirically improving the success rates of TBAs [3, 17]. However, to the best of our knowledge, existing theoretical analyses of TBAs [18–20] are either based on simple models (e.g., linear classifiers) or strong assumptions (e.g., natural data are drawn from the spherical Gaussian distribution). The theoretical studies of TBAs are falling behind the engineering practice, which motivates us to propose an explanatory model that analyzes and explains the existing empirical results. As is discussed in Appendix A.1, we argue that conducting experiments (on either real-world or synthetic datasets) is unnecessary for evaluating an explanatory model. Therefore, we do not include experiments in our paper. ▲

A.2 Visualization of the Non-Adversarial Region

We provide a visualization of Example 4.10 in Figure A.1.

B Complete Proofs

Proposition 4.3 (semantic classifier, binary case)**.** *Given 2λ-separated sets $A_f, B_f \subset \mathcal{M}$. Define:*

$$f_b(\mathbf{x}) = f_b(\mathbf{x}; A_f, B_f) := \frac{d_p(\mathbf{x}, B_f) - d_p(\mathbf{x}, A_f)}{d_p(\mathbf{x}, B_f) + d_p(\mathbf{x}, A_f)}. \tag{B.1}$$

Then, f_b is a semantic classifier. In particular, we can obtain from Equation (B.1) that $f_b(\mathbf{x}) > 0$ if $\mathbf{x}$ is closer (w.r.t. d_p) to A_f than B_f and $f_b(\mathbf{x}) < 0$ otherwise.

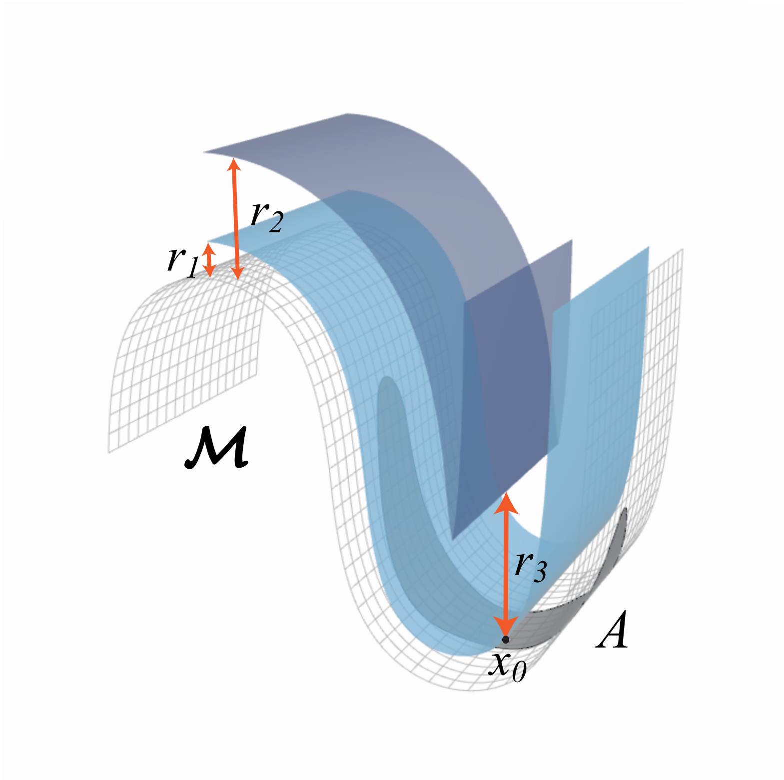

Figure A.1: A visualization of Example 4.10. The data manifold $\mathcal{M}$ is represented by the grid surface. Let the surface in light blue (or dark blue) be the contour surface that $\phi_1 = 0$ (or $\phi_2 = 0$). The distance between $\mathbf{x}_0$ and the dark blue surface is r_3, which is greater than δ and r_2.

Proof of Proposition 4.3. It is easy to check that $f_b(\mathbf{x}) = 1$ when $\mathbf{x} \in A_f$ and $f_b(\mathbf{x}) = -1$ when $\mathbf{x} \in B_f$. By definition, we know that f_b is a semantic classifier. □

Proposition 4.4. *Take $A_f = A$ and $B_f = B$ in Equation (B.1) and denote the corresponding classifier by f_b^*. Then, for any given $\lambda \geq \delta > 0$, we have $R_{\text{std}}(f_b^*) = R_{\text{adv}}(f_b^*, \delta) = 0$.*

Proof of Proposition 4.4. By Equation (B.1), we have

$$f_b^*(\mathbf{x}) = \frac{d_p(\mathbf{x}, B) - d_p(\mathbf{x}, A)}{d_p(\mathbf{x}, B) + d_p(\mathbf{x}, A)}, \ \forall \mathbf{x} \in \mathbb{R}^d. \tag{B.2}$$

Clearly, we have $f_b^*(\mathbf{x}) = 1$ when $\mathbf{x} \in A$ and $f_b^*(\mathbf{x}) = -1$ when $\mathbf{x} \in B$. Then, the standard risk of f_b^* w.r.t. $D(\mathbf{x})$ is

$$R_{\text{std}}(f_b^*) = \mathbb{P}_D\left[f_b^*(\mathbf{x})y < 0 \mid \mathbf{x} \in A\right] + \mathbb{P}_D\left[f_b^*(\mathbf{x})y < 0 \mid x \in B\right] = 0 \tag{B.3}$$

Recall that A and B are 2λ-separated (cf. Definition 3.2). For $\forall \mathbf{x} \in A$ and $\mathbf{x}' \in B(\mathbf{x}, \delta)$, we have $d_p(\mathbf{x}, B) > \delta$, which implies that $d_p(\mathbf{x}', B) - d_p(\mathbf{x}', A) > 0$, and thus $f_b^*(\mathbf{x}')f_b^*(\mathbf{x}) = f_b^*(\mathbf{x}') > 0$. For $\forall \mathbf{x} \in B$, a similar deduction shows that $f_b^*(\mathbf{x}')f_b^*(\mathbf{x}) > 0$ holds for $\forall \mathbf{x}' \in B(\mathbf{x}, \delta)$. Together, we have

$$\begin{aligned} R_{\text{adv}}(f_b^*, \delta) :=& \mathbb{P}\left[\exists \mathbf{x}' \in B(\mathbf{x}; \delta) \ s.t. \ f_b^*(\mathbf{x}')f_b^*(\mathbf{x}) < 0 \mid \mathbf{x} \in A\right] \\ &+ \mathbb{P}\left[\exists \mathbf{x}' \in B(\mathbf{x}; \delta) \ s.t. \ f_b^*(\mathbf{x}')f_b^*(\mathbf{x}) < 0 \mid \mathbf{x} \in B\right] = 0, \end{aligned} \tag{B.4}$$

which completes the proof. □

Remark B.1. The construction of Equation (B.2) can be found in previous works [6, 21]. In particular, Li et al. [6] uses the ReLU-approximation of f_b^* to study the robust generalization of deep NNs.

Proposition 4.5 (Concentration multiplier, binary case)**.** *For any given $r > 0$ and $G \subset \mathbb{R}^d$, denote*

$$\phi(\mathbf{x}) = \phi(\mathbf{x}; r, G) := \frac{r - d_p(\mathbf{x}, G)}{r + d_p(\mathbf{x}, G)},\ \forall \mathbf{x} \in \mathbb{R}^d. \tag{B.5}$$

Then $\phi(\mathbf{x})$ is a concentration multiplier around G.

Proof of Proposition 4.5. For $\forall \mathbf{x} \in G$, we have $d_p(\mathbf{x}, G) = 0$. That is, $\phi(\mathbf{x}) = 1$ for $\forall \mathbf{x} \in G$. For $\forall \mathbf{x}_1, \mathbf{x}_2$ s.t. $d_p(\mathbf{x}_1, G) > d_p(\mathbf{x}_2, G)$, it is easy to check that $\phi(\mathbf{x}_1) < \phi(\mathbf{x}_2)$. $\square$

Proposition 4.6. *Let $f = f_b \cdot \phi$ and A_f, B_f be the semantic information of f_b. We can obtain that*

1. *if $R_{\text{adv}}(f; \delta) \neq 0$, then f suffers from off-manifold adversarial examples.*
2. *if $R_{\text{adv}}(f; \delta) \neq 0$ and $d_p(A \cup B, (A_f \cup B_f)^c) > \delta$, then all the adversarial examples of f are off the manifold.*[1]

Proof of Proposition 4.6. We first prove the first result. By definition, there are $r > 0$ and $G \subset \mathbb{R}^d$ such that

$$f(\mathbf{x}) = \frac{d_p(\mathbf{x}, B_f) - d_p(\mathbf{x}, A_f)}{d_p(\mathbf{x}, B_f) + d_p(\mathbf{x}, A_f)} \cdot \frac{r - d_p(\mathbf{x}, G)}{r + d_p(\mathbf{x}, G)} \tag{B.6}$$

Given $R_{\text{adv}}(f; \delta) \neq 0$, then $\exists \mathbf{x} \in A \cup B$ and $\mathbf{x}_0 \in B(\mathbf{x}, \delta)$ such that $f(\mathbf{x}) f(\mathbf{x}_0) < 0$. If $x_0 \in \mathcal{M}^c$, there is nothing to prove.

Otherwise, we have $\mathbf{x}_0 \in \mathcal{M}$. Without loss of generality (WLOG), we assume that such $\mathbf{x} \in A$ and $f(\mathbf{x}_0) < 0$, which implies that either $f_b(\mathbf{x}_0) < 0$ or $\phi(\mathbf{x}_0) < 0$. We first consider the case when $f_b(\mathbf{x}_0) < 0$. Then, we have $d_p(\mathbf{x}, B_f) - d_p(\mathbf{x}, A_f) < 0$ and $r - d_p(\mathbf{x}, S) > 0$. Denote

$$r_0 := \frac{1}{3} \min\{|d_p(\mathbf{x}_0, A_f) - d_p(\mathbf{x}_0, B_f)|, |r - d_p(\mathbf{x}_0, S)|, \delta\}. \tag{B.7}$$

Consider the non-empty set

$$B(\mathbf{x}, \delta) \cap B(\mathbf{x}_0, r_0) \cap \mathcal{M}^c.$$

For $\forall x_0' \in B(\mathbf{x}_0, r_0)$, there is

$$d_p(\mathbf{x}_0', A_f) \geq d_p(\mathbf{x}_0, A_f) - d_p(\mathbf{x}_0, \mathbf{x}_0'), \tag{B.8}$$

and

$$d_p(\mathbf{x}_0', B_f) \leq d_p(\mathbf{x}_0, B_f) + d_p(\mathbf{x}_0, \mathbf{x}_0'), \tag{B.9}$$

which implies that

$$d_p(\mathbf{x}_0', A_f) - d_p(\mathbf{x}_0', B_f) \geq d_p(\mathbf{x}_0, A_f) - d_p(\mathbf{x}_0, B_f) - 2d_p(\mathbf{x}_0, \mathbf{x}_0') \geq r_0 > 0. \tag{B.10}$$

Similarly, we can obtain $r - d_p(\mathbf{x}_0', S) > 0$. Together, these two inequalities lead us to $f(\mathbf{x}_0') = f(\mathbf{x}_0') f(\mathbf{x}) < 0$, i.e., x_0' is an off-manifold adversarial example of x_0. Some tedious manipulation yields the same result when $\phi(\mathbf{x}_0) < 0$, which is omitted here.

As for the second result, since $R_{\text{adv}}(f; \delta) \neq 0$, we can obtain from the first result that off-manifold adversarial examples exist. It remains to show that f has no on-manifold adversarial examples.

Since $d_p(A \cup B, (A_f \cup B_f)^c) > \delta$ and by assumption $A_f \cup B_f \subset G$, we have

$$\frac{r - d_p(\mathbf{x}, G)}{r + d_p(\mathbf{x}, G)} = 1 \tag{B.11}$$

and

$$\frac{r - d_p(\mathbf{x}', G)}{r + d_p(\mathbf{x}', G)} = 1 \tag{B.12}$$

for $\forall \mathbf{x} \in A \cup B$ and $\mathbf{x}' \in B(\mathbf{x}, \delta) \cap \mathcal{M}$. We can easily obtain that $f_b(\mathbf{x}) = f_b(\mathbf{x}')$, which implies that f has no on-manifold adversarial examples. $\square$

[1] **Notice:** In the main part of the paper, we made a typo in this result. Here, we provide the corrected version. The other results in the main paper are based on the corrected version of this result.

Proposition 4.7. *Consider TBAs with perturbation radius $\delta \in (0, \lambda]$, target model $f_t = f_b^* \cdot \phi_{\text{off}}$ and source model $f_s = f_b \cdot \phi_{\text{off}}$, $f_b \in \mathcal{F}_b$. Denote the semantic information of f_b by A_f and B_f. Then, for $\forall \mathbf{x} \in A \cup B$, all adversarial examples (if exist) of f_s at $\mathbf{x}$ are transferable if $\mathbf{x} \in A_f \cup B_f$.*

Proof of Proposition 4.7. Consider $\mathbf{x} \in A_f$ WLOG. By Equation (6), denote

$$f_s(\mathbf{x}) = f_b(\mathbf{x}) \cdot \phi_{\text{off}}(\mathbf{x}) = \frac{d_p(\mathbf{x}, B_f) - d_p(\mathbf{x}, A_f)}{d_p(\mathbf{x}, B_f) + d_p(\mathbf{x}, A_f)} \cdot \frac{\alpha\delta - d_p(\mathbf{x}, \mathcal{M})}{\alpha\delta + d_p(\mathbf{x}, \mathcal{M})}. \tag{B.13}$$

If f_s is robust against adversarial examples at $\mathbf{x} \in B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A_f$, then there is nothing to prove. If not, denote the adversarial example of f_s at $\mathbf{x}$ by $\mathbf{x}_a$, and we have $f_s(\mathbf{x}_a) < 0$. It is not hard to verify that $d_p(\mathbf{x}_a, B_f) - d_p(\mathbf{x}_a, A_f) > 0$ since $\delta < \lambda$, which implies that $f_b(\mathbf{x}_a) > 0$. To obtain $f_s(\mathbf{x}_a) < 0$, there must be $\phi_{\text{off}}(\mathbf{x}_a) < 0$. We thus have $\alpha\delta < d_p(\mathbf{x}_a, \mathcal{M})$, which implies that $\mathbf{x}_a$ is off the manifold and the distance between $\mathbf{x}_a$ and $\mathcal{M}$ is greater than $\alpha\delta$. In particular, we have

$$\phi_{\text{off}}(\mathbf{x})\phi_{\text{off}}(\mathbf{x}_a) < 0, \tag{B.14}$$

which is independent of the choice of f_b. Now consider $f_t(\mathbf{x})$ and $f_t(\mathbf{x}_a)$, where

$$f_t(\mathbf{x}) = f_b^*(\mathbf{x}) \cdot \phi_{\text{off}}(\mathbf{x}) = \frac{d_p(\mathbf{x}, B) - d_p(\mathbf{x}, A)}{d_p(\mathbf{x}, B) + d_p(\mathbf{x}, A)} \cdot \frac{\alpha\delta - d_p(\mathbf{x}, \mathcal{M})}{\alpha\delta + d_p(\mathbf{x}, \mathcal{M})}. \tag{B.15}$$

No matter $\mathbf{x} \in A$ or $\mathbf{x} \in B$, we have for $\forall \mathbf{x}' \in B(\mathbf{x}, \delta)$, there is $f_b^*(\mathbf{x}) = f_b^*(\mathbf{x}')$ (by the 2λ-separated property of A and B). By Equation (B.14), we have

$$f_t(\mathbf{x})f_t(\mathbf{x}_a) = f_b^*(\mathbf{x})f_b^*(\mathbf{x}_a) \cdot \phi_{\text{off}}(\mathbf{x})\phi_{\text{off}}(\mathbf{x}_a) < 0, \tag{B.16}$$

i.e., $\mathbf{x}_a$ transfers to f_t, which completes the proof. □

Proposition 4.8. *Consider TBA with perturbation radius $\delta \in (0, \lambda]$, target model $f_t = f_b^* \cdot \phi_{\text{on}}(\cdot, f_b^*)$ and source model $f_s = f_b \cdot \phi_{\text{on}}(\cdot, f_b)$, $f_b \in \mathcal{F}_b$. Denote*

$$S_{\text{crt}} := (A \cap A_f) \cup (B \cap B_f),\ S_{\text{wrg}} := (A \cap B_f) \cup (B \cap A_f). \tag{B.17}$$

Then, for $\forall \mathbf{x} \in A \cup B$, we have

1. *if $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset S_{\text{crt}} \cup S_{\text{wrg}}$, then f_t and f_s are both robust against adversarial examples;*
2. *if $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A \cup B$ and $\mathbf{x} \in A_f \cup B_f$, then the adversarial examples of f_s at $\mathbf{x}$ (if exists) cannot transfer to f_t.*

Proof of Proposition 4.8. The proof of the first result is also straightforward, which is omitted here. For $\forall \mathbf{x} \in A \cup B$ such that $(B(\mathbf{x}, \delta) \cap \mathcal{M}) \subset S_{\text{crt}}$, it is easy to check that $\phi_{\text{on}}(\mathbf{x}) = 1$ and $\phi_{\text{on}}(\mathbf{x}_a) = 1$ for $\forall \mathbf{x}_a \in B(\mathbf{x}, \delta)$, which implies that f is robust against adversarial examples. It remains to prove the second result. By Equation (7), denote

$$f_s(\mathbf{x}) = f_b(\mathbf{x}) \cdot \phi_{\text{on}}(\mathbf{x}; f_b) = \frac{d_2(\mathbf{x}, B_f) - d_2(\mathbf{x}, A_f)}{d_2(\mathbf{x}, B_f) + d_2(\mathbf{x}, A_f)} \cdot \frac{\alpha\delta - d_2(\mathbf{x}, \mathcal{N}_\delta(A_f \cup B_f))}{\alpha\delta + d_2(\mathbf{x}, \mathcal{N}_\delta(A_f \cup B_f))}. \tag{B.18}$$

For $\forall \mathbf{x} \in A \cup B$ such that $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A \cup B$ and $\mathbf{x} \in A_f \cup B_f$, denote the unspecific adversarial example (if exist) of f_s at $\mathbf{x}$ by $\mathbf{x}_a$. Assume that $\mathbf{x} \in A \cap B_f$ WLOG. By definition, we have

$$f_b(\mathbf{x}) < 0. \tag{B.19}$$

By the 2λ-separated assumption of A_f and B_f, we have and

$$f_b(\mathbf{x}_a) < 0. \tag{B.20}$$

From $f_s(\mathbf{x})f_s(\mathbf{x}_a) < 0$, we can obtain that $\phi_{\text{on}}(\mathbf{x}; f_b)\phi_{\text{on}}(\mathbf{x}_a; f_b) < 0$. Since $\mathbf{x} \in B_f$, we have

$$d_2(\mathbf{x}, \mathcal{N}_\delta(A_f \cup B_f)) = 0, \tag{B.21}$$

i.e., $\phi_{\text{on}}(\mathbf{x}; f_b) = 1$. Combine this with $\phi_{\text{on}}(\mathbf{x}; f_b)\phi_{\text{on}}(\mathbf{x}_a; f_b) < 0$, we have $\phi_{\text{on}}(\mathbf{x}_a; f_b) < 0$, i.e.,

$$\alpha\delta < d_2(\mathbf{x}_a, \mathcal{N}_\delta(A_f \cup B_f)) < d_2(\mathbf{x}_a, \mathbf{x}) \leq \delta. \tag{B.22}$$

Since $f_t(\mathbf{x}) = 1$ and $\mathbf{x}_a$ is unspecific, it remains to show that $f_t(\mathbf{x}_a) > 0$. By Equation (7), denote

$$f_t(\mathbf{x}) = f_b^*(\mathbf{x}) \cdot \phi_{\text{on}}(\mathbf{x}; f_b^*) = \frac{d_2(\mathbf{x}, B) - d_2(\mathbf{x}, A)}{d_2(\mathbf{x}, B) + d_2(\mathbf{x}, A)} \cdot \frac{\alpha\delta - d_2(\mathbf{x}, \mathcal{N}_\delta(A \cup B))}{\alpha\delta + d_2(\mathbf{x}, \mathcal{N}_\delta(A \cup B))}. \tag{B.23}$$

By $\mathbf{x} \in A$ and the 2λ-separated assumption of A and B, we have $f_b^*(\mathbf{x}_a) > 0$. By $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A$, we have $\mathbf{x}_a \in \mathcal{N}_\delta(A \cup B))$, i.e., $\phi_{\text{on}}(\mathbf{x}_a; f_b^*) > 0$. Together, we have $f_t(\mathbf{x}_a) = f_b^*(\mathbf{x}_a) \cdot \phi_{\text{on}}(\mathbf{x}_a; f_b^*) > 0$, which completes the proof. □

Proposition 4.12. *Given perturbation radius $\delta \in (0, \lambda]$ and target model $f_t = f_b^* \cdot \phi_{\text{off}}(\cdot; r, \mathcal{M})$. Let Δ be the constant specified in Lemma 4.11. Then, for $\forall \mathbf{x} \in A \cup B$, the off-manifold adversarial example of f_t at $\mathbf{x}$ exists if $r < \Delta$.*

Proof of Proposition 4.12. For $\forall \mathbf{x} \in A \cup B$, let $\mathbf{u} \in N_{\mathbf{x}}(\mathcal{M})$ be the normal direction at $\mathbf{x}$ with $\|\mathbf{u}\|_2 = 1$. Since $r < \Delta$, we can find $r_0 > r$ such that $r_0 < \Delta$ and $r_0 < \delta$. Denote

$$\mathbf{x}_a := \mathbf{x} + r_0 \mathbf{u}. \tag{B.24}$$

Clearly, we have $\mathbf{x}_a \in B(\mathbf{x}, \delta)$. Since $\mathcal{N}_\Delta(\mathcal{M})$ is a tubular neighborhood of $\mathcal{M}$, we have

$$d_2(\mathbf{x}_a, \mathcal{M}) = r_0 > r, \tag{B.25}$$

which implies that $\mathbf{x}_a$ is an off-manifold adversarial example of f_t at $\mathbf{x}$. □

Corollary 5.3. *Let f_b be a semantic classifier with semantic information A_f and B_f that satisfy a 2λ-separated property. Given $\epsilon > 0$, there is a ReLU network $\tilde{f}$ with $O((1/\lambda\epsilon)^d) \cdot O(d^2 + d\log(1/\epsilon))$ parameters such that $\|f - \tilde{f}\|_\infty \leq \epsilon$.*

Proof of Corollary 5.3. According to Lemma 5.2, our goal is to upper bound the Lipschitz constant l of f_b. By definition, it suffices to upper bound the supremum of

$$\begin{aligned} s :=& \frac{|f_b(\mathbf{x}_1; A_f, B_f) - f_b(\mathbf{x}_2; A_f, B_f)|}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \\ =& \frac{1}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left| \frac{d_p(\mathbf{x}_1, A_f)}{d_p(\mathbf{x}_1, A_f) + d_p(\mathbf{x}_1, B_f)} - \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \right|. \end{aligned} \tag{B.26}$$

We only need to consider three cases:

1. both of $\mathbf{x}_1, \mathbf{x}_2 \in A_f \cup B_f$, or
2. both of $\mathbf{x}_1, \mathbf{x}_2 \in (A_f \cup B_f)^c$, and
3. either $\mathbf{x}_1$ or $\mathbf{x}_2$ is in $A_f \cup B_f$.

When $\mathbf{x}_1, \mathbf{x}_2 \in A_f \cup B_f$, a trivial verification shows that $s \leq \frac{1}{\lambda}$. We now turn to the second case. By symmetry, let

$$\frac{d_p(\mathbf{x}_1, A_f)}{d_p(\mathbf{x}_1, A_f) + d_p(\mathbf{x}_1, B_f)} - \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} > 0. \tag{B.27}$$

By simplifying Equation (B.26), we can obtain that

$$\begin{aligned} & \frac{|f_b(\mathbf{x}_1; A_f, B_f) - f_b(\mathbf{x}_2; A_f, B_f)|}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \\ =& \frac{1}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left(\frac{d_p(\mathbf{x}_1, A_f)}{d_p(\mathbf{x}_1, A_f) + d_p(\mathbf{x}_1, B_f)} - \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \right) \\ \leq& \frac{1}{2\lambda} \cdot \left(\frac{d_p(\mathbf{x}_1, A_f) - d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \frac{d_p(\mathbf{x}_2, B_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \right. \\ & \left. + \frac{d_p(\mathbf{x}_1, B_f) - d_p(\mathbf{x}_2, B_f)}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \right) \\ \leq& \frac{1}{2\lambda} \cdot (1 \cdot 1 + 1 \cdot 1) = \frac{1}{\lambda}, \end{aligned} \tag{B.28}$$

which implies that $s \leq \frac{1}{\lambda}$ in this case. Finally, we consider the third case. We assume WLOG that $\mathbf{x}_1 \in A_f$ and $\mathbf{x}_2 \in (A_f \cup B_f)^c$. Substitute into Equation (B.26), we have

$$\begin{aligned} s =& \frac{1}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left| \frac{d_p(\mathbf{x}_1, A_f)}{d_p(\mathbf{x}_1, A_f) + d_p(\mathbf{x}_1, B_f)} - \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \right| \\ =& \frac{1}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \frac{d_p(\mathbf{x}_2, A_f)}{d_p(\mathbf{x}_2, A_f) + d_p(\mathbf{x}_2, B_f)} \leq \frac{1}{2\lambda} \end{aligned} \tag{B.29}$$

To sum up above, we have $\sup_{\mathbf{x}_1 \neq \mathbf{x}_2} s = \frac{1}{\lambda}$, which implies that f_b is $\frac{1}{\lambda}$-Lipschitz continuous, as is required. □

Corollary 5.4. *Given $\epsilon > 0$, $r > 0$ and $S \subset [0,1]^d$, there is a ReLU network $\tilde{\phi}$ with $O((1/r\epsilon)^d) \cdot O(d^2 + d\log(1/\epsilon))$ parameters that can approximate $\phi(\cdot; r, S)$ to precision ϵ.*

Proof of Corollary 5.4. We prove this corollary in a similar manner as Corollary 5.3, i.e., we upper bound the supremum of

$$s := \frac{|\phi(\mathbf{x}_1; r, S) - \phi(\mathbf{x}_2; r, S)|}{d_p(\mathbf{x}_1, \mathbf{x}_2)} = \frac{1}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left| \frac{d_p(\mathbf{x}_1, S)}{r + d_p(\mathbf{x}_1, S)} - \frac{d_p(\mathbf{x}_2, S)}{r + d_p(\mathbf{x}_2, S)} \right| = \frac{r}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left| \frac{1}{r + d_p(\mathbf{x}_1, S)} - \frac{1}{r + d_p(\mathbf{x}_2, S)} \right|. \tag{B.30}$$

We also consider three cases in this proof:

1. both of $\mathbf{x}_1, \mathbf{x}_2 \in S$, or
2. both of $\mathbf{x}_1, \mathbf{x}_2 \in S^c$, and
3. either $\mathbf{x}_1$ or $\mathbf{x}_2$ is in S.

In case 1, we see at once that $s = 0$. When both of $\mathbf{x}_1, \mathbf{x}_2 \in S^c$, we assume WLOG that $d_p(\mathbf{x}_1, S) > d_p(\mathbf{x}_2, S)$. By Equation (B.30), we have

$$s = \frac{d_p(\mathbf{x}_1, S) - d_p(\mathbf{x}_2, S)}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \frac{r}{(r + d_p(\mathbf{x}_1, S))(r + d_p(\mathbf{x}_2, S))} \leq \frac{1}{r}. \tag{B.31}$$

Analysis similar to Equation (B.31) shows that

$$s = \frac{r}{d_p(\mathbf{x}_1, \mathbf{x}_2)} \cdot \left(\frac{1}{r} - \frac{1}{r + d_p(\mathbf{x}_2, S)} \right) \leq \frac{1}{r}. \tag{B.32}$$

To sum up above, we have $\sup_{\mathbf{x}_1 \neq \mathbf{x}_2} s = \frac{1}{\lambda}$, which implies that ϕ is $\frac{1}{r}$-Lipschitz continuous, as is required. □

Proposition 5.6. *Given $\epsilon, \lambda, \delta, r > 0$, for any $f \in \mathcal{F}_{\mathcal{M}}$, there is a ReLU network $\tilde{f}$ with*

$$O(\max\{\frac{1}{\lambda\epsilon}, \frac{2}{r\epsilon}\}^d) \cdot O(d^2 + d\log(\frac{1}{\epsilon})) + O(\log^2(\frac{1}{\epsilon})) \tag{B.33}$$

parameters that satisfies $\|f - \tilde{f}\|_\infty \leq \epsilon$.

Proof of Proposition 5.6. This proposition can be derived directly from Lemma 5.5, Corollary 5.3, and Corollary 5.4. □

Theorem 5.7. *Consider TBAs with perturbation radius $\delta \in (0, \lambda/2]$, target model $f_t = f_b^* \cdot \phi_{\text{off}}$ and source model $f_s = f_b \cdot \phi_{\text{off}}$, $f_b \in \mathcal{F}_b$. Denote the semantic information of f_b by A_f and B_f. Given $\epsilon \leq 0.1$, let $\tilde{f}_t$ and $\tilde{f}_s$ be ReLU networks that satisfy*

$$\|\tilde{f}_t - f_t\|_\infty \leq \epsilon, \|\tilde{f}_s - f_s\|_\infty \leq \epsilon \tag{B.34}$$

Then, for $\forall \mathbf{x} \in (A \cup B) \cap (A_f \cup B_f)$, the adversarial examples $\mathbf{x}_a$ (if exist) of $\tilde{f}_s$ satisfies

$$\tilde{f}_t(\mathbf{x}) \cdot \tilde{f}_t(\mathbf{x}_a) \leq 2\epsilon(1 + \epsilon)^2 + 2\epsilon^2. \tag{B.35}$$

Proof of Theorem 5.7. Consider $\mathbf{x} \in A_f$ WLOG. If f_s is robust against adversarial examples at $\mathbf{x} \in B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A_f$, then there is nothing to prove. If not, denote the adversarial example of f_s at $\mathbf{x}$ by $\mathbf{x}_a$. Since $\mathbf{x} \in A_f \subset \mathcal{M}$, there is

$$\tilde{f}_s(\mathbf{x}) \geq \tilde{f}_b(\mathbf{x}) \cdot \tilde{\phi}_{\text{off}}(\mathbf{x}) - \epsilon \geq (1 - \epsilon)^2 - \epsilon > 0 \tag{B.36}$$

and we thus have $\tilde{f}_s(\mathbf{x}_a) < 0$. By $\mathbf{x}_a \in B(\mathbf{x}; \delta)$ and the assumption $\delta < \frac{\lambda}{2}$, we have

$$\tilde{f}_b(\mathbf{x}_a) = 1 - \frac{2d_p(\mathbf{x}_a, A_f)}{d_p(\mathbf{x}_a, B_f) + d_p(\mathbf{x}_a, A_f)} \geq 1 - \frac{2\delta}{2\lambda} \geq \frac{1}{2}. \tag{B.37}$$

To obtain $\tilde{\times}(\tilde{f}_b, \tilde{\phi}_{\text{off}})(\mathbf{x}_a) < 0$, there must be $\tilde{f}_b(\mathbf{x}_a) \cdot \tilde{\phi}_{\text{off}}(\mathbf{x}_a) < \epsilon$, which implies that

$$\tilde{\phi}_{\text{off}}(\mathbf{x}_a) < 2\epsilon. \tag{B.38}$$

Now consider $\tilde{f}_t(\mathbf{x})$ and $\tilde{f}_t(\mathbf{x}_a)$. By definition, we have $\tilde{f}_b^*(\mathbf{x}) \in [1-\epsilon, 1+\epsilon]$, $\tilde{\phi}_{\text{off}}(\mathbf{x}) \in [1-\epsilon, 1+\epsilon]$, and thus

$$\tilde{f}_t(\mathbf{x}) = \tilde{\times}(\tilde{f}_b^*, \tilde{\phi}_{\text{off}})(\mathbf{x}) \le (1+\epsilon)^2 + \epsilon. \tag{B.39}$$

Similar to Equation (B.37), there is

$$\tilde{f}_b^*(\mathbf{x}_a) = 1 - \frac{2d_p(\mathbf{x}_a, A)}{d_p(\mathbf{x}_a, B) + d_p(\mathbf{x}_a, A)} \le 1 \tag{B.40}$$

Combining Equations (B.38) to (B.40) together, we have

$$\tilde{f}_t(\mathbf{x}) \cdot \tilde{f}_t(\mathbf{x}_a) \le 2\epsilon(1+\epsilon)^2 + 2\epsilon^2. \tag{B.41}$$

as is required. □

Theorem 5.8. *Consider TBAs with perturbation radius $\delta \in (0, \lambda/2]$, target model $f_t = f_b^* \cdot \phi_{\text{off}}$ and source model $f_s = f_b \cdot \phi_{\text{off}}$, $f_b \in \mathcal{F}_b$. Denote the semantic information of f_b by A_f and B_f. Given $\epsilon \le 0.1$, let $\tilde{f}_t$ and $\tilde{f}_s$ be ReLU networks that satisfy Equation (13). Then, for $\forall \mathbf{x} \in A \cup B$, we have*

1. *if $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset S_{\text{crt}} \cup S_{\text{wrg}}$, then $\tilde{f}_t$ and $\tilde{f}_s$ are both robust against adversarial examples;*
2. *if $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A \cup B$ and $\mathbf{x} \in A_f \cup B_f$, then the adversarial examples of $\tilde{f}_s$ at $\mathbf{x}$ (if exists) cannot transfer to $\tilde{f}_t$.*

Proof of Theorem 5.8. The proof of the first result is also straightforward, which is omitted here. It remains to prove the second result. By Equation (7), denote

$$f_s(\mathbf{x}) = f_b(\mathbf{x}) \cdot \phi_{\text{on}}(\mathbf{x}; f_b) = \frac{d_2(\mathbf{x}, B_f) - d_2(\mathbf{x}, A_f)}{d_2(\mathbf{x}, B_f) + d_2(\mathbf{x}, A_f)} \cdot \frac{\alpha\delta - d_2(\mathbf{x}, \mathcal{N}_\delta(A_f \cup B_f))}{\alpha\delta + d_2(\mathbf{x}, \mathcal{N}_\delta(A_f \cup B_f))}. \tag{B.42}$$

For $\forall \mathbf{x} \in A \cup B$ such that $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A \cup B$ and $\mathbf{x} \in A_f \cup B_f$, denote the unspecific adversarial example (if exist) of f_s at $\mathbf{x}$ by $\mathbf{x}_a$. Assume that $\mathbf{x} \in A \cap A_f$ WLOG. By definition, we have

$$\tilde{f}_b(\mathbf{x}) \in [1-\epsilon, 1+\epsilon]. \tag{B.43}$$

By the 2λ-separated assumption of A_f and B_f, we have and

$$\tilde{f}_b(\mathbf{x}_a) \in [1-\epsilon, 1+\epsilon]. \tag{B.44}$$

Since $\mathbf{x} \in A_f \cup B_f$, we have $\mathbf{x} \in \mathcal{N}_\delta(A_f \cup B_f)$ and

$$\tilde{\phi}_{\text{on}}(\mathbf{x}; f_b) \in [1-\epsilon, 1+\epsilon], \tag{B.45}$$

which implies that

$$\tilde{f}_s(\mathbf{x}) = \tilde{\times}(\tilde{f}_b, \tilde{\phi}_{\text{on}}(\cdot; f_b))(\mathbf{x}) \ge (1-\epsilon)^2 - \epsilon > 0. \tag{B.46}$$

From $\tilde{f}_s(\mathbf{x})\tilde{f}_s(\mathbf{x}_a) < 0$, we can obtain that $\tilde{f}_s(\mathbf{x}_a) < 0$, which implies that

$$\tilde{\phi}_{\text{on}}(\mathbf{x}_a; f_b) \cdot \tilde{f}_b(\mathbf{x}) < \epsilon, \tag{B.47}$$

which implies that

$$\tilde{\phi}_{\text{on}}(\mathbf{x}_a; f_b) < \frac{\epsilon}{1-\epsilon} < 2\epsilon. \tag{B.48}$$

By definition, we have

$$\tilde{f}_t(\mathbf{x}) = \tilde{\times}(\tilde{f}_b^*, \tilde{\phi}_{\text{on}}(\cdot; f_b^*))(\mathbf{x}) \ge (1-\epsilon)^2 - \epsilon > 0. \tag{B.49}$$

By $\mathbf{x} \in A$ and the 2λ-separated assumption of A and B, we have

$$\tilde{f}_b^*(\mathbf{x}_a) = 1 - \frac{2d_p(\mathbf{x}_a, A)}{d_p(\mathbf{x}_a, B) + d_p(\mathbf{x}_a, A)} \ge 1 - \frac{2\delta}{2\lambda} \ge \frac{1}{2}. \tag{B.50}$$

By $B(\mathbf{x}, \delta) \cap \mathcal{M} \subset A$, we have $\mathbf{x}_a \in \mathcal{N}_\delta(A \cup B))$, i.e.,

$$\phi_{\text{on}}(\mathbf{x}_a; f_b^*) \in [1-\epsilon, 1+\epsilon] \tag{B.51}$$

Together, we have

$$\tilde{f}_t(\mathbf{x}_a) = \tilde{\times}(\tilde{f}_b^*, \tilde{\phi}_{\text{on}}(\cdot; f_b^*))(\mathbf{x}_a) \ge \frac{1-\epsilon}{2} > 0, \tag{B.52}$$

which completes the proof. □

Proposition 5.9. *For any classifier f^* with $R_{\text{std}}(f^*) = 0$ and perturbation radius $\delta \in (r_\delta(f^*), \lambda)$, there is $f \in \mathcal{F}_\mathcal{M}$ such that*

1. $R_{\text{std}}(f) = R_{\text{std}}(f^*)$*, and*
2. *for $\forall \mathbf{x} \in A \cup B$, if $\mathbf{x}_a$ is an adversarial example of f^* at $\mathbf{x}$, then exists $\mathbf{x}'_a \in B(\mathbf{x}_a, r_\delta(f^*)/4)$ such that $\mathbf{x}'_a$ is an adversarial example of f.*

Proof of Proposition 5.9. Define the following set

$$S_a := \{\mathbf{x} \in [0,1]^d \cap (A \cup B)^c : \exists \mathbf{x}' \in A \cup B \ s.t.\ \mathbf{x}' \in B(\mathbf{x}; \delta), f^*(\mathbf{x}) f^*(\mathbf{x}') < 0\}, \tag{B.53}$$

and let

$$G = \left(\bigcup_{\mathbf{x} \in S_a} B(\mathbf{x}, r_\delta(f^*)/2) \right)^c. \tag{B.54}$$

By definition, $A \cup B \in G$. Consider

$$f(\mathbf{x}) = f_b^*(\mathbf{x}) \cdot \phi(\mathbf{x}; r_\delta(f^*)/4, G). \tag{B.55}$$

Since $A \cup B \in G$, we have $R_{\text{std}}(f) = R_{\text{std}}(f_b^*) = 0 = R_{\text{std}}(f)^*$. For $\forall \mathbf{x} \in A \cup B$, if $\mathbf{x}_a$ is an adversarial example of f^* at $\mathbf{x}$, then

$$\frac{r_\delta(f^*)}{4} \le d_p(\mathbf{x}_a, G) \le \frac{r_\delta(f^*)}{2}, \tag{B.56}$$

which implies that exists $\mathbf{x}'_a \in B(\mathbf{x}_a, r_\delta(f^*)/4)$ such that $\mathbf{x}'_a$ is an adversarial example of f. □

C Analyses in Multi-Class Classification Problems

This section some of the results in Sections 4 and 5 to k-class classification problems. We first extend Propositions 4.3 and 4.5 to multi-class classification.

Proposition C.1 (Semantic classifier, multi-class case)**.** *Given 2λ-separated sets $A_f^1, A_f^2, \cdots, A_f^k \subset \mathcal{M}$. Consider $f_b(\mathbf{x}) = (f_b^{(1)}(\mathbf{x}), f_b^{(2)}(\mathbf{x}), \cdots, f_b^{(k)}(\mathbf{x}))^T$ and define:*

$$f_b^{(i)}(\mathbf{x}) := \frac{\left(\sum_{j \ne i} d_p(\mathbf{x}, A_f^j)\right) - d_p(\mathbf{x}, A_f^i)}{\left(\sum_{j \ne i} d_p(\mathbf{x}, A_f^j)\right) + d_p(\mathbf{x}, A_f^i)} \tag{C.57}$$

for $\forall 1 \le i \le k$. Then, f_b is a semantic classifier.

Proof of Proposition C.1. By Equation (C.57), we have

$$f_b^{(i)}(\mathbf{x}) = \frac{\left(\sum_{j=1}^k d_p(\mathbf{x}, A_f^j)\right) - d_p(\mathbf{x}, A_f^i)}{\sum_{j=1}^k d_p(\mathbf{x}, A_f^j)} \tag{C.58}$$

for $\forall 1 \le i \le k$. Then, there is

$$y(f_b, \mathbf{x}) = \arg\max_{1 \le i \le k} f_b^{(i)}(\mathbf{x}) = \arg\max_{1 \le i \le k} \left(-d_p(\mathbf{x}, A_f^i)\right) = \arg\min_{1 \le i \le k} d_p(\mathbf{x}, A_f^i). \tag{C.59}$$

Given that $A_f^1, A_f^2, \cdots, A_f^k$ are 2λ-separated, we have

$$0 = d_p(\mathbf{x}, A_f^i) < d_p(\mathbf{x}, A_f^j) \tag{C.60}$$

for $\forall j \ne i$ if $\mathbf{x} \in A_f^j$, which completes the proof. □

Next, we specify a family of concentration multipliers for multi-class TBAs.

Proposition C.2 (Concentration multiplier, multi-class case)**.** *For any given $r > 0$ and $G \subset \mathbb{R}^d$, denote*

$$\phi(\mathbf{x}) = \phi(\mathbf{x}; r, G) := \frac{r - d_p(\mathbf{x}, G)}{r + d_p(\mathbf{x}, G)}, \ \forall \mathbf{x} \in \mathbb{R}^d. \tag{C.61}$$

Then $\phi(\mathbf{x})$ is a concentration multiplier around G.

Note that Equation (C.61) is identical to Equation (5). The proof of Proposition C.2 is therefore omitted. The following proposition extends Proposition 4.4 to the multi-class case.

Proposition C.3. *Take $A_f^i = A^i$ in Equation (C.57) for $\forall 1 \leq i \leq k$. Denote the corresponding classifier by f_b^*. Then, for any given $\lambda \geq \delta > 0$, we have $R_{\text{std}}(f_b^*) = R_{\text{adv}}(f_b^*, \delta) = 0$.*

Proof of Proposition C.3. By Equation (C.57), we have

$$f_b^*(\mathbf{x}) = \frac{\left(\sum_{j \neq i} d_p(\mathbf{x}, A^j)\right) - d_p(\mathbf{x}, A^i)}{\left(\sum_{j \neq i} d_p(\mathbf{x}, A^j)\right) + d_p(\mathbf{x}, A^i)} \tag{C.62}$$

Apparently, we have $y(f_b^*, \mathbf{x}) = i$ when $\mathbf{x} \in A^i$ for $\forall 1 \leq i \leq k$. The standard risk of f_b^* w.r.t. D is

$$R_{\text{std}}(f_b^*) = \sum_{i=1}^{k} \left(\mathbb{P}_D \left[y(f_b^*, \mathbf{x}) \neq y(\mathbf{x}) \mid \mathbf{x} \in A^i\right]\right) = 0. \tag{C.63}$$

For $\forall i \neq j$, recall that A and B are 2λ-separated (cf. Definition 3.2). For $\forall \mathbf{x} \in A^i$ and $\mathbf{x}' \in B(\mathbf{x}, \delta)$, we have $d_p(\mathbf{x}', A^j) > \delta$, which implies that $d_p(\mathbf{x}', A_j) > d_p(\mathbf{x}', A^i)$ and

$$\left(\sum_{l=1}^{k} d_p(\mathbf{x}', A^l)\right) - d_p(\mathbf{x}', A^i) > \left(\sum_{l=1}^{k} d_p(\mathbf{x}', A^l)\right) - d_p(\mathbf{x}', A^j). \tag{C.64}$$

Since j is arbitrarily chosen, and according to Equation (C.58), we have

$$f_b^{(i)}(\mathbf{x}') > f_b^{(j)}(\mathbf{x}') \tag{C.65}$$

holds for $\forall j \neq i$, i.e., $y(f_b^*, \mathbf{x}) = y(f_b^*, \mathbf{x}')$ for $\forall \mathbf{x}' \in B(\mathbf{x}, \delta)$. Then, the adversarial risk of f_b^* is

$$R_{\text{adv}}(f_b^*, \delta) = \sum_{i=1}^{k} \left(\mathbb{P}_D \left[\exists \mathbf{x}_a \in B(\mathbf{x}; \delta) \ s.t. \ y(f, \mathbf{x}) \neq y(f, \mathbf{x}_a) \mid \mathbf{x} \in A^i\right]\right) = 0, \tag{C.66}$$

which completes the proof. □

Next, we go straight for the two explanatory results. We first note that the non-existence of off-manifold adversarial examples is due to the "sharp curvature" of the data manifold. The analyses in Example 4.10 are regardless of whether the task is binary or multi-class. Here, we extend Proposition 4.12 to multi-class cases. Consider TBAs with perturbation radius $\delta \in (0, \lambda]$ For any unspecified $\alpha \in (0, 1)$, let

$$\phi_{\text{off}}(\mathbf{x}) := \phi(\mathbf{x}; \alpha\delta, \mathcal{M}) = \frac{\alpha\delta - d_p(\mathbf{x}, \mathcal{M})}{\alpha\delta + d_p(\mathbf{x}, \mathcal{M})}. \tag{C.67}$$

Recall that Proposition 4.12 is restricted to $p = 2$. The following proposition provides a sufficient condition for the existence of off-manifold adversarial examples in multi-class classification tasks.

Proposition C.4. *Given perturbation radius $\delta \in (0, \lambda]$ and target model $f_t = f_b^* \cdot \phi_{\text{off}}$. Let Δ be the constant specified in Lemma 4.11. Then, for $\forall \mathbf{x} \in \cup_{i=1}^k A^i$, the off-manifold adversarial example of f_t at $\mathbf{x}$ exists if $\alpha\delta < \Delta$.*

Proof of Proposition C.4. For $\forall \mathbf{x} \in \cup_{i=1}^k A^i$, let $\mathbf{u} \in N_{\mathbf{x}}(\mathcal{M})$ be the normal direction at $\mathbf{x}$ with $\|\mathbf{u}\|_2 = 1$. Since $\alpha\delta < \Delta$, we can find $r_0 > \alpha\delta$ such that $r_0 < \Delta$ and $r_0 < \delta$. Denote

$$\mathbf{x}_a := \mathbf{x} + r_0 \mathbf{u}. \tag{C.68}$$

Clearly, we have $\mathbf{x}_a \in B(\mathbf{x}, \delta)$. Since $\mathcal{N}_\Delta(\mathcal{M})$ is a tubular neighborhood of $\mathcal{M}$, we have

$$d_2(\mathbf{x}_a, \mathcal{M}) = r_0 > \alpha\delta. \tag{C.69}$$

For $\forall i \in \{1, 2, \cdots, k\}$, by definition, we have

$$f_t^{(l)}(\mathbf{x}_a) = f_b^{*,(l)}(\mathbf{x}_a) \cdot \phi_{\text{off}}(\mathbf{x}_a) = \frac{\left(\sum_{l \neq i} d_2(\mathbf{x}_a, A^l)\right) - d_2(\mathbf{x}_a, A^i)}{\left(\sum_{l \neq i} d_2(\mathbf{x}_a, A^l)\right) + d_2(\mathbf{x}_a, A^i)} \cdot \frac{\alpha\delta - d_2(\mathbf{x}, \mathcal{M})}{\alpha\delta + d_2(\mathbf{x}, \mathcal{M})}. \tag{C.70}$$

Since $A_f^1, A_f^2, \cdots, A_f^k$ are 2λ-separated, we can easily obtain that

$$y(f_b^*, \mathbf{x}_a) = y(f_b^*, \mathbf{x}). \tag{C.71}$$

Combining Equations (C.69) to (C.71) together, we can obtain that $\mathbf{x}_a$ is an off-manifold adversarial example of f_t at $\mathbf{x}$, since ϕ_{off} is negative and thus turn the arg max of f_b^* to the arg min. □

Let $\mathcal{F}_b$ and Φ be the function class defined in Proposition C.1 and Proposition C.2, respectively. The following proposition extends Proposition 4.7 to multi-class classification tasks based on the results in Proposition C.4.

Proposition C.5. *Denote the target model by $f_t = f_b^* \cdot \phi_{\text{off}}$ and the source model $f_s = f_b \cdot \phi_{\text{off}}$, $f_b \in \mathcal{F}_b$. With some abuse of notation, let the semantic information of f_b be $A_f^1, A_f^2, \cdots, A_f^k$. Let Δ be the constant specified in Lemma 4.11. Assume that $\alpha\delta < \Delta$. Then, for $\forall \mathbf{x} \in \cup_{i=1}^k A^i$, exists adversarial example of f_s at $\mathbf{x}$ that is transferable if $\mathbf{x} \in \cup_{i=1}^k A_f^i$.*

In the main part of our paper, Proposition 4.7 proves that adversarial examples are transferable even if the source model is accurate, which is consistent with the empirical results in Papernot et al. [1]. Proposition C.5 also explains this phenomenon, even though it is weaker than Proposition 4.7.

Proof of Proposition C.5. For $\forall i \in \{1, 2, \cdots, k\}$, we first consider those $\mathbf{x} \in A^i$. By definition, we have

$$f_s^{(l)}(\mathbf{x}) = f_b^{(l)}(\mathbf{x}) \cdot \phi_{\text{off}}(\mathbf{x}) = \frac{\left(\sum_{l \neq i} d_p(\mathbf{x}, A_f^l)\right) - d_p(\mathbf{x}, A_f^i)}{\left(\sum_{l \neq i} d_p(\mathbf{x}, A_f^l)\right) + d_p(\mathbf{x}, A_f^i)} \cdot \frac{\alpha\delta - d_p(\mathbf{x}, \mathcal{M})}{\alpha\delta + d_p(\mathbf{x}, \mathcal{M})}. \tag{C.72}$$

According to Proposition C.4, we know that off-manifold adversarial examples exist. In fact, for $\forall \mathbf{x} \in \cup_{i=1}^k A_f^i$, let $\mathbf{x}_a$ be as defined in Equation (C.68). Similar to the proof of Proposition C.4, we have $\phi_{\text{off}}(\mathbf{x}_a) < 0$ and

$$y(f_b^*, \mathbf{x}_a) = y(f_b^*, \mathbf{x}), \; y(f_b, \mathbf{x}_a) = y(f_b, \mathbf{x}). \tag{C.73}$$

which implies that $\mathbf{x}_a$ is a transferable adversarial example. □

References

[1] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *AsiaCCS*, 2017.

[2] Yinpeng Dong, Shuyu Cheng, Tianyu Pang, Hang Su, and Jun Zhu. Query-efficient black-box adversarial attacks guided by a transfer-based prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[3] Shuman Fang, Jie Li, Xianming Lin, and Rongrong Ji. Learning to learn transferable attack. In *AAAI*, 2022.

[4] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *ICLR*, 2020.

[5] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. In *NeurIPS*, pages 8172–8182, 2019.

[6] Binghui Li, Jikai Jin, Han Zhong, John E. Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power. *CoRR*, abs/2205.13863, 2022.

[7] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. In *NeurIPS*, 2020.

[8] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021.

[9] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[10] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *FOCS*, 2021.

[11] Sanjeev R. Kulkarni, Sanjoy K. Mitter, and John N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Mach. Learn.*, 1993.

[12] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[13] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *CVPR*, 2020.

[14] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[16] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[17] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *AAAI*, 2022.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[19] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. Adversarial spheres. In *ICLR*, 2018.

[20] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019.

[21] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *NeurIPS*, 2020.