# Theoretical Explanation for Generalization from Adversarial Perturbations

**Soichiro Kumano**
The University of Tokyo
kumano@cvm.t.u-tokyo.ac.jp

**Hiroshi Kera**
Chiba University
kera@chiba-u.jp

**Toshihiko Yamasaki**
The University of Tokyo
yamasaki@cvm.t.u-tokyo.ac.jp

## Abstract

It is not fully understood why adversarial examples can deceive neural networks and transfer between different networks. To elucidate this, several studies hypothesized that adversarial perturbations contain data features that are imperceptible to humans but still recognizable by neural networks. Empirical evidence has shown that neural networks trained on mislabeled samples with these perturbations can generalize to natural test data. However, a theoretical understanding of this counterintuitive phenomenon is limited. In this study, assuming orthogonal training samples, we first prove that one-hidden-layer neural networks can learn natural data structures from adversarial perturbations. Our results indicate that, under mild conditions, the decision boundary from learning perturbations aligns with that from natural data, except for specific points in the input space.

## 1 Introduction

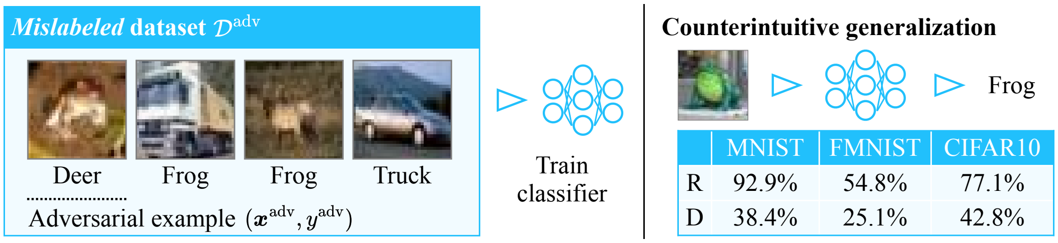
It is well known that a small malicious perturbation, or an adversarial perturbation, can change a classifier's prediction from the correct class to an incorrect class [16]. An interesting observation by [8] has shown that a classifier, trained on natural samples with adversarial perturbations labeled by such incorrect classes, can generalize to unperturbed data. Specifically, the procedure is as follows:

**Definition 1.1** (Learning from adversarial perturbations (later redefined) [8]**).** Let $\mathcal{D} := \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ be a training dataset, where $\boldsymbol{x}_n$ denotes an input (e.g., an image) and $y_n$ denotes the corresponding label. Let $f$ be a classifier trained on $\mathcal{D}$. For each $n$, the adversarial example $\boldsymbol{x}_n^{\mathrm{adv}}$ is produced by imposing an adversarial perturbation on $\boldsymbol{x}_n$ to increase the probability for a target label $y_n^{\mathrm{adv}} \neq y_n$ given by $f$, constructing $\mathcal{D}^{\mathrm{adv}} := \{(\boldsymbol{x}_n^{\mathrm{adv}}, y_n^{\mathrm{adv}})\}_{n=1}^N$. Training a classifier from scratch on $\mathcal{D}^{\mathrm{adv}}$ is called *learning from adversarial perturbations*.

Notably, a training sample $\boldsymbol{x}^{\mathrm{adv}}$ appears almost identical to $\boldsymbol{x}$ for humans but is labeled as a different class $y^{\mathrm{adv}} \neq y$. Nevertheless, neural networks can learn to accurately classify benign samples $\{\boldsymbol{x}_n\}_{n=1}^N$ from such adversarially perturbed samples with seemingly incorrect labels.

The unexpected success of learning from adversarial perturbations suggests that they may contain data features that are class-specific but imperceptible to humans, helping classifiers understand data structures. This hypothesis suggests intriguing properties of adversarial examples. For example, classifier misclassifications may be caused by their sensitivity to the features in perturbations. In addition, transferability across different classifiers [2, 5, 7] can be interpreted as classifiers responding to the same features in perturbations.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.


*Mislabeled* dataset $\mathcal{D}^{\text{adv}}$
Deer
Frog
Frog
Truck
Adversarial example $(\boldsymbol{x}^{\text{adv}}, y^{\text{adv}})$
Train classifier
Counterintuitive generalization
Frog

| | MNIST | FMNIST | CIFAR10 |
|---|---|---|---|
| R | 92.9% | 54.8% | 77.1% |
| D | 38.4% | 25.1% | 42.8% |

Figure 1: Learning from adversarial perturbations (cf. Definition 1.1). Classifiers trained on a mislabeled dataset with adversarial perturbations achieve high test accuracy on correctly labeled test datasets, as shown in the table. "R" denotes random selection of the adversarial target label $y_n^{\text{adv}}$ from nine non-original labels, while "D" denotes deterministic selection one after the original. Detailed experimental settings can be found in Appendix C.

The hypothesis that adversarial perturbations contain data features has drawn the attention of the research community, leading to extensive discussions [1, 4, 9, 14, 18]. However, many of these discussions are empirical and their theoretical understanding remains limited. For example, it is still unknown the manner in which an adversarial perturbation encapsulates features, similarity between the decision boundaries of learning from standard data and perturbations, and feasibility of learning from perturbations in a high-dimensional dataset with numerous samples.

In this study, we provide the first theoretical validation of the learnability from adversarial perturbations. By leveraging recent results on the decision boundary of a one-hidden-layer neural network trained on orthogonal data [6], we prove that, under mild conditions, the decision boundary from adversarial perturbations becomes consistent with that from natural data, except for specific points in the input space. That is, for most test samples, a one-hidden-layer network trained on seemingly mislabeled data produces predictions consistent with those of a normally trained network.

## 2 Related Work

Ilyas et al. first claimed that an adversarial perturbation contains data features, called non-robust features [8]. These features are highly predictive and generalizable, yet brittle and incomprehensible to humans. This idea is supported by neural networks that learn from perturbations (cf. Definition 1.1) achieving good test accuracies on standard datasets [8]. Subsequent studies deepened the discussion of non-robust features. While Ilyas et al. considered robust and non-robust features separately [8], Springer et al. claimed their potential entanglement [14]. Some studies have attempted to separate robust and non-robust features using the information bottleneck [9] and neural tangent kernel [18]. Engstrom et al. conducted a broad discussion on topics such as robust neural style transfer and robust feature leakage [4]. Other studies leveraged the feature hypothesis to generate highly transferable adversarial examples [13–15], understand the behavior of batch normalization in adversarial training [1], and degrade the robustness of adversarially trained models [17]. However, the nature of adversarial perturbations as data features and the theoretical explanation for the counterintuitive success of perturbation learning remain unclear.

In this study, we first justify learning from perturbations, which is an essential foundation for validating the feature hypothesis. Our results support those of the aforementioned studies based on the feature hypothesis. We do not consider whether adversarial perturbations are robust or non-robust features or their entanglement. We primarily discuss whether adversarial perturbations are features or bugs and why classifiers can obtain generalization ability from perturbations.

## 3 Preliminary

### 3.1 Settings

**Network.** Our network settings follow [6]. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a one-hidden-layer neural network. The number of hidden neurons is even and is denoted by $m$. We assume that the hidden layer is trainable and that the last layer is frozen to constant weights $\boldsymbol{a} \in \mathbb{R}^m$. The first half elements of $\boldsymbol{a}$

are $1/\sqrt{m}$ and the latter half are $-1/\sqrt{m}$. Let $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)^\top \in \mathbb{R}^m \times \mathbb{R}^d$ be the weights of the hidden layer. Let $\phi(z) := \max(z, \gamma z)$ be the element-wise leaky ReLU for a constant $\gamma \in (0, 1)$. That is, $f(\boldsymbol{x}) := \boldsymbol{a}^\top \boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x})$. The assumption that the positive and negative values of $\boldsymbol{a}$ are equal is introduced for notational simplicity and is fundamentally unnecessary.

**Training.** Let $\mathcal{D} := \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \{\pm 1\}$ be a training dataset, where $N \in \mathbb{N}$ is the number of training samples. With a loss function $\ell : \mathbb{R} \to \mathbb{R}$, the loss of $f(\boldsymbol{x}; \boldsymbol{W})$ over $\mathcal{D}$ is defined as $\mathcal{L}(\boldsymbol{W}; \mathcal{D}) := \frac{1}{N} \sum_{n=1}^N \ell(y_n f(\boldsymbol{x}_n; \boldsymbol{W}))$. We consider the exponential loss $\ell(z) = \exp(-z)$ and logistic loss $\ell(z) = \ln(1 + \exp(-z))$. The network parameters are updated by gradient flow, gradient descent with an infinitesimal step size. Namely, $\boldsymbol{W}$ is updated as $\mathrm{d}\boldsymbol{W}(t)/\mathrm{d}t = -\boldsymbol{\nabla}_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}(t); \mathcal{D})$, where $t \geq 0$ is a continuous training step. Finally, we summarize the training setting as follows:

**Setting 3.1** (Training)**.** Consider training a one-hidden-layer neural network $f$ on dataset $\mathcal{D}$. The network parameter $\boldsymbol{W}$ is updated by minimizing the exponential or logistic loss over $\mathcal{D}$, using gradient flow. The training runs for a sufficiently long time, $t \to \infty$.

**Learning from Adversarial Perturbations.** In the following, we remove the restriction of $y_n^{\text{adv}} \neq y_n$ from Definition 1.1 to consider a wider variety of cases.

### 3.2 Decision Boundary of One-Hidden-Layer Neural Network

To understand learning from perturbations, we employ the following result on the implicit bias of gradient flow [6] (similar results are shown in [12]).

**Theorem 3.2** (Rearranged from [6])**.** *Let* $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \{\pm 1\}$ *be a training dataset. Let* $R_{\max} := \max_n \|\boldsymbol{x}_n\|$, $R_{\min} := \min_n \|\boldsymbol{x}_n\|$, *and* $p_{\max} := \max_{n \neq k} |\langle \boldsymbol{x}_n, \boldsymbol{x}_k \rangle|$. *Assume* $\gamma^3 R_{\min}^4 / (3N R_{\max}^2) \geq p_{\max}$. *A one-hidden-layer neural network* $f : \mathbb{R}^d \to \mathbb{R}$ *is trained on the dataset following Setting 3.1. Then, as* $t \to \infty$, $\operatorname{sgn}(f(\boldsymbol{z})) = \operatorname{sgn}(f^{\text{bdy}}(\boldsymbol{z}))$ *holds, where* $f^{\text{bdy}}(\boldsymbol{z}) := \sum_{n=1}^N \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle$ *and* $\lambda_n \in \big(\frac{1}{2R_{\max}^2}, \frac{3}{2\gamma^2 R_{\min}^2}\big)$ *for every* $n \in [N]$.

Appendix A provides a more detailed background. This theorem claims that the binary decision from $f(\boldsymbol{z})$ equals that from the linear function $f^{\text{bdy}}(\boldsymbol{z})$; that is, $f(\boldsymbol{z})$ has a linear decision boundary. This theorem only requires training data to be nearly-orthogonal, which is a common property of high-dimensional data. Although this theorem is not directly related to learning from perturbations, we utilize it to easily observe the decision boundary derived from perturbation learning as follows:

**Corollary 3.3** (Learning from adversarial perturbations)**.** *Let* $\{(\boldsymbol{x}_n^{\text{adv}}, y_n^{\text{adv}})\}_{n=1}^{N^{\text{adv}}}$ *be a training dataset with adversarial perturbations (cf. Definition 1.1). Let* $R_{\max}^{\text{adv}} := \max_n \|\boldsymbol{x}_n^{\text{adv}}\|$, $R_{\min}^{\text{adv}} := \min_n \|\boldsymbol{x}_n^{\text{adv}}\|$, $p_{\max}^{\text{adv}} := \max_{n \neq k} |\langle \boldsymbol{x}_n^{\text{adv}}, \boldsymbol{x}_k^{\text{adv}} \rangle|$, *and* $\lambda_n^{\text{adv}} \in \big(\frac{1}{2{R_{\max}^{\text{adv}}}^2}, \frac{3}{2\gamma^2 {R_{\min}^{\text{adv}}}^2}\big)$ *for every* $n \in [N^{\text{adv}}]$. *Then, the orthogonality assumption and decision boundary in Theorem 3.2 are given by* $\gamma^3 {R_{\min}^{\text{adv}}}^4 / (3N^{\text{adv}} {R_{\max}^{\text{adv}}}^2) \geq p_{\max}^{\text{adv}}$ *and* $f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z}) := \sum_{n=1}^{N^{\text{adv}}} \lambda_n^{\text{adv}} y_n^{\text{adv}} \langle \boldsymbol{x}_n^{\text{adv}}, \boldsymbol{z} \rangle$*, respectively.*

## 4 Theoretical Results

**Perturbation Definition.** Recall that a one-hidden-layer network trained on orthogonal data with Setting 3.1 has a linear decision boundary. We focus on adversarial attacks to this boundary rather than the network itself, called geometry-inspired attacks. Let $\epsilon > 0$ be the perturbation constraint. A geometry-inspired adversarial example $\boldsymbol{x}_n^{\text{adv}}$ maximizes $y_n^{\text{adv}} f^{\text{bdy}}(\boldsymbol{x}_n^{\text{adv}})$ under $\|\boldsymbol{x}_n^{\text{adv}} - \boldsymbol{x}_n\| \leq \epsilon$ as:

$$\boldsymbol{x}_n^{\text{adv}} := \boldsymbol{x}_n + \boldsymbol{\eta}_n, \qquad \boldsymbol{\eta}_n = \epsilon y_n^{\text{adv}} \frac{\boldsymbol{\nabla}_{\boldsymbol{x}_n} f^{\text{bdy}}(\boldsymbol{x}_n)}{\|\boldsymbol{\nabla}_{\boldsymbol{x}_n} f^{\text{bdy}}(\boldsymbol{x}_n)\|} = \epsilon y_n^{\text{adv}} \frac{\sum_{k=1}^N \lambda_k y_k \boldsymbol{x}_k}{\|\sum_{k=1}^N \lambda_k y_k \boldsymbol{x}_k\|}. \tag{1}$$

We observe that the perturbation $\boldsymbol{\eta}_n$ is expressed as a weighted sum of the training samples. Because the training samples $\{\boldsymbol{x}_n\}_{n=1}^N$ are nearly-orthogonal, and $\boldsymbol{x}_n$ and $\boldsymbol{x}_k$ do not negate each other for $n \neq k$, the perturbation contains rich training data information.

**Decision Boundary.** Using Corollary 3.3, the decision boundary can be derived as follows. Appendix B provides the proofs of the theorems.

**Theorem 4.1** (Decision boundary by learning from geometry-inspired perturbations on natural data)**.** *Let $f$ be a one-hidden-layer neural network trained on geometry-inspired perturbations on natural data (cf. Eq. (1) and Definition 1.1) with Setting 3.1. Assume sufficiently large $N$. If*

$$\frac{\gamma^3(R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)^2}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} \tag{2}$$

*with $C := \frac{3R_{\max}^4 + \gamma^3 R_{\min}^4}{\gamma^2 R_{\min}^3 \sqrt{1-\gamma}}$, then, as $t \to \infty$, the decision boundary of $f$ is given by*

$$f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z}) := \underbrace{\frac{\sum_{n=1}^N \lambda_n^{\text{adv}} y_n^{\text{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle}{\sum_{n=1}^N \lambda_n^{\text{adv}}}}_{\text{Effect of learning from mislabeled natural data}} + \underbrace{\epsilon \frac{f^{\text{bdy}}(\boldsymbol{z})}{\| \sum_{n=1}^N \lambda_n y_n \boldsymbol{x}_n \|}}_{\text{Effect of learning from perturbations}} . \tag{3}$$

The decision boundary, Eq. (3), includes two components that explain the effects of mislabeled data and geometry-inspired perturbations. The sign of the first term is determined by the sum of the weighted inner products, $\sum_{n=1}^N \lambda_n^{\text{adv}} y_n^{\text{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle$. Because $y_n^{\text{adv}}$ is mislabeled, the sign (binary decision) of the first term is not always consistent with human perception. The sign of the second term depends only on that of the standard decision boundary $f^{\text{bdy}}(\boldsymbol{z})$. When the magnitude of the second term is dominant, then $\text{sgn}(f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z}))$ matches $\text{sgn}(f^{\text{bdy}}(\boldsymbol{z}))$. This suggests that, although the dataset appears mislabeled to humans, the classifier can still provide a reasonable prediction. A more general version of Theorem 4.1 without assuming large $N$ is given in Theorem B.2. The assumption, Ineq. (2), requires orthogonal training data and $\epsilon = \mathcal{O}(\sqrt{d/N})$.

**Random Label Learning.** Let us consider the limiting behavior of the first and second terms, denoted as $T_1(\boldsymbol{z})$ and $T_2(\boldsymbol{z})$, when $y_n^{\text{adv}}$ is randomly sampled from $\{\pm 1\}$.

**Theorem 4.2** (Consistent decision of learning from geometry-inspired perturbations on natural data)**.** *Suppose that Ineq. (2), $\|\boldsymbol{x}_n\| = \Theta(\sqrt{d})$ for all $n \in [N]$, and $\|\boldsymbol{z}\| = \Theta(\sqrt{d})$ hold. Consider $N \to \infty$ and $d \to \infty$ while keeping $d/N = \Theta(1)$. Suppose that $y_n^{\text{adv}}$ is randomly sampled from $\{\pm 1\}$ for each $n \in [N]$. Assume $|\sum_{n=1}^N \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle| = \Theta(g(N,d))$ if $\sum_{n=1}^N \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z} \rangle| = \Theta(g(N,d))$, where $g$ is a positive function of $N$ and $d$. Let $Q \subset [N]$ be a set of indices such that $|Q| = \Theta(1)$, and let $\boldsymbol{r} \in \mathbb{R}^d$ be a vector such that $\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{r} \rangle| = \mathcal{O}(d)$. If $\boldsymbol{z}$ is **not** represented as $\boldsymbol{z} = \sum_{n \in Q} \pm\Theta(1)\boldsymbol{x}_n + \boldsymbol{r}$, then $\text{sgn}(f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z})) = \text{sgn}(f^{\text{bdy}}(\boldsymbol{z}))$ holds with probability at least 99.99%.*

Given labels $y_n$ freely selected from $\{\pm 1\}$, estimating the growth rate for $T_2(\boldsymbol{z})$ is challenging. Therefore, we assume that $|\sum_{n=1}^N \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle| = \Theta(g(N,d))$ if $\sum_{n=1}^N \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z} \rangle| = \Theta(g(N,d))$ and instead estimate the growth rate of $\sum_{n=1}^N \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z} \rangle|$ rather than $|\sum_{n=1}^N \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle|$. Notably, this theorem suggests that classifiers trained on an apparently mislabeled dataset can produce decisions consistent with standard classifiers, except for a specific $\boldsymbol{z} = \sum_{n \in Q} \pm\Theta(1)\boldsymbol{x}_n + \boldsymbol{r}$. Such $\boldsymbol{z}$ could be, for example, $\boldsymbol{z} = \boldsymbol{x}_1$, $\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3$, and $\boldsymbol{x}_1 + \mathcal{O}(1/N)\mathbf{1}$, where $\mathbf{1}$ denotes an all-ones vector. The first term represents a strong correlation with a few samples. Note that a strong correlation with many samples is invalid because of the orthogonality of $\{\boldsymbol{x}_n\}_{n=1}^N$ and $\|\boldsymbol{z}\| = \Theta(\sqrt{d})$ (cf. Lemma B.4). The second term represents a small vector pointing towards the neighborhood of the first term. For such $\boldsymbol{z}$, the impact of learning from mislabeled samples, $T_1(\boldsymbol{z})$, becomes dominant, and the decisions are not always aligned. Essentially, for inputs that do not strongly correlate with a few training samples, the network decisions derived from perturbation learning align with those of a standard network. Because test datasets typically exclude samples similar to the training data, network learning from perturbations is expected to produce reasonable predictions for many test samples. This confirms the high test accuracy of perturbation learning [8].

## 5 Conclusion

We provided the first theoretical justification of learning from adversarial perturbations for one-hidden-layer networks, assuming orthogonal training data. We showed that networks learning from perturbations produce decisions consistent with normally trained networks, except for specific inputs.

## References

[1] P. Benz, C. Zhang, and I. S. Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *ICCV*, pages 7818–7827, 2021.

[2] S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn. Adversarial examples from computational constraints. In *ICML*, pages 831–840, 2019.

[3] L. Deng. The MNIST database of handwritten digit images for machine learning research. *Signal Process. Mag.*, 29(6):141–142, 2012.

[4] L. Engstrom, J. Gilmer, G. Goh, D. Hendrycks, A. Ilyas, A. Madry, R. Nakano, P. Nakkiran, S. Santurkar, B. Tran, D. Tsipras, and E. Wallace. A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, 2019. https://distill.pub/2019/advex-bugs-discussion.

[5] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In *NeurIPS*, volume 31, 2018.

[6] S. Frei, G. Vardi, P. L. Bartlett, N. Srebro, and W. Hu. Implicit bias in leaky relu networks trained on high-dimensional data. In *ICLR*, 2023.

[7] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. Adversarial spheres. In *ICLR WS*, 2018.

[8] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, pages 125–136, 2019.

[9] J. Kim, B.-K. Lee, and Y. M. Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. In *NeurIPS*, volume 34, pages 17148–17159, 2021.

[10] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[12] R. Sarussi, A. Brutzkus, and A. Globerson. Towards understanding learning in neural networks with linear teachers. In *ICML*, pages 9313–9322, 2021.

[13] J. Springer, M. Mitchell, and G. Kenyon. A little robustness goes a long way: Leveraging robust features for targeted transfer attacks. In *NeurIPS*, volume 34, pages 9759–9773, 2021.

[14] J. M. Springer, M. Mitchell, and G. T. Kenyon. Adversarial perturbations are not so weird: Entanglement of robust and non-robust features in neural network classifiers. *arXiv:2102.05110*, 2021.

[15] J. M. Springer, M. Mitchell, and G. T. Kenyon. Uncovering universal features: How adversarial training improves adversarial transferability. In *ICML WS*, 2021.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[17] L. Tao, L. Feng, H. Wei, J. Yi, S.-J. Huang, and S. Chen. Can adversarial training be manipulated by non-robust features? In *NeurIPS*, volume 35, pages 26504–26518, 2022.

[18] N. Tsilivis and J. Kempe. What can the neural tangent kernel tell us about adversarial robustness? In *NeurIPS*, volume 35, pages 18116–18130, 2022.

[19] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

# A Background

To formulate learning from adversarial perturbations, we utilize the theorem presented in [6], which addresses the implicit bias of one-hidden-layer neural networks under gradient flow with an exponential loss. Note that this theorem does not directly pertain to adversarial attacks, adversarial examples, or learning from perturbations. We leverage this because of the tractable form of a decision boundary. The key findings of their study are summarized as follows:

**Theorem A.1** (Rearranged from [6])**.** *Let* $\mathcal{D} := \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N} \subset \mathbb{R}^d \times \{\pm 1\}$ *be a training dataset. Let* $R_{\max} := \max_n \|\boldsymbol{x}_n\|$, $R_{\min} := \min_n \|\boldsymbol{x}_n\|$, *and* $p_{\max} := \max_{n \neq k} |\langle \boldsymbol{x}_n, \boldsymbol{x}_k \rangle|$. *Assume* $\gamma^3 R_{\min}^4 / (3 N R_{\max}^2) \geq p_{\max}$. *A one-hidden neural network* $f : \mathbb{R}^d \to \mathbb{R}$ *is trained on* $\mathcal{D}$ *following Setting 3.1. Then, gradient flow on* $f$ *converges to* $\lim_{t \to \infty} \frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|_F} = \frac{\boldsymbol{W}^{\mathrm{std}}}{\|\boldsymbol{W}^{\mathrm{std}}\|_F}$, *where* $\boldsymbol{W}^{\mathrm{std}} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m/2}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{m/2})^\top$ *satisfies*

$$\forall n \in [N] : y_n f(\boldsymbol{x}_n; \boldsymbol{W}^{\mathrm{std}}) = 1, \tag{4}$$

$$\boldsymbol{v}_1 = \cdots = \boldsymbol{v}_{m/2} = \boldsymbol{v} := \frac{1}{\sqrt{m}} \sum_{n : y_n = +1} \lambda_n \boldsymbol{x}_n - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n = -1} \lambda_n \boldsymbol{x}_n, \tag{5}$$

$$\boldsymbol{u}_1 = \cdots = \boldsymbol{u}_{m/2} = \boldsymbol{u} := \frac{1}{\sqrt{m}} \sum_{n : y_n = -1} \lambda_n \boldsymbol{x}_n - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n = +1} \lambda_n \boldsymbol{x}_n, \tag{6}$$

*where* $\lambda_n \in \left( \frac{1}{2R_{\max}^2}, \frac{3}{2\gamma^2 R_{\min}^2} \right)$ *for every* $n \in [N]$. *The binary decision of* $f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{std}})$ *is also given by:*

$$\operatorname{sgn}\left(f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{std}})\right) = \operatorname{sgn}\left(f^{\mathrm{bdy}}(\boldsymbol{z})\right), \qquad \text{where} \qquad f^{\mathrm{bdy}}(\boldsymbol{z}) := \sum_{n=1}^{N} \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle. \tag{7}$$

The theorem provides three insights: (i) Although there might be many possible directions $\boldsymbol{W}/\|\boldsymbol{W}\|_F$ that can accurately classify the training dataset, gradient flow consistently converges in direction to $\boldsymbol{W}^{\mathrm{std}}$ irrespective of the initial weight configurations. (ii) Given that $\boldsymbol{W}^{\mathrm{std}}$ is composed of a maximum of two unique row vectors, its rank is constrained to two or less, highlighting the implicit bias of the gradient flow. (iii) The binary decision of $f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{std}})$ is the same as the sign of the linear function $f^{\mathrm{bdy}}(\boldsymbol{z})$, indicating that $f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{std}})$ has a linear decision boundary. The assumption of the theorem requires nearly orthogonal data, which is a typical characteristic of high-dimensional data.

Note that in [6], the binary decision boundary is given by:

$$f^{\mathrm{bdy}}(\boldsymbol{z}) = \frac{\sqrt{m}}{2} \boldsymbol{v} - \frac{\sqrt{m}}{2} \boldsymbol{u}. \tag{8}$$

To derive Eq. (7), we rearrange the above equation as:

$$\begin{aligned}
f^{\mathrm{bdy}}(\boldsymbol{z}) =& \frac{\sqrt{m}}{2} \left( \frac{1}{\sqrt{m}} \sum_{n : y_n = +1} \lambda_n \boldsymbol{x}_n - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n = -1} \lambda_n \boldsymbol{x}_n \right) && (9) \\
& - \frac{\sqrt{m}}{2} \left( \frac{1}{\sqrt{m}} \sum_{n : y_n = -1} \lambda_n \boldsymbol{x}_n - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n = +1} \lambda_n \boldsymbol{x}_n \right) && (10) \\
=& \frac{1 + \gamma}{2} \left( \sum_{n : y_n = +1} \lambda_n \boldsymbol{x}_n - \sum_{n : y_n = -1} \lambda_n \boldsymbol{x}_n \right) && (11) \\
=& \frac{1 + \gamma}{2} \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n. && (12)
\end{aligned}$$

Thus,

$$\operatorname{sgn}\left(f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{std}})\right) = \operatorname{sgn}\left(f^{\mathrm{bdy}}(\boldsymbol{z})\right) = \operatorname{sgn}\left( \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \right). \tag{13}$$

Note that this theorem does not impose any assumptions on the training data other than orthogonality. Thus, it can be adapted to a dataset with adversarial perturbations as follows:

**Corollary A.2** (Learning from adversarial perturbations)**.** *Let* $\mathcal{D}^{\text{adv}} := \{(\boldsymbol{x}_n^{\text{adv}}, y_n^{\text{adv}})\}_{n=1}^{N^{\text{adv}}} \subset \mathbb{R}^d \times \{\pm 1\}$ *be a training dataset. Let* $R_{\max}^{\text{adv}} := \max_n \|\boldsymbol{x}_n^{\text{adv}}\|$, $R_{\min}^{\text{adv}} := \min_n \|\boldsymbol{x}_n^{\text{adv}}\|$, *and* $p_{\max}^{\text{adv}} := \max_{n \neq k} |\langle \boldsymbol{x}_n^{\text{adv}}, \boldsymbol{x}_k^{\text{adv}} \rangle|$. *Assume* $\gamma^3 {R_{\min}^{\text{adv}}}^4 / (3N {R_{\max}^{\text{adv}}}^2) \geq p_{\max}^{\text{adv}}$. *A one-hidden neural network* $f : \mathbb{R}^d \to \mathbb{R}$ *is trained on the dataset following Setting 3.1. Then, gradient flow on* $f$ *converges to* $\lim_{t \to \infty} \frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|_F} = \frac{\boldsymbol{W}^{\text{adv}}}{\|\boldsymbol{W}^{\text{adv}}\|_F}$, *where* $\boldsymbol{W}^{\text{adv}} := (\boldsymbol{v}_1^{\text{adv}}, \ldots, \boldsymbol{v}_{m/2}^{\text{adv}}, \boldsymbol{u}_1^{\text{adv}}, \ldots, \boldsymbol{u}_{m/2}^{\text{adv}})^\top$ *satisfies*

$$\forall n \in [N] : y_n^{\text{adv}} f(\boldsymbol{x}_n^{\text{adv}}; \boldsymbol{W}^{\text{adv}}) = 1, \tag{14}$$

$$\boldsymbol{v}_1^{\text{adv}} = \cdots = \boldsymbol{v}_{m/2}^{\text{adv}} = \frac{1}{\sqrt{m}} \sum_{n : y_n^{\text{adv}} = +1} \lambda_n^{\text{adv}} \boldsymbol{x}_n^{\text{adv}} - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n^{\text{adv}} = -1} \lambda_n^{\text{adv}} \boldsymbol{x}_n^{\text{adv}}, \tag{15}$$

$$\boldsymbol{u}_1^{\text{adv}} = \cdots = \boldsymbol{u}_{m/2}^{\text{adv}} = \frac{1}{\sqrt{m}} \sum_{n : y_n^{\text{adv}} = -1} \lambda_n^{\text{adv}} \boldsymbol{x}_n^{\text{adv}} - \frac{\gamma}{\sqrt{m}} \sum_{n : y_n^{\text{adv}} = +1} \lambda_n^{\text{adv}} \boldsymbol{x}_n^{\text{adv}}, \tag{16}$$

*where* $\lambda_n^{\text{adv}} \in \left( \frac{1}{2{R_{\max}^{\text{adv}}}^2}, \frac{3}{2\gamma^2 {R_{\min}^{\text{adv}}}^2} \right)$ *for every* $n \in [N]$. *The binary decision of* $f(\boldsymbol{z}; \boldsymbol{W}^{\text{adv}})$ *is also given by:*

$$\operatorname{sgn}\left(f(\boldsymbol{z}; \boldsymbol{W}^{\text{adv}})\right) = \operatorname{sgn}\left(f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z})\right), \quad \text{where} \quad f_{\text{adv}}^{\text{bdy}}(\boldsymbol{z}) := \sum_{n=1}^{N} \lambda_n^{\text{adv}} y_n^{\text{adv}} \langle \boldsymbol{x}_n^{\text{adv}}, \boldsymbol{z} \rangle. \tag{17}$$

The presented theorem establishes the foundation for learning from adversarial perturbations. The orthogonality assumption, model weights, and decision boundary are influenced by the definition of adversarial perturbations.

## B Proofs of Theorems in Section 4

In Lemma B.1, we first restructure the orthogonality condition required by Theorem 3.2 and Corollary 3.3, which is applicable to any dataset configuration. To this end, we leverage the lower and upper bounds of $\lambda_n$ and the upper bound of $p_{\max}$. Then, in Theorem B.2, we formulate learning from geometry-inspired perturbations without assuming the number of training samples $N$. Theorem 4.1 is the special case of Theorem B.2 for a sufficiently large $N$. Lemmas B.3 and B.4 describe the preliminary on limiting behavior. Propositions B.5 and B.6 give the limiting behavior for two cases: when $y_n^{\text{adv}}$ is chosen deterministically or randomly. Finally, Theorem 4.2 is introduced.

We denote the gradient of the decision boundary of the normally trained one-hidden-neural network by $\boldsymbol{q}$; namely,

$$\boldsymbol{q} := \boldsymbol{\nabla}_{\boldsymbol{z}} f^{\text{bdy}}(\boldsymbol{z}) = \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n. \tag{18}$$

Using $\boldsymbol{q}$, we can represent the geometry-inspired adversarial example as:

$$\boldsymbol{x}_n^{\text{adv}} := \boldsymbol{x}_n + \epsilon y_n^{\text{adv}} \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}. \tag{19}$$

**Lemma B.1** (orthogonality condition for learning from geometry-inspired perturbations on natural data)**.** *Consider the geometry-inspired perturbation defined in Eq. (1). Let*

$$C := \frac{3R_{\max}^4 + \gamma^3 R_{\min}^4}{\gamma^2 R_{\min}^3 \sqrt{1 - \gamma}}. \tag{20}$$

*Suppose that the following inequalities are satisfied:*

$$\begin{cases} \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max} + \epsilon)^2} - 2\epsilon R_{\max} - \epsilon^2 \geq p_{\max} & \left(N \leq \frac{C^2}{R_{\max}^2}\right) \\ \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(\frac{C^2}{R_{\max}^2} < N \leq \frac{C^2}{R_{\min}^2}\right) \\ \frac{\gamma^3 (R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)^2}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(N > \frac{C^2}{R_{\min}^2}\right) \end{cases}. \tag{21}$$

*Then, the following inequality holds for any* $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ *and* $\{y_n^{\mathrm{adv}}\}_{n=1}^N$*:*

$$\frac{\gamma^3 {R_{\min}^{\mathrm{adv}}}^4}{3N {R_{\max}^{\mathrm{adv}}}^2} \geq p_{\max}^{\mathrm{adv}}. \tag{22}$$

*Proof.* First, we exclude cases where $\epsilon > R_{\min}$. Then, given $\epsilon \leq R_{\min}$, we establish the three primary inequalities. Finally, we prove that these inequalities intrinsically presuppose $\epsilon \leq R_{\min}$.

**Pruning.** In this lemma, we consider the orthogonality condition valid for any $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $\{y_n^{\mathrm{adv}}\}_{n=1}^N$. Here, we assert that if $\epsilon > R_{\min}$, this condition cannot be sustained.

*(Lower bound of maximum inner product)* Let $n, k$ be different data indices that satisfy $y_n = y_k = y_n^{\mathrm{adv}} = y_k^{\mathrm{adv}}$. Consider the case with $p_{\max} = 0$. For $n$,

$$\operatorname{sgn}\left(\left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle\right) = \operatorname{sgn}\left(\lambda_n y_n \|\boldsymbol{x}_n\|\right) = y_n. \tag{23}$$

The lower bound of the maximum inner product is calculated as:

$$p_{\max}^{\mathrm{adv}} \geq \langle \boldsymbol{x}_n^{\mathrm{adv}}, \boldsymbol{x}_k^{\mathrm{adv}} \rangle = \epsilon y_n^{\mathrm{adv}} \left\langle \boldsymbol{x}_k, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle + \epsilon y_k^{\mathrm{adv}} \left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle + \epsilon^2 y_n^{\mathrm{adv}} y_k^{\mathrm{adv}} \geq \epsilon^2. \tag{24}$$

*(Upper bound of minimum norm)* Let $l := \arg\min_l \|\boldsymbol{x}_l\|$ and $y_l^{\mathrm{adv}} = -\operatorname{sgn}\left(\langle \boldsymbol{x}_l, \boldsymbol{q}/\|\boldsymbol{q}\| \rangle\right)$. Note that this is trivially compatible with the settings of $y_n = y_k = y_n^{\mathrm{adv}} = y_k^{\mathrm{adv}}$ and $p_{\max} = 0$. The upper bound of the minimum norm can be calculated as:

$$\left\|\boldsymbol{x}_l^{\mathrm{adv}}\right\| = \sqrt{R_{\min}^2 - 2\epsilon \left| \left\langle \boldsymbol{x}_l, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle \right| + \epsilon^2} \geq R_{\min}^{\mathrm{adv}}. \tag{25}$$

*(orthogonality condition)* Finally, we rearrange the orthogonality condition using the above two bounds as:

$$\frac{\gamma^3 {R_{\min}^{\mathrm{adv}}}^4}{3N {R_{\max}^{\mathrm{adv}}}^2} - p_{\max}^{\mathrm{adv}} \leq \frac{{R_{\min}^{\mathrm{adv}}}^2}{3} - p_{\max}^{\mathrm{adv}} \leq \frac{R_{\min}^2 + \epsilon^2}{3} - \epsilon^2 < -\frac{\epsilon^2}{3} < 0. \tag{26}$$

This inequality indicates that the orthogonality condition that holds for any $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $\{y_n^{\mathrm{adv}}\}_{n=1}^N$ does not exist.

**Main proof.** Assume $\epsilon \leq R_{\min}$. We define the lower and upper bounds of $\lambda_n$ as $\lambda_{\min} := \frac{1}{2R_{\max}^2}$ and $\lambda_{\max} := \frac{3}{2\gamma^2 R_{\min}^2}$, respectively.

*(Preliminary)* The lower bound of the norm of $\boldsymbol{q}$ is derived as:

$$\begin{aligned} \|\boldsymbol{q}\| =& \sqrt{\sum_{n=1}^N \lambda_n \left(\lambda_n \|\boldsymbol{x}_n\|^2 + \sum_{k \neq n} \lambda_k y_n y_k \langle \boldsymbol{x}_n, \boldsymbol{x}_k \rangle\right)} && (27) \\ \geq& \sqrt{N \lambda_{\min}(\lambda_{\min} R_{\min}^2 - N \lambda_{\max} p_{\max})} && (28) \\ =& \frac{R_{\min}\sqrt{(1-\gamma)N}}{2R_{\max}^2}. && (29) \end{aligned}$$

The upper bound of the inner product between $\boldsymbol{x}_n$ and $\boldsymbol{q}$ is derived as:

$$\langle \boldsymbol{x}_n, \boldsymbol{q} \rangle = \sum_{k=1}^{N} \lambda_k y_k \langle \boldsymbol{x}_n, \boldsymbol{x}_k \rangle \leq \lambda_{\max}(R_{\max}^2 + N p_{\max}) = \frac{3R_{\max}^2}{2\gamma^2 R_{\min}^2} + \frac{\gamma R_{\min}^2}{2R_{\max}^2}. \tag{30}$$

The naive upper bound of the inner product between $\boldsymbol{x}_n$ and $\boldsymbol{q}/\|\boldsymbol{q}\|$ is given by:

$$\left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle \leq R_{\max}. \tag{31}$$

Alternatively, that can be also obtained as:

$$\left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle \leq \frac{3R_{\max}^4 + \gamma^3 R_{\min}^4}{\gamma^2 R_{\min}^3 \sqrt{(1-\gamma)N}} =: \frac{C}{\sqrt{N}}. \tag{32}$$

Note that

$$\begin{cases} \frac{C}{\sqrt{N}} \geq R_{\max} & \left(N \leq \frac{C^2}{R_{\max}^2}\right) \\ R_{\min} \leq \frac{C}{\sqrt{N}} < R_{\max} & \left(\frac{C^2}{R_{\max}^2} < N \leq \frac{C^2}{R_{\min}^2}\right) \\ \frac{C}{\sqrt{N}} < R_{\min} & \left(N > \frac{C^2}{R_{\min}^2}\right) \end{cases}. \tag{33}$$

Thus,

$$\left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle \leq \begin{cases} R_{\max} & \left(N \leq \frac{C^2}{R_{\max}^2}\right) \\ \frac{C}{\sqrt{N}} & \text{(otherwise)} \end{cases}. \tag{34}$$

*(Lower and upper bounds of norm)* The norm of the geometry-inspired adversarial example can be represented as:

$$\left\|\boldsymbol{x}_n^{\text{adv}}\right\| = \sqrt{\|\boldsymbol{x}_n\|^2 + 2\epsilon y_n^{\text{adv}} \left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle + \epsilon^2}. \tag{35}$$

Under $\epsilon \leq R_{\min}$, the trivial lower and upper bounds of the above norm is written as:

$$R_{\min} - \epsilon \leq \left\|\boldsymbol{x}_n^{\text{adv}}\right\| \leq R_{\max} + \epsilon. \tag{36}$$

Now, we have the following three lower bounds of the norm of $\boldsymbol{x}_n$: (i) $\sqrt{R_{\min}^2 - 2\epsilon R_{\max} + \epsilon^2}$ for $N \leq \frac{C^2}{R_{\max}^2}$. (ii) $\sqrt{R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2}$ for $N > \frac{C^2}{R_{\max}^2}$. (iii) $R_{\min} - \epsilon$ for $\epsilon \leq R_{\min}$. Since $(R_{\min} - \epsilon)^2 - (R_{\min}^2 - 2\epsilon R_{\max} + \epsilon^2) \geq 0$, (iii) is always tighter than (i). In addition, since $(R_{\min} - \epsilon)^2 - (R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2) \geq 0$ under $\frac{C^2}{R_{\max}^2} < N \leq \frac{C^2}{R_{\min}^2}$, (iii) is always tighter than (ii). Thus, under $\epsilon \leq R_{\min}$,

$$\left\|\boldsymbol{x}_n^{\text{adv}}\right\| \geq \begin{cases} R_{\min} - \epsilon & (N \leq \frac{C^2}{R_{\min}^2}) \\ \sqrt{R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2} & \text{(otherwise)} \end{cases}. \tag{37}$$

The upper bound of the norm is given by:

$$\left\|\boldsymbol{x}_n^{\text{adv}}\right\| \leq \begin{cases} R_{\max} + \epsilon & (N \leq \frac{C^2}{R_{\max}^2}) \\ \sqrt{R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2} & \text{(otherwise)} \end{cases}. \tag{38}$$

*(Upper bound of inner product)* The upper bound of the inner product between $\boldsymbol{x}_n^{\text{adv}}$ and $\boldsymbol{x}_k^{\text{adv}}$ for $n \neq k$ is represented as:

$$\langle \boldsymbol{x}_n^{\text{adv}}, \boldsymbol{x}_k^{\text{adv}} \rangle \leq p_{\max} + 2\epsilon \left\langle \boldsymbol{x}_n, \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\rangle + \epsilon^2. \tag{39}$$

Thus,

$$\langle \boldsymbol{x}_n^{\mathrm{adv}}, \boldsymbol{x}_k^{\mathrm{adv}} \rangle \leq \begin{cases} p_{\max} + 2\epsilon R_{\max} + \epsilon^2 & (N \leq \frac{C^2}{R_{\max}^2}) \\ p_{\max} + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2 & (\text{otherwise}) \end{cases}. \tag{40}$$

*(orthogonality condition)* Using above bounds, we can derive the orthogonality condition that holds for any $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $\{y_n^{\mathrm{adv}}\}_{n=1}^N$ as:

$$\begin{cases} \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max}+\epsilon)^2} - 2\epsilon R_{\max} - \epsilon^2 \geq p_{\max} & \left(N \leq \frac{C^2}{R_{\max}^2}\right) \\ \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(\frac{C^2}{R_{\max}^2} < N \leq \frac{C^2}{R_{\min}^2}\right) \\ \frac{\gamma^3 (R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)^2}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(N > \frac{C^2}{R_{\min}^2}\right) \end{cases}. \tag{41}$$

**Compatibility with pruned condition.** Here, we prove that the above inequalities implicitly suggest $\epsilon \leq R_{\min}$; that is, they do not hold under $\epsilon > R_{\min}$. The common upper bound of the left term of the inequalities is

$$\frac{\gamma^3 (R_{\min}^2 + \epsilon^2)}{3N} - \epsilon^2. \tag{42}$$

This bound monotonically decreases with $\epsilon$ and is under zero at $\epsilon = R_{\min}$. Thus, the inequalities are not satisfied under $\epsilon > R_{\min}$ □

**Theorem B.2** (Decision boundary by learning from geometry-inspired perturbations on natural data)**.** *Let $f$ be a one-hidden-layer neural network trained on geometry-inspired perturbations on natural data (cf. Eq. (1)) with Setting 3.1. If*

$$\begin{cases} \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max}+\epsilon)^2} - 2\epsilon R_{\max} - \epsilon^2 \geq p_{\max} & \left(N \leq \frac{C^2}{R_{\max}^2}\right) \\ \frac{\gamma^3 (R_{\min} - \epsilon)^4}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(\frac{C^2}{R_{\max}^2} < N \leq \frac{C^2}{R_{\min}^2}\right) \\ \frac{\gamma^3 (R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)^2}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} & \left(N > \frac{C^2}{R_{\min}^2}\right) \end{cases}. \tag{43}$$

*with $C := \frac{3R_{\max}^4 + \gamma^3 R_{\min}^4}{\gamma^2 R_{\min}^3 \sqrt{1-\gamma}}$, then, as $t \to \infty$, the decision boundary of $f$ is given by*

$$f_{\mathrm{adv}}^{\mathrm{bdy}}(\boldsymbol{z}) := \underbrace{\frac{\sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle}{\sum_{n=1}^N \lambda_n^{\mathrm{adv}}}}_{\text{Learning from mislabeled natural data}} + \underbrace{\epsilon \frac{f^{\mathrm{bdy}}(\boldsymbol{z})}{\| \sum_{n=1}^N \lambda_n y_n \boldsymbol{x}_n \|}}_{\text{Learning from perturbations}}. \tag{44}$$

*Proof.* By Lemma B.1, if Ineq. (43) holds, then $\gamma^3 {R_{\min}^{\mathrm{adv}}}^4 / (3N {R_{\max}^{\mathrm{adv}}}^2) \geq p_{\max}^{\mathrm{adv}}$ holds for any dataset configuration, i.e., for any $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $\{y_n^{\mathrm{adv}}\}_{n=1}^N$. Thus, we can apply Corollary 3.3 to this dataset. By Corollary 3.3, the decision boundary is given by:

$$\begin{aligned} \operatorname{sgn}(f(\boldsymbol{z}; \boldsymbol{W}^{\mathrm{adv}})) &= \operatorname{sgn}\left( \sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n^{\mathrm{adv}}, \boldsymbol{z} \rangle \right) && (45) \\ &= \operatorname{sgn}\left( \sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle + \sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \epsilon y_n^{\mathrm{adv}} \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}, \boldsymbol{z} \right\rangle \right) && (46) \\ &= \operatorname{sgn}\left( \sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle + \left( \sum_{n=1}^N \lambda_n^{\mathrm{adv}} \right) \epsilon \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}, \boldsymbol{z} \right\rangle \right) && (47) \\ &= \operatorname{sgn}\left( \frac{\sum_{n=1}^N \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle}{\sum_{n=1}^N \lambda_n^{\mathrm{adv}}} + \epsilon \frac{\langle \boldsymbol{q}, \boldsymbol{z} \rangle}{\|\boldsymbol{q}\|} \right) && (48) \end{aligned}$$

$$= \operatorname{sgn}\left( \frac{\sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle}{\sum_{n=1}^{N} \lambda_n^{\mathrm{adv}}} + \epsilon \frac{f^{\mathrm{bdy}}(\boldsymbol{z})}{\|\boldsymbol{q}\|} \right). \tag{49}$$

□

**Theorem 4.1** (Decision boundary by learning from geometry-inspired perturbations on natural data)**.** *Let $f$ be a one-hidden-layer neural network trained on geometry-inspired perturbations on natural data (cf. Eq. (1) and Definition 1.1) with Setting 3.1. Assume sufficiently large $N$. If*

$$\frac{\gamma^3 (R_{\min}^2 - 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)^2}{3N(R_{\max}^2 + 2\frac{C}{\sqrt{N}}\epsilon + \epsilon^2)} - 2\frac{C}{\sqrt{N}}\epsilon - \epsilon^2 \geq p_{\max} \tag{2}$$

*with $C := \frac{3R_{\max}^4 + \gamma^3 R_{\min}^4}{\gamma^2 R_{\min}^3 \sqrt{1-\gamma}}$, then, as $t \to \infty$, the decision boundary of $f$ is given by*

$$f_{\mathrm{adv}}^{\mathrm{bdy}}(\boldsymbol{z}) := \underbrace{\frac{\sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z} \rangle}{\sum_{n=1}^{N} \lambda_n^{\mathrm{adv}}}}_{\text{Effect of learning from mislabeled natural data}} + \underbrace{\epsilon \frac{f^{\mathrm{bdy}}(\boldsymbol{z})}{\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \|}}_{\text{Effect of learning from perturbations}} . \tag{3}$$

*Proof.* This is the special case of Theorem B.2 when the number of training samples $N$ is sufficiently large. □

**Lemma B.3** (Order of norm of weighted sum of training data)**.** *Assume $\gamma^3 R_{\min}^4 / (3NR_{\max}^2) \geq p_{\max}$ and $\|\boldsymbol{x}_n\| = \Theta(\sqrt{d})$ for any $n \in [N]$. Then,*

$$\left\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \right\| = \Theta\left( \sqrt{\frac{N}{d}} \right). \tag{50}$$

*Proof.* By definition in Theorem 3.2, $\lambda_n = \Theta(1/d)$. By the assumption, $p_{\max} = \mathcal{O}(d/N)$. The lower bound is derived as:

$$\left\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \right\| = \sqrt{\sum_{n=1}^{N} \lambda_n \left( \lambda_n \|\boldsymbol{x}_n\|^2 + \sum_{k \neq n} \lambda_k y_n y_k \langle \boldsymbol{x}_n, \boldsymbol{x}_k \rangle \right)} \tag{51}$$

$$\geq \sqrt{\sum_{n=1}^{N} \lambda_n \left( \lambda_n \|\boldsymbol{x}_n\|^2 - \sum_{k \neq n} \lambda_k p_{\max} \right)} \tag{52}$$

$$= \Omega\left( \sqrt{\frac{N}{d}} \right). \tag{53}$$

The upper bound is derived as:

$$\left\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \right\| \leq \sqrt{\sum_{n=1}^{N} \lambda_n \left( \lambda_n \|\boldsymbol{x}_n\|^2 + \sum_{k \neq n} \lambda_k p_{\max} \right)} = \mathcal{O}\left( \sqrt{\frac{N}{d}} \right). \tag{54}$$

Note that the radicand of the lower bound is positive since the following inequality holds:

$$\lambda_n \|\boldsymbol{x}_n\|^2 - \sum_{k \neq n} \lambda_k p_{\max} \geq \frac{R_{\min}^2}{2R_{\max}^2} - \frac{\gamma R_{\min}^2}{2R_{\max}^2} \geq \frac{(1-\gamma) R_{\min}^2}{2R_{\max}^2} > 0. \tag{55}$$

Thus,

$$\left\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \right\| = \Theta\left( \sqrt{\frac{N}{d}} \right). \tag{56}$$

□

**Lemma B.4** (Order of inner product)**.** *Let $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{R}^d$ and $\boldsymbol{z} \in \mathbb{R}^d$ be the $d$-dimensional data. For any $n, k \in [N], k \neq n$, assume $\|\boldsymbol{x}_n\| = \Theta(\sqrt{d})$, $\|\boldsymbol{z}\| = \Theta(\sqrt{d})$, and $|\langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle| = \mathcal{O}(d/N)$. Let $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(d^{\beta^{\max}})$. Then, the following statements hold:*

*(a)*

$$\left|\sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\right| \leq \sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(\sqrt{N} d^{\beta^{\max}}). \tag{57}$$

*(b)*

$$\sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 = \mathcal{O}(d^{2\beta^{\max}}). \tag{58}$$

*(c) There are at most $\mathcal{O}(\min(N^{-2\alpha}, N))$ instances of $n$ that satisfy $\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle = \pm\Theta(N^\alpha d^\beta)$, where $\alpha \leq 0$ and $\beta \leq 1$. There is no $n$ that satisfies $\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle = \pm\Omega(N^\alpha d^\beta)$, where $\alpha > 0$ or $\beta > 1$.*

*(d) (i) The growth rate of $\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle|$ is faster than or equal to $\Theta(1/d) \sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2$. (ii) The growth rate of $\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle|$ is equal to $\Theta(1/d) \sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2$ if and only if there exists $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(1)$ and $\sum_{n:|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| \neq \Theta(1)} |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(1)$ with respect to $N$.*

*Proof.* For $n \in [N]$, let $\psi_n := \operatorname{sgn}(\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle)$.

**(a)** The left inequality is trivial. By the Cauchy–Schwarz inequality,

$$\begin{aligned}\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| =& \sum_{n=1}^N \langle \psi_n \boldsymbol{x}_n, \boldsymbol{z}\rangle & (59)\\ \leq& \left\|\sum_{n=1}^N \psi_n \boldsymbol{x}_n\right\| \|\boldsymbol{z}\| & (60)\\ =& \sqrt{\sum_{n=1}^N \|\boldsymbol{x}_n\|^2 + \sum_{n=1}^N \sum_{k\neq n} \psi_n \psi_k \langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle} \|\boldsymbol{z}\| & (61)\\ =& \mathcal{O}(\sqrt{N} d). & (62)\end{aligned}$$

If $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(d^{\beta^{\max}})$, then $\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle|$ is trivially constrained to $\mathcal{O}(\sqrt{N} d^{\beta^{\max}})$.

**(b)** By the Cauchy–Schwarz inequality,

$$\begin{aligned}\sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 =& \sum_{n=1}^N \langle \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \boldsymbol{x}_n, \boldsymbol{z}\rangle & (63)\\ =& \left\langle \sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \boldsymbol{x}_n, \boldsymbol{z} \right\rangle & (64)\\ \leq& \sqrt{\sum_{n=1}^N \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 \|\boldsymbol{x}_n\|^2 + \sum_{n=1}^N \sum_{k\neq n} \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \langle \boldsymbol{x}_k, \boldsymbol{z}\rangle \langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle} \|\boldsymbol{z}\|. & (65)\end{aligned}$$

Now,

$$\begin{aligned}\sum_{n=1}^N \sum_{k\neq n} \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \langle \boldsymbol{x}_k, \boldsymbol{z}\rangle \langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle \leq& \sum_{n=1}^N \sum_{k=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| |\langle \boldsymbol{x}_k, \boldsymbol{z}\rangle| |\langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle| & (66)\\ =& \mathcal{O}\left(\frac{d}{N}\right) \left(\sum_{n=1}^N |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle|\right)^2. & (67)\end{aligned}$$

By (a),

$$\sum_{n=1}^{N}\sum_{k\neq n}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\langle \boldsymbol{x}_k, \boldsymbol{z}\rangle\langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle = \mathcal{O}(d^3). \tag{68}$$

Thus,

$$\sum_{n=1}^{N}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 = \sqrt{\Theta(d^2)\sum_{n=1}^{N}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 + \mathcal{O}(d^4)}. \tag{69}$$

Let $\sum_{n=1}^{N}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 = \mathcal{O}(N^{\zeta}d^2)$ for a constant $\zeta \in \mathbb{R}^d$. Using this,

$$\underbrace{\sum_{n=1}^{N}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2}_{\mathcal{O}(N^{\zeta}d^2)} = \mathcal{O}(\max(N^{\zeta/2}d^2, d^2)). \tag{70}$$

If $\zeta > 0$, the left term grows faster than the right term, which contradicts the equation. Thus, $\zeta \leq 0$. If $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(d^{\beta^{\max}})$, then $\sum_{n=1}^{N}\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2$ is trivially constrained to $\mathcal{O}(d^{2\beta^{\max}})$. Thus, the claim is established.

**(c)** By the Cauchy–Schwarz inequality, $\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \leq \|\boldsymbol{x}_n\|\|\boldsymbol{z}\| = \mathcal{O}(d)$. Let $[N]_{\alpha,\beta} := \{n : |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^{\alpha}d^{\beta})\}$, where $\alpha \leq 0$ and $\beta \leq 1$. We define $\delta \leq 1$ to satisfy $|[N]_{\alpha,\beta}| = \Theta(N^{\delta})$. Then,

$$\sum_{n\in[N]_{\alpha,\beta}}\langle \psi_n\boldsymbol{x}_n, \boldsymbol{z}\rangle = \sum_{n\in[N]_{\alpha,\beta}}\Theta(N^{\alpha}d^{\beta}) = \Theta(N^{\alpha+\delta}d^{\beta}). \tag{71}$$

By the Cauchy–Schwarz inequality,

$$\sum_{n\in[N]_{\alpha,\beta}}\langle \psi_n\boldsymbol{x}_n, \boldsymbol{z}\rangle = \left\langle \sum_{n\in[N]_{\alpha,\beta}}\psi_n\boldsymbol{x}_n, \boldsymbol{z}\right\rangle \leq \left\|\sum_{n\in[N]_{\alpha,\beta}}\psi_n\boldsymbol{x}_n\right\|\|\boldsymbol{z}\|. \tag{72}$$

Note that

$$\begin{aligned}\left\|\sum_{n\in[N]_{\alpha,\beta}}\psi_n\boldsymbol{x}_n\right\| &= \sqrt{\sum_{n\in[N]_{\alpha,\beta}}\|\boldsymbol{x}_n\|^2 + \sum_{n\in[N]_{\alpha,\beta}}\sum_{k\neq n}\psi_n\psi_k\langle \boldsymbol{x}_n, \boldsymbol{x}_k\rangle} &(73)\\ &= \sqrt{\Theta(N^{\delta}d) \pm \Theta(N^{2\delta})\mathcal{O}\left(\frac{d}{N}\right)} &(74)\\ &= \Theta(\sqrt{N^{\delta}d}). &(75)\end{aligned}$$

Thus, $\sum_{n\in[N]_{\alpha,\beta}}\langle \psi_n\boldsymbol{x}_n, \boldsymbol{z}\rangle = \mathcal{O}(\sqrt{N^{\delta}}d)$. Comparing this with Eq. (71), $\alpha + \delta \leq \delta/2 \Leftrightarrow \delta \leq -2\alpha$.

**(d)** Since $\Theta(1/d)\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle^2 = \Theta(N^{2\alpha}d^{2\beta-1})$ if $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^{\alpha}d^{\beta})$, and $\alpha \leq 0$ and $\beta \leq 1$, the first claim is trivial. The second claim is trivial by (c). □

**Proposition B.5** (Limiting behavior for learning from geometry-inspired perturbations on natural data (deterministic label)). *Suppose that Ineq. (2) holds. Assume* $\|\boldsymbol{x}_n\| = \Theta(\sqrt{d})$ *for any* $n \in [N]$ *and* $\|\boldsymbol{z}\| = \Theta(\sqrt{d})$. *Assume*

$$\sum_{n=1}^{N}\lambda_n|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g_1(N,d)) \Rightarrow \left|\sum_{n=1}^{N}\lambda_n y_n\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\right| = \Theta(g_1(N,d)), \tag{76}$$

$$\sum_{n=1}^{N}\lambda_n|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g_2(N,d)) \Rightarrow \left|\sum_{n=1}^{N}\lambda_n y_n\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\right| = \Theta(g_2(N,d)). \tag{77}$$

*where $g_1$ and $g_2$ are positive functions of $N$ and $d$. Then, the following statements hold:*

- ***Consistent growth rate.*** *For any $\boldsymbol{z}$, $|T_1(\boldsymbol{z})| = \Theta(g_3(N,d)) \Leftrightarrow |T_2(\boldsymbol{z})| = \Theta(g_3(N,d))$, where $g_3$ is a positive function of $N$ and $d$.*
- ***Consistent upper bound.*** *For any $\boldsymbol{z}$, $|T_1(\boldsymbol{z})| = \mathcal{O}(d/\sqrt{N})$ and $|T_2(\boldsymbol{z})| = \mathcal{O}(d/\sqrt{N})$.*
- ***Test sample strongly correlated with few training samples.*** *If there exists $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(d)$ and $\sum_{n:|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle| \neq \Theta(d)} |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(d)$ holds, then $|T_1(\boldsymbol{z})| = \Theta(d/N)$ and $|T_2(\boldsymbol{z})| = \Theta(d/N)$.*
- ***Test sample weakly correlated with many training samples.*** *If there are $\Theta(N)$ instances of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(d/\sqrt{N})$, then $|T_1(\boldsymbol{z})| = \Theta(d/\sqrt{N})$ and $|T_2(\boldsymbol{z})| = \Theta(d/\sqrt{N})$.*

*Proof.* By definition in Theorem 3.2 and Corollary 3.3, $\lambda_n = \Theta(1/d)$ and $\lambda_n^{\mathrm{adv}} = \Theta(1/d)$. Under Ineq. (2), $\epsilon = \mathcal{O}(\sqrt{d/N})$. Because we can set $\epsilon$ freely under Ineq. (2), we can consider $\epsilon = \Theta(\sqrt{d/N})$. By Lemma B.3, $\| \sum_{n=1}^{N} \lambda_n y_n \boldsymbol{x}_n \| = \Theta(\sqrt{N/d})$. Let $S := \{(\alpha,\beta) : n \in [N], |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^\alpha d^\beta)\}$. As shown in Lemma B.4, $\alpha$ and $\beta$ is constrained to $\alpha \leq 0$ and $\beta \leq 1$. Denote the number of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^\alpha d^\beta)$ by $C(\alpha,\beta)$. By Lemma B.4, the upper bound of $C(\alpha,\beta)$ is constrained to $C(\alpha,\beta) = \mathcal{O}(\min(N^{-2\alpha}, N))$. Let $[N]_{\alpha,\beta}$ be the subset of $[N]$ such that $\{n \in [N] : |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^\alpha d^\beta)\}$.

**Consistent growth rate.** Since $\lambda_n = \Theta(1/d)$ and $\lambda_n^{\mathrm{adv}} = \Theta(1/d)$,

$$\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g(N,d)) \Leftrightarrow \sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g(N,d)). \tag{78}$$

Under the assumption,

$$\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g(N,d)), \quad \sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g(N,d)) \tag{79}$$

$$\Rightarrow \left| \sum_{n=1}^{N} \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \right| = \Theta(g(N,d)), \quad \left| \sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle \right| = \Theta(g(N,d)). \tag{80}$$

Using the above orders,

$$\begin{aligned} |T_1(\boldsymbol{z})| =& \frac{\Theta(g(N,d))}{\Theta(\frac{N}{d})} = \Theta\left(\frac{dg(N,d)}{N}\right), \\ |T_2(\boldsymbol{z})| =& \Theta\left(\sqrt{\frac{d}{N}}\right) \frac{\Theta(g(N,d))}{\Theta\left(\sqrt{\frac{N}{d}}\right)} = \Theta\left(\frac{dg(N,d)}{N}\right) \end{aligned} \tag{81}$$

**Consistent upper bound.** By Lemma B.4, $\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \mathcal{O}(\sqrt{N})$. Thus, $T_1(\boldsymbol{z}) = \mathcal{O}(d/\sqrt{N})$ and $T_2(\boldsymbol{z}) = \mathcal{O}(d/\sqrt{N})$.

**Test sample strongly correlated with few training samples.** By Lemma B.4, there are at most $\mathcal{O}(1)$ instance of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(d)$, i.e., $C(0,1) = \mathcal{O}(1)$. Now, since we assume that there exists $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(d)$, $C(0,1) = \Theta(1)$ holds. Lemma B.4 guarantees that there can exist at most $\mathcal{O}(\min(N^{-\alpha}, N))$ instances of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^\alpha d^\beta)$ for $\alpha \leq 0$ and $\beta \leq 1$, i.e., $C(\alpha,\beta) = \mathcal{O}(\min(N^{-\alpha}, N))$. Let $S' := S \setminus \{(0,1)\}$. Now,

$$\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \sum_{(\alpha,\beta)\in S} \sum_{n\in[N]_{\alpha,\beta}}^{N} \Theta(N^\alpha d^{\beta-1}) \tag{82}$$

$$= \sum_{n\in[N]_{0,1}}^{N} \Theta(N^\alpha d^{\beta-1}) \tag{83}$$

$$+ \sum_{(\alpha,\beta)\in S'} \Theta(N^\alpha d^{\beta-1}) \tag{84}$$

$$
\begin{aligned}
&= c(0,1)\Theta(1) + \mathcal{O}(1) && (85)\\
&= \Theta(1). && (86)
\end{aligned}
$$

Thus, $T_1(\boldsymbol{z}) = \Theta(d/N)$ and $T_2(\boldsymbol{z}) = \Theta(d/N)$.

**Test sample weakly correlated with many training samples.** Note that Lemma B.4 guarantees that there can exist $\Theta(N)$ instances of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(d/\sqrt{N})$. Let $S' := S \setminus \{(-1/2, 1)\}$. By the result of (consistent upper bound),

$$
\begin{aligned}
\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = &\sum_{(\alpha,\beta)\in S} C(\alpha,\beta)\Theta(N^{\alpha} d^{\beta-1}) && (87)\\
= &C\left(-\frac{1}{2}, 1\right)\Theta(N^{-1/2}) + \sum_{(\alpha,\beta)\in S'} \mathcal{O}(\min(N^{-2\alpha}, N))\Theta(N^{\alpha} d^{\beta-1}) && (88)\\
= &\Theta(\sqrt{N}) + \mathcal{O}(\sqrt{N}) && (89)\\
= &\Theta(\sqrt{N}). && (90)
\end{aligned}
$$

Thus, $g(N,d) = \sqrt{N}$ holds, indicating $T_1(\boldsymbol{z}) = \Theta(d/\sqrt{N})$ and $T_2(\boldsymbol{z}) = \Theta(d/\sqrt{N})$. □

**Proposition B.6** (Limiting behavior for learning from geometry-inspired perturbations on natural data (random label))**.** *Suppose that Ineq. (2) holds. Assume $\|\boldsymbol{x}_n\| = \Theta(\sqrt{d})$ for any $n \in [N]$ and $\|\boldsymbol{z}\| = \Theta(\sqrt{d})$. Suppose that $y_n^{\mathrm{adv}}$ is randomly sampled from $\{\pm 1\}$ for each $n \in [N]$. Assume*

$$
\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(g_1(N,d)) \Rightarrow \left|\sum_{n=1}^{N} \lambda_n y_n \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\right| = \Theta(g_1(N,d)).. \tag{91}
$$

*where $g_1$ is a positive function of $N$ and $d$. Let $S := \{(\alpha,\beta) : n \in [N], |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^{\alpha} d^{\beta})\}$. Denote the number of $n$ such that $|\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle| = \Theta(N^{\alpha} d^{\beta})$ by $C(\alpha,\beta)$. Let $(\alpha^{\max}, \beta^{\max}) := \arg\max_{(\alpha,\beta)\in S} C(\alpha,\beta) N^{\alpha} d^{\beta-1}$ as $N \to \infty$ and $d \to \infty$. By abuse of notation, we denote $f(x) > \mathcal{O}(g(x))$ if $f(x)$ grows faster than $g(x)$. Similarly, we denote $f(x) < \Omega(g(x))$ if $f(x)$ shrinks faster than $g(x)$. Then, the following statements hold with probability at least 99.99%:*

- ***Growth rate.*** *For any $\boldsymbol{z}$, the growth rate of $|T_2(\boldsymbol{z})|$ is larger than or equal to $|T_1(\boldsymbol{z})|$.*
- ***Upper bound of effect of mislabeled data.*** *For any $\boldsymbol{z}$, $|T_1(\boldsymbol{z})| = \mathcal{O}(d/N)$.*
- ***Faster growth.*** *The growth rate of $|T_2(\boldsymbol{z})|$ is larger than $|T_1(\boldsymbol{z})|$ If either of the following conditions holds:*

$$
\begin{aligned}
&(i)\ C(\alpha^{\max}, \beta^{\max}) N^{\alpha^{\max}} d^{\beta^{\max}-1} > \mathcal{O}(1) && (92)\\
&(ii)\ C(\alpha^{\max}, \beta^{\max}) N^{\alpha^{\max}} d^{\beta^{\max}-1} = \Theta(1)\\
&\quad \text{and } C(\alpha^{\max}, \beta^{\max}) > \mathcal{O}(1) \text{ for any } (\alpha^{\max}, \beta^{\max}). && (93)
\end{aligned}
$$

- ***Consistent shrinks.*** *If $C(\alpha^{\max}, \beta^{\max}) N^{\alpha^{\max}} d^{\beta^{\max}-1} = \mathcal{O}(g_2(N,d)) < \Omega(1)$, then $|T_1(\boldsymbol{z})| = \mathcal{O}(dg_2(N,d)/N)$ and $|T_2(\boldsymbol{z})| = \mathcal{O}(dg_2(N,d)/N)$, where $g_2(N,d)$ is a positive function of $N$ and $d$. In addition, as $N \to \infty$ and $d \to \infty$ while preserving $d/N = \Theta(1)$, $f^{\mathrm{bdy}}(\boldsymbol{z}) = 0$ and $f^{\mathrm{bdy}}_{\mathrm{adv}}(\boldsymbol{z}) = 0$.*
- ***Consistent growth rate.*** *If there exists $(\alpha^{\max}, \beta^{\max})$ such that*

$$
C(\alpha^{\max}, \beta^{\max}) N^{\alpha^{\max}} d^{\beta^{\max}-1} = \Theta(1) \text{ and } C(\alpha^{\max}, \beta^{\max}) = \Theta(1), \tag{94}
$$

*then $|T_1(\boldsymbol{z})| = \mathcal{O}(d/N)$ and $|T_2(\boldsymbol{z})| = \Theta(d/N)$.*

*Proof.* As a preliminary, please refer to the proof of Proposition B.5. The notations follow it. Similar to Eq. (81),

$$
|T_1(\boldsymbol{z})| = \Theta\left(\frac{d}{N}\right)\left|\sum_{n=1}^{N} \lambda_n^{\mathrm{adv}} y_n^{\mathrm{adv}} \langle \boldsymbol{x}_n, \boldsymbol{z}\rangle\right|, \qquad |T_2(\boldsymbol{z})| = \Theta\left(\frac{d}{N}\right)\sum_{n=1}^{N} \lambda_n |\langle \boldsymbol{x}_n, \boldsymbol{z}\rangle|. \tag{95}
$$

By Hoeffding's inequality,

$$\mathbb{P}\left[\left|\sum_{n=1}^{N}\lambda_n^{\mathrm{adv}}y_n^{\mathrm{adv}}\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle\right|>t\right]\le 2\exp\left(-\frac{2t^2}{\sum_{n=1}^{N}{\lambda_n^{\mathrm{adv}}}^2\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle^2}\right). \tag{96}$$

Thus, for a constant $C>0$, $|\sum_{n=1}^{N}\lambda_n^{\mathrm{adv}}y_n^{\mathrm{adv}}\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|$ is larger than $C\sqrt{\sum_{n=1}^{N}{\lambda_n^{\mathrm{adv}}}^2\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle^2}$ with probability at most $2\exp\bigl(-2C^2\bigr)$. Therefore, $|\sum_{n=1}^{N}\lambda_n^{\mathrm{adv}}y_n^{\mathrm{adv}}\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\mathcal{O}(\sqrt{g(N,d)})$ holds with sufficiently high probability if $\sum_{n=1}^{N}{\lambda_n^{\mathrm{adv}}}^2\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle^2=\mathcal{O}(g(N,d))$.

**Growth rate.** By Lemma B.4, the claims is established.

**Upper bound of effect of mislabeled data.** By Lemma B.4, $\sum_{n=1}^{N}{\lambda_n^{\mathrm{adv}}}^2\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle^2=\mathcal{O}(1)$. Thus, the claim is established.

**Consistent shrinks.** If $T_2(\boldsymbol{z})$ shrinks with $d$ and $N$, by the result of (growth rate), $T_1(\boldsymbol{z})$ shrinks more faster or equally. Since $f^{\mathrm{bdy}}(\boldsymbol{z})\le\sum_{n=1}^{N}\lambda_n|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\mathcal{O}(g_2(N,d))$, $f^{\mathrm{bdy}}(\boldsymbol{z})=0$ holds.

**Consistent growth rate.** Consistent growth rate between $T_1(\boldsymbol{z})$ and $T_2(\boldsymbol{z})$ can be found only in the case with $\sum_{n=1}^{N}{\lambda^{\mathrm{adv}}}^2\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle^2=\mathcal{O}(1)$ and $\sum_{n=1}^{N}\lambda|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(1)$. This holds only when the following condition holds:

$$C(\alpha^{\max},\beta^{\max})N^{\alpha^{\max}}d^{\beta^{\max}-1}=\Theta(1)\text{ and }C(\alpha^{\max},\beta^{\max})=\Theta(1). \tag{97}$$

**Faster growth.** For cases except for (consistent shrinks) and (Consistent growth rate), by Lemma B.4, $T_2(\boldsymbol{z})$ grows faster than $T_1(\boldsymbol{z})$.

□

**Lemma B.7** (Representation of test sample strongly correlated with few training samples)**.** *Suppose that Ineq. (2) holds. Assume $\|\boldsymbol{x}_n\|=\Theta(\sqrt{d})$ for any $n\in[N]$ and $\|\boldsymbol{z}\|=\Theta(\sqrt{d})$. Let $S:=\{(\alpha,\beta):n\in[N],|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(N^{\alpha}d^{\beta})\}$. Denote the number of $n$ such that $|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(N^{\alpha}d^{\beta})$ by $C(\alpha,\beta)$. Let $(\alpha^{\max},\beta^{\max}):=\arg\max_{(\alpha,\beta)\in S}C(\alpha,\beta)N^{\alpha}d^{\beta-1}$ as $N\to\infty$ and $d\to\infty$. Let $Q\subset[N]$ be the set of indices such that $|Q|=\Theta(1)$. Let $\boldsymbol{r}\in\mathbb{R}^d$ be the vector such that $\sum_{n=1}^{N}|\langle \boldsymbol{x}_n,\boldsymbol{r}\rangle|=\mathcal{O}(d)$. Then, the following statement holds:*

$$\begin{aligned}&\text{There exists }(\alpha^{\max},\beta^{\max})\text{ such that}\\ &C(\alpha^{\max},\beta^{\max})N^{\alpha^{\max}}d^{\beta^{\max}-1}=\Theta(1)\text{ and }C(\alpha^{\max},\beta^{\max})=\Theta(1), &(98)\\ &\Rightarrow\boldsymbol{z}=\sum_{n\in Q}\pm\Theta(1)\boldsymbol{x}_n+\boldsymbol{r}. &(99)\end{aligned}$$

*Proof.* Note that $\alpha$ and $\beta$ are constrained to $\alpha\le 0$ and $\beta\le 1$, respectively, by Lemma B.4. In addition, $C(\alpha,\beta)$ is also constrained to $C(\alpha,\beta)=\mathcal{O}(\min(N^{-2\alpha},N))$ by Lemma B.4. If $C(\alpha^{\max},\beta^{\max})N^{\alpha^{\max}}d^{\beta^{\max}-1}=\Theta(1)$ and $C(\alpha^{\max},\beta^{\max})=\Theta(1)$, $N^{\alpha^{\max}}d^{\beta^{\max}-1}=\Theta(1)$. This holds with only $\alpha^{\max}=0$ and $\beta^{\max}=1$. Thus, there must exist $n$ such that $|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(d)$. Note that there are at most $\mathcal{O}(1)$ instances of such $n$. Therefore, $\boldsymbol{z}$ has the term $\sum_{n\in Q}\pm\Theta(1)\boldsymbol{x}_n$ for $|Q|=\Theta(1)$. Note that $|\langle \boldsymbol{x}_1,\boldsymbol{z}\rangle|=\Theta(d)$ also holds for $\boldsymbol{z}=\pm\Theta(1)\boldsymbol{x}_2\pm\Theta(1)\boldsymbol{x}_3+\cdots\pm\Theta(1)\boldsymbol{x}_N$ under $\gamma^3R_{\min}^4/(3NR_{\max}^2)\ge p_{\max}$. However, as shown in Lemma B.4, such $\boldsymbol{z}$ does not satisfy $\|\boldsymbol{z}\|=\Theta(\sqrt{d})$. Then, let us consider the properties of $\boldsymbol{r}$. To maintain $\alpha^{\max}=0$ and $\beta^{\max}=1$, $\max_{(\alpha,\beta)\in S\setminus\{0,1\}}C(\alpha,\beta)N^{\alpha}d^{\beta-1}=\mathcal{O}(1)$. In other words, $\max_{(\alpha,\beta)\in S\setminus\{0,1\}}C(\alpha,\beta)N^{\alpha}d^{\beta}=\mathcal{O}(d)$. This holds if $\sum_{n=1}^{N}|\langle \boldsymbol{x}_n,\boldsymbol{r}\rangle|=\mathcal{O}(d)$. □

**Theorem 4.2** (Consistent decision of learning from geometry-inspired perturbations on natural data)**.** *Suppose that Ineq. (2), $\|\boldsymbol{x}_n\|=\Theta(\sqrt{d})$ for all $n\in[N]$, and $\|\boldsymbol{z}\|=\Theta(\sqrt{d})$ hold. Consider $N\to\infty$ and $d\to\infty$ while keeping $d/N=\Theta(1)$. Suppose that $y_n^{\mathrm{adv}}$ is randomly sampled from $\{\pm1\}$ for each $n\in[N]$. Assume $|\sum_{n=1}^{N}\lambda_n y_n\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(g(N,d))$ if $\sum_{n=1}^{N}\lambda_n|\langle \boldsymbol{x}_n,\boldsymbol{z}\rangle|=\Theta(g(N,d))$,* *where $g$ is a positive function of $N$ and $d$. Let $Q \subset [N]$ be a set of indices such that $|Q| = \Theta(1)$, and let $\boldsymbol{r} \in \mathbb{R}^d$ be a vector such that $\sum_{n=1}^{N} |\langle \boldsymbol{x}_n, \boldsymbol{r} \rangle| = \mathcal{O}(d)$. If $\boldsymbol{z}$ is **not** represented as $\boldsymbol{z} = \sum_{n \in Q} \pm\Theta(1)\boldsymbol{x}_n + \boldsymbol{r}$, then $\mathrm{sgn}(f_{\mathrm{adv}}^{\mathrm{bdy}}(\boldsymbol{z})) = \mathrm{sgn}(f^{\mathrm{bdy}}(\boldsymbol{z}))$ holds with probability at least 99.99%.*

*Proof.* By Proposition B.6 and Lemma B.7, this is trivial. □

# C Experimental Settings

In this section, we present the experimental settings in Fig. 1. An NVIDIA A100 GPU was used. A six-layer convolutional neural network was employed for both MNIST [3] and Fashion-MNIST [19], whereas WideResNet-28-10 with a dropout ratio of 0.3 was used for CIFAR-10 [10]. The batch size was set to 128. While no data augmentation was applied to MNIST and Fashion-MNIST, CIFAR-10 utilized random cropping and random horizontal flipping. Training was conducted using stochastic gradient descent with Nesterov momentum set at 0.9 and a weight decay of $5 \times 10^{-4}$. The initial learning rates are 0.01 for MNIST and Fashion-MNIST and 0.1 for CIFAR-10. The perturbation constraint $\epsilon$ was set to 2.0 for MNIST and Fashion-MNIST and 0.5 for CIFAR-10. As a loss function for both training and adversarial attacks, we adopted the cross-entropy loss. The number of epochs was set at 100 for MNIST and 200 for both Fashion-MNIST and CIFAR-10. The scheduler reduced the learning rate to 10% of its original value if the training loss did not decrease over 10 consecutive epochs.

Adversarial attacks were performed using projected gradient descent [11]. The final output was selected as the adversarial example that maximized the loss over all steps. The step size of projected gradient descent is $\epsilon/5$. The number of steps is 100. It should be noted that although we primarily considered geometry-inspired perturbations (cf. Eq. (1)) in the theoretical discussion, these perturbations are not computationally feasible in practice. This is because we cannot practically obtain the decision boundary of a one-hidden-layer neural network (cf. Eq. (7)) and the explicit value of $\lambda_n$ (cf. Theorem A.1).