Origins of Low-Dimensional Adversarial Perturbations

Elvis Dohmatob
Facebook AI Research

Chuan Guo
Facebook AI Research

Morgane Goibert
Criteo AI Lab

Abstract

Machine learning models are known to be susceptible to adversarial perturbations. Even more concerning is the fact that these adversarial perturbations can be found by black-box search using surprisingly few queries, which essentially restricts the perturbation to a subspace of dimension k—much smaller than the dimension d of the image space. This intriguing phenomenon raises the question: *Is the vulnerability to black-box attacks inherent or can we hope to prevent them?* In this paper, we initiate a rigorous study of the phenomenon of low-dimensional adversarial perturbations (LDAPs). Our result characterizes precisely the sufficient conditions for the existence of LDAPs, and we show that these conditions hold for neural networks under practical settings, including the so-called lazy regime wherein the parameters of the trained network remain close to their values at initialization. Our theoretical results are confirmed by experiments on both synthetic and real data.

Contents

Despite their widespread use and success in solving real-life tasks like speech recognition, face recognition, assisted driving, etc., neural networks (NNs) are known to be vulnerable to *adversarial perturbations*, i.e. imperceptible modifications of input data causing the model to fail (Szegedy et al., 2013). This vulnerability can be exploited by an adversary to manipulate the model's decision at test-time and can constitute a serious security risk if left unchecked.

Although the majority of attacks using adversarial perturbation rely on access to the model parameters for gradient-based search, a more recent series of works showed that query-based black-box attacks are also possible. Methods such as Boundary Attack (Brendel et al., 2017), NES (Ilyas et al., 2018), SimBA (Guo et al., 2019) and HopSkipJump (Chen et al., 2020) approximate the full gradient via a Monte-Carlo finite-difference estimate which sub-samples the coordinates randomly. Surprisingly, existing black-box attacks can be carried out using a very small number of queries, which suggests that adversarial examples are abundant in low-dimensional subspaces.

For instance, Chen et al. (2017) initiated the study on black-box attacks using a finite-difference approximation for the gradient to perform gradient-ascent search. This method inspired others such as Boundary Attack (Brendel et al., 2017), NES (Ilyas et al., 2018), SimBA (Guo et al., 2019) and HopSkipJump (Chen et al., 2020) that approximate the full finite-difference gradient via a Monte-Carlo estimate which sub-samples the coordinates randomly. This approach only requires sampling a very small fraction of the total input space, e.g., on ImageNet where the input dimensionality is approximately 150K, SimBA perturbs as few as 1665 random coordinates and succeeds with over 98.6% probability (Guo et al., 2019). Subsequent works also performed adversarial search in a *fixed* subspace such as the low-frequency subspace (Yin et al., 2019; Guo et al., 2018) or by selecting the subspace in a distribution-dependent manner using an independently-trained NN (Tu et al., 2019; Yan et al., 2019; Huang and Zhang, 2019).

These empirical findings lead us to hypothesize that adversarial perturbations exist with high probability in low-dimensional subspaces. Our work initiates a rigorous study to understand low-dimensional adversarial perturbations (LDAPs). We provide rigorous explanations for the empirical success of some powerful heuristics that have appeared in the literature (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018; Guo et al., 2018; Yin et al., 2019; Chen et al., 2020).

1.1 Main Contributions

Our main results are as follows. In Sections 4 and 5, we consider different realistic notions of smoothness for a binary classifier. These smoothness assumptions allow us to

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

linearize the decision boundary locally and derive generic lower-bounds on the fooling rate of any subspace V of the feature space $\mathbb{R}^d$. Our bounds reveal the role of

- the **alignment** of the subspace V with the unit-normals at the classifier's decision-boundary,
- the distribution of classifier's **pointwise margin**,
- the **local smoothness** of the classifier's decision-boundary, and
- the **attacker's budget** ε (measured in Euclidean norm).

We formalize a notion of alignment in Section 3.

For random subspaces of sufficiently high dimension (Guo et al., 2019) and subspaces obtained via SVD on the gradients (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018), our results provide transparent lower-bounds on the fooling rate, which explain the empirical success of the very efficient heuristic methods that have been proposed in the literature for constructing LDAPs; see section 6. Moreover, the lower-bounds only depend on the distributions of the predictions and the gradients of the model and so can be empirically estimated on held-out data, making them a practical predictor for the adversarial vulnerability of classifiers. Our theoretical results are confirmed by numerous experiments on real and simulated data (Section 7). In all cases, the bounds can be easily evaluated and are close to the actual fooling rates.

1.2 Literature Overview

Earlier experiments showed that adversarial attacks based on a single direction of feature space (i.e., UAPS) can be designed to effectively fool neural networks (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018). UAPs are often more transferable across datasets and architectures than classical attacks, making them interesting for use in practice. Their theoretical analysis has been initiated in Moosavi-Dezfooli et al. (2018), where the authors establish lower-bounds for the fooling rate of UAPs under certain curvature conditions on the decision boundary. The aforementioned work has two fundamental limitations. First, the notions of curvature used are stated in terms of unconstrained optimal adversarial perturbation (i.e., the closest point) for an arbitrary input point, and thus are not easy to verify in practice. Also, the existence of the UAP is only guaranteed within a subspace which is required to satisfy a global alignment property with the gradients of the model. In contrast, we use a more flexible curvature requirement (refer to Definition 3), which is adapted to any subspace under consideration, and we prove results that are strong enough to provide a satisfactory theory of LDAPs, and UAPs in particular, under very general settings.

Guo (2020) studied LDAPs when the attacker is constrained to a uniformly random k-dimensional subspace. For classifiers whose decision regions are half-spaces and spheres in $\mathbb{R}^d$, they established the existence of low-dimensional adversarial subspaces under a Gaussian concentration assumption on the data. Our work considers more general decision regions (e.g. of certain neural networks) and more general data distributions and subspaces. Our results recover the findings of Guo (2020) as special cases.

1.3 Need for New Theoretical Tools

Classical theoretical works on understanding adversarial examples (Tsipras et al., 2019; Shafahi et al., 2018a; Mahloujifar et al., 2018; Gilmer et al., 2018; Dohmatob, 2019) focus on the case of adversarial attacks on the full feature space. They use the concentration property of certain high-dimensional (e.g., multivariate Gaussians, distributions satisfying log-Sobolev inequalities, etc.), to establish that an imperfect classifier will admit adversarial examples. However, such techniques cannot be used directly when we add the constraint that the attacks only live in a low-dimensional subspace. Thus, new techniques are needed. Such techniques were initiated in Guo (2020) for the case of linear models, and are extended in our paper to non-linear models.

2 Preliminaries

Notations. We denote by $[d]$ the set of integers from 1 to d inclusive. The notation t_+ is for the maximum of t and 0, $\|u\|$ the L_2-norm (unless otherwise stated) of a vector u, and $\|A\|_{op}$ denotes the operator norm of a real matrix A. The unit-sphere (resp. closed unit-ball) in $\mathbb{R}^d$ is written $\mathcal{S}_{d-1}$ (resp. B_d). The orthogonal projection of a vector $z \in \mathbb{R}^d$ onto the subspace $V \subseteq \mathbb{R}^d$ is denoted $\Pi_V z$. As usual, asymptotic notation $F(d) = O(G(d))$ (also written $F(d) \lesssim G(d)$) means there exists a constant c such that $F(d) \leq c \cdot G(d)$ for sufficiently large d, while $F(d) = \Omega(G(d))$ means $G(d) = O(F(d))$, and $F(d) = \Theta(G(d))$ or $F(d) \asymp G(d)$ means $F(d) \lesssim G(d) \lesssim F(d)$. Finally, $F(d) = o(G(d))$ means $F(d)/G(d) \to 0$ as $d \to \infty$.

2.1 Binary Classification and Adversarial Examples

We consider a binary classification setup, where $X = (X_1, ..., X_d) \in \mathbb{R}^d$ denotes an input of dimension d (e.g. for the MNIST dataset, $d = 784$) drawn from a probability distribution $\mathbb{P}_X$ on $\mathbb{R}^d$. We will denote by $f : \mathbb{R}^d \to \mathbb{R}$ the feature map, and $h_f = \text{sign} \circ f$ the corresponding classifier, with the arbitrary convention that $\text{sign}(0) = -1$. For example, for NNs, $f(x)$ would be the predicted *logit*, for a closed ball of radius $r > 0$ in $\mathbb{R}_d$, $f(x) := (\|x\|^2 - r^2)/2$, and for a half-space (linear classifier), $f(x) := x^\top w - b$. The binary classifier h_f can be unambiguously identified

with a measurable subset of $\mathbb{R}^d$

$$C = \{x \in \mathbb{R}^d \mid h_f(x) = -1\} = \{x \in \mathbb{R}^d \mid f(x) \leq 0\}, \quad (1)$$

called the negative *decision-region* of h. Thus, the complement $C' := \mathbb{R}^d \setminus C$ of C is the positive *decision-region* of h. Of course, the terms "negative" or "positive" are interchangeable, as we can always consider the classifier $-h$ instead. Therefore, without loss of generality, we shall focus our attention on adversarial attacks on the positive decision-region C'.

Given an input $x \in C'$ classified by h_f as positive, an adversarial perturbation for x is a vector $a \in \mathbb{R}^d$ of size $\|a\|$ such that $x + a \in C$. The goal of the attacker is to move points from C' to C with small perturbations. Note that we are not interested in the true labels of the inputs, just the robustness of the classifier w.r.t. its own predictions. However, note that this distinction is not important for classifiers which are already very accurate in the classical sense.

The notion of *margin* will be important in the sequel.

Definition 1 (Margin at a point)**.** *If f is differentiable at a non-critical point $x \in \mathbb{R}^d$, its margin at x, denoted $m_f(x)$, is defined by*

$$m_f(x) := (f(x))_+ / \|\nabla f(x)\|. \quad (2)$$

For example, if $f(x) \equiv x^\top w - b$ for some scalar $b \in \mathbb{R}$ non-zero $w \in \mathbb{R}^d$, as in the case where the classifier is a half-space, then $m_f(x) = (x^\top w - b)_+ / \|w\|$. In this case, $m_f(x)$ also corresponds to the distance of x from the negative decision-region of the classifier.

2.2 Low-Dimensional Adversarial Perturbations

In this paper, we focus on *low-dimensional* perturbations (LDAPs) (Guo et al., 2018, 2019; Tu et al., 2019; Yan et al., 2019; Huang and Zhang, 2019; Guo, 2020), meaning that the perturbations a are limited to a k-dimensional subspace V of $\mathbb{R}^d$ whose choice is left to the attacker. Here, k can be much smaller than d. The special case where $k = 1$ corresponds to the scenario where the attacker is allowed to operate in one dimension only (e.g. modify the same pixel in all images of the same class), also famously known as *universal adversarial perturbations (UAPs)* (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018). More generally, given a subspace V of $\mathbb{R}^d$, let C_V^ε be the set of all points in $\mathbb{R}^d$ which can be pushed into the negative decision-region C by adding a perturbation of size ε in V, that is

$$C_V^\varepsilon := \{x \in \mathbb{R}^d \mid \exists v \in V \text{ with } \|v\| \leq \varepsilon \text{ s.t } x + v \in C\}, \quad (3)$$

where $B_V := V \cap B_d$ is the unit-ball in V. Note that by definition, $x \in C_V^\varepsilon$ iff $(x + \varepsilon B_V) \cap C \neq \emptyset$. In the particular case of full-dimensional attacks where $V = \mathbb{R}^d$, the set C_V^ε corresponds to the usual ε-expansion C^ε of C, i.e., the set of points in $\mathbb{R}^d$ which are at a distance at most ε from C. This case has been extensively studied in Shafahi et al. (2018b); Fawzi et al. (2018); Mahloujifar et al. (2019); Dohmatob (2019). Note that it always holds that $C \subseteq C_V^\varepsilon \subseteq C^\varepsilon$.

Definition 2 (Fooling rate of a subspace)**.** *Given an attack budget $\varepsilon \geq 0$, the fooling rate $\mathrm{FR}(V; \varepsilon)$ of a subspace $V \subseteq \mathbb{R}^d$ is the proportion of test data which can be moved from the positive decision-region C' to the negative decision-region C by moving a distance ε along V, that is*

$$\mathrm{FR}(V; \varepsilon) := \mathbb{P}_X(X \in C_V^\varepsilon \mid X \in C'). \quad (4)$$

Note that by definition of C_V^ε, the fooling rate $\mathrm{FR}(V; \varepsilon)$ is a supremum over all possible attackers operating in the subspace V, and with L_2-norm budget ε. In particular, $\mathrm{FR}(\mathbb{R}^d; \varepsilon)$ is the usual optimal fooling rate of an adversarial attack with budget ε, without any subspace constraint, and already studied extensively in the literature (Shafahi et al., 2018b; Fawzi et al., 2018; Mahloujifar et al., 2019; Dohmatob, 2019).

2.3 Warm-up: Insights from Linear models

We start with the simple case of a linear binary classifier on $\mathbb{R}^d$, for which the negative decision-region C (and therefore the positive decision region too) is a half-space given by

$$H_{w,b} := \{x \in \mathbb{R}^d \mid x^\top w - b \leq 0\}, \quad (5)$$

on with unit-normal vector $w \in \mathbb{R}^d$ and bias parameter $b \in \mathbb{R}$. This corresponds to taking $f(x) := x^\top w - b$ in (1). The following result generalizes a result of Guo (2020) (see Lemma 2.2 therein) which was only established in the case where the marginal distribution of the features P_X is the standard Gaussian distribution on $\mathbb{R}^d$.

Proposition 1. *Consider the scenario where C is the half-space $H_{w,b}$ defined in (5). For any subspace V of $\mathbb{R}^d$ and $\varepsilon \geq 0$, it holds $\mathrm{FR}(V; \varepsilon) \geq \mathbb{P}_X(X^\top w - b \leq \|\Pi_V w\| \varepsilon \mid X \in C')$.*

In particular, if V is a uniformly random k-dimensional subspace of $\mathbb{R}^d$, then for any $t \in (0, \sqrt{k/d})$ it holds w.p $1 - 2e^{-t^2 d/2}$ over V that

$$\mathrm{FR}(V; \varepsilon) \geq \mathbb{P}_X(X^\top w - b \leq (\sqrt{k/d} - t)\varepsilon \mid X \in C'). \quad (6)$$

Interpretation of Proposition 1. To understand the power of the the above proposition, consider the cause where $P_X = \mathcal{N}(0, I_d)$ and $b = 0$ so that $\mathbb{P}_X(C) = \mathbb{P}_X(C') = 1/2$. Note that a typical $x \sim P_X$ has norm of order $N = \mathbb{E}\|x\| \asymp \sqrt{d}$. Thus a random perturbation of dimension $k = \sqrt{d} \ll d$ and of ℓ_2-norm $\varepsilon = \sqrt{d/k} = d^{1/4} \ll N$ is sufficient to change the decision of the classifier on a proportion

$$\begin{aligned} \mathrm{FR}(V; \epsilon) &\geq \mathbb{P}_X(X^\top w \leq 1 \mid X^\top w \geq 0) \\ &= (\Phi(1) - \Phi(0))/(1/2) \approx 68\% \end{aligned}$$

from negative to positive.

Proof of Proposition 1. Indeed, one computes

$$\begin{aligned}\mathrm{FR}(V;\varepsilon) &:= \mathbb{P}_X(X \in C_V^\varepsilon \mid X \in C') \\ &\geq \sup_{v \in V} \mathbb{P}_X(X \in C_v^\varepsilon \mid X \in C') \\ &= \sup_{v \in V \cap \mathcal{S}_{d-1}} \mathbb{P}_X(X^\top w + \varepsilon v^\top w - b \leq 0 \mid X \in C') \\ &= \mathbb{P}_X(X^\top w - b \leq \varepsilon \|\Pi_V w\| \mid X \in C'),\end{aligned}$$

which proves the first part of the claim. The second part follows from the first part combined with the fact that

$$\|\Pi_V w\| \geq \sqrt{k/d} - t \text{ w.p.} 1 - 2e^{-t^2 d/2}, \tag{7}$$

by basic concentration arguments.

Lifting the Core Ideas to the Non-Linear Setting. In the results of this manuscript, we will emulate the lower-bound (6), for the case of non-linear classifiers. In this direction, first observe that, since the margin for the linear classifier is $m_f(x) := \max(f(x), 0)/\|\nabla f(x)\| = (x^\top w + b)_+$, the lower-bound (6) can be written in expectation-form as

$$\mathbb{E}_V \, \mathrm{FR}(V;\varepsilon) \geq \mathbb{P}(m_f(x) \leq \alpha\varepsilon \mid X \in C') - \delta, \tag{8}$$

with $\alpha = \sqrt{k/d} - t$ and $\delta = 2e^{-t^2 d/2}$. The pair of scalars (α, δ) capture the alignment between the random subspace V, and the gradients of the linear classifier at a random point $X \in C'$, i.e with the normal vector $\eta(X) = \nabla f(x)/\|\nabla f(X)\| = w$, in the sense that

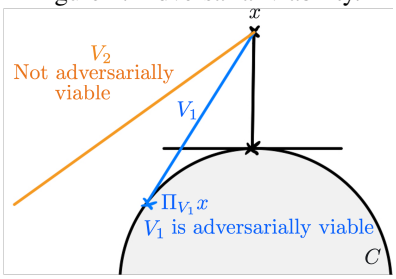
$$\mathbb{P}_{X,V}(\|\Pi_V \eta(X)\| \geq \alpha \mid X \in C') \geq 1 - \delta. \tag{9}$$

Since $\eta(X) = w$ here, and is independent of the feature vector X, (9) is just a restatement of (7). In the general case of non-linear models f (e.g neural nets) and arbitrary subspaces V, inequalities such as (9) will be the basis of so-called adversarially viable subspaces, studied in detail in Section 3.

3 Adversarially Viable Subspaces

We will formalize the notion of an *adversarially viable* subspace which is a subspace V that has a non-negligible inner product with the classifier's gradient, hence it is possible to significantly alter the value of $f(x)$ by moving strictly within V. Intriguingly, such subspaces are pivotal to the empirical success of LDAPs, and we show that popular heuristics lead to adversarially viable subspaces. Then, we prove that when the classifier satisfies certain smoothness conditions, adversarially viable subspaces allow the attacker to follow the gradient direction within V to reach the decision boundary of C for most points $x \in C'$, hence achieving a high fooling rate. Restricting the adversarial perturbation to a given subspace V presents a particular challenge to the attacker. If $\dim V < d$ and $x \in C' := \mathbb{R}^d \setminus C \neq \emptyset$, it is possible that $x \notin C_V^\varepsilon$ for all $\varepsilon > 0$. In particular, if f is convex and the subspace V is orthogonal to the gradient of f at a point $x \in \mathbb{R}^d$, then no amount of perturbation within V will make x closer to the boundary of C, in an effort to flip its predicted class label. See Figure 1 for underlying geometric intuition. Thus, we can hope to establish nontrivial fooling rates only for certain subspaces.

Figure 1: Adversarial viability.



Our first contribution is a crisp characterization of subspaces for which we can hope to achieve a nonzero fooling rate. These are so-called *adversarially viable* subspaces and are a generalization of the subspaces considered in Moosavi-Dezfooli et al. (2018); Moosavi-Dezfooli et al. (2017); Guo (2020).

Definition 3 (Adversarially viable subspace)**.** *Given $\alpha \in (0, 1]$ and $\delta \in [0, 1)$, a possibly random subspace $V \subseteq \mathbb{R}^d$ is said to be adversarially (α, δ)-viable if*

$$\mathbb{P}_{X,V}(\|\Pi_V \eta(X)\| \geq \alpha \mid X \in C') \geq 1 - \delta, \tag{10}$$

where $\eta(x) := \nabla f(x)/\|\nabla f(x)\|$ is the gradient direction at x.

The above definition captures the essence of (7), which was the crucial piece in the proof of Proposition 1. To see that this is a generalization of (7), note that $\eta(x) \equiv w$ when C is a half-space (i.e when f is a linear function $f(x) \equiv x^\top w - b$).

We now provide some important examples of adversarially viable subspaces.

3.1 Random Subspaces

Consider the case of a uniformly random k-dimensional subspace V of $\mathbb{R}^d$. Such subspaces have been proposed in the literature (Moosavi-Dezfooli et al., 2017; Guo, 2020), for constructing low-dimensional adversarial perturbations.

Lemma 1. *The random subspace as given above is $(\sqrt{k/d} - t, 2e^{-t^2 d/2})$-viable for any $t \in (0, \sqrt{k/d})$.*

Indeed, this is just a restatement of (7), in the language of Definition 3.

3.2 Top Eigenvectors of Gradient Covariance Matrix

Let $\Sigma_\eta \in \mathbb{R}^{d\times d}$ be the covariance matrix of the gradient direction $\eta(X)$ conditioned on $X \in C'$.

Theorem 1. *For any $k \in [d]$, let $s_k \in (0,1]$ be the sum of first the k eigenvalues of Σ_η. Then, for any $\alpha \in (0, \sqrt{s_k})$, the (deterministic) subspace $V_{\text{eigen},k}$ of $\mathbb{R}^d$ corresponding to the top k eigendirections of Σ_η is adversarially $(\alpha, (1-s_k)/(1-\alpha^2))$-viable.*

Thus, if the histogram of eigenvalues of Σ_η is "spiked" in the sense that $s_k \geq s = \Omega(1)$ for some $k = o(d)$, then $V_{\text{eigen},k}$ is a $o(d)$-dimensional adversarially $(\Omega(1), O(1-s))$-viable subspace! Combined with the results established in the following sections, the preceding observation provides a rigorous justification for the heuristic in Moosavi-Dezfooli et al. (2017); Khrulkov and Oseledets (2018) which proposed UAPs based on eigenvectors of the covariance matrix Σ_η. Our experiments in Section 7 also support this.

4 Lipschitz Decision-Boundary

Consider a binary classifier on $\mathbb{R}^d$ for which the negative decision-region C of the classifier is given by (1), where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function. Let us start by observing that, thanks to a classical result from optimization theory (see Proposition 3.2 of Azé and Corvellec (2017)), if the following condition is satisfied, then any $x \in C'$ is at a distance $d_C(x)$ at most $f(x)/\beta$ from C.

Condition 1 (Uniformly strong gradients)**.** *There exists a constant $\beta > 0$ such that $\|\nabla f(x)\| \geq \beta$ for all $x \in C'$*

Intuitively, under Condition 1, the gradient of f at any point $x \in C'$ is strong enough: gradient-flow started at x then escapes the region C' after travelling a distance $O(f(x))$. This is formalized in the following result which will be extended to the case of subspace attacks in the rest of this section.

Theorem 2 (A lower-bound for full-dimensional attacks)**.** *Under Condition 1, it holds for any $\varepsilon \geq 0$ that*

$$\mathrm{FR}(\mathbb{R}^d; \varepsilon) \geq \mathbb{P}_X(f(X) \leq \beta\varepsilon \mid X \in C'). \tag{11}$$

As an illustration, if we consider f to be a randomly initialized[1] finite-depth ReLU neural-network, one can show (see Daniely and Shacham (2020); Bubeck et al. (2021); Bartlett et al. (2021)) that for any $x \in \mathbb{R}^d$, we have $f(x) = \mathcal{O}(\|x\|/\sqrt{d})$ and $\inf_x \|\nabla f(x)\| = \Omega(1)$ w.h.p. over the weights. The above theorem immediately predicts the existence of adversarial examples of size $\sqrt{d}$ times smaller than the typical L_2-norm of a data point.

[1]With layer widths within $\mathrm{poly}(\log d)$ factors of one another, and weights initialized in the standard way.

4.1 Main Result under Lipschitz Smoothness

We will extend Theorem 2 to the case of subspace attacks, under the following smoothness condition.

Condition 2 (Lipschitz gradients)**.** *There exists a constant $L \geq 0$ such that for all $x, x' \in \mathbb{R}^d$,*

$$\|\nabla f(x') - \nabla f(x)\| \leq L\|x' - x\|. \tag{12}$$

This condition stipulates that the gradient of f varies smoothly on the positive decision-region $C' = \mathbb{R}^d \setminus C$ of the classifier (1). Note that when f is twice-differentiable on C', Condition 2 holds with $L = \sup_{x\in C'} \|\nabla^2 f(x)\|_{op}$, where $\nabla^2 f(x) \in \mathbb{R}^{d\times d}$ is the Hessian of f at x. For example, a feed-forward neural net with bounded weights and twice-differentiable activation function with bounded Hessian (e.g. sigmoid, quadratic, tanh, GELU, cos, sin, etc.) will satisfy Condition 2.

To obtain simplified / more transparent lower-bounds for the fooling rates, we will also need the following natural condition which ensures that there is a strong descent direction at a constant fraction of points in the positive decision-region C', to allow for gradient-based attacks.

Condition 3 (Strong gradients)**.** *For some constants $\beta > 0$ and $\gamma \in [0,1)$, it holds that*

$$\mathbb{P}_X(\|\nabla f(X)\| \geq \beta \mid X \in C') \geq 1 - \gamma. \tag{13}$$

Note that Condition 1 is a special case of Condition 3 corresponding to $\gamma = 0$. The following is one of our main results. It generalizes both Proposition 1 and Theorem 2.

Theorem 3. *Suppose Condition 2 holds. Let V be a possibly random adversarially (α, δ)-viable subspace of $\mathbb{R}^d$. Then,*

(A) For any $\varepsilon \geq 0$, the average fooling rate of V is lower-bounded as follows

$$\mathbb{E}_V[\mathrm{FR}(V;\varepsilon)] \geq \mathbb{P}_X\left(m_f(X) \leq \widetilde{\alpha}(X) \mid X \in C'\right) - \delta, \tag{14}$$

where $\widetilde{\alpha}(X) := \min\left(\alpha\varepsilon/2, \alpha^2\|\nabla f(X)\|/(2L)\right)$. (B) If in addition Condition 3 is in order, then for any $\varepsilon \in [0, \alpha\beta/L]$,

$$\begin{aligned}\mathbb{E}_V \mathrm{FR}(V;\varepsilon) \geq \mathbb{P}_X(m_f(X) \leq \alpha\varepsilon/2 \mid X \in C') \\ - \delta - \gamma.\end{aligned} \tag{15}$$

Remark 1. *Note that the condition "$\varepsilon \leq \alpha\beta/L$" in part (B) of the theorem cannot be removed in general, as is seen in the case where $C = B_d$, and considering any subspace V with $\dim V < d$.*

4.2 Sketch of Proof of Theorem 3

We give a vivid sketch of the proof here. It is an elementary fact in optimization theory that a function $f : \mathbb{R}^d \to \mathbb{R}$

which has the structure stated in Condition 2 admits the following first-order approximation: for all $x, x' \in \mathbb{R}^d$,

$$|f(x') - f(x) - \nabla f(x)^\top (x' - x)| \leq \frac{L}{2}\|x' - x\|^2. \quad (16)$$

Now, starting at a point $x \in C'$, let us move a distance ε in the direction $\Pi_V \nabla f(x)$ to arrive at a point $x' = x - \varepsilon \Pi_V \nabla f(x) \in \mathbb{R}^d$, the above inequality gives the quadratic approximation

$$f(x') \leq f(x) - \varepsilon \|\Pi_V \nabla f(x)\|^2 + \frac{L}{2}\varepsilon^2 \|\Pi_V \nabla f(x)\|^2. \quad (17)$$

After some calculations, the RHS of (17) can be made ≤ 0 by guaranteeing that

(1) **Alignment:** $\|\Pi_V \nabla f(x)\| \geq \alpha \|\nabla f(x)\|$.

(2) **Small Margin:** $m_f(x) \leq \min(\alpha\varepsilon/2, \frac{\alpha^2 \|\nabla f(x)\|}{2L})$.

The requirement (1) holds because the subspace V is assumed to be (α, δ)-viable (see Definition 3). (2) is obtained from (1) and a careful analysis of (17). In particular, if $0 \leq \varepsilon \leq \alpha\beta/L$, then conditioned on $\|\nabla f(x)\| \geq \beta$ the "small margin" condition reduces to: $m_f(x) \leq \alpha\epsilon/2$. The full proof is given in the supplemental / appendix.

4.3 Applications: Some Special Cases of Theorem 3

We provide a non-exhaustive list of examples to illustrate the power of Theorem 3.

Linear Models. Proposition 1 which is a generalization of Lemma 2.2 of Guo (2020) is itself a special case of part (B) of Theorem 3. Indeed the linear function $f(x) \equiv x^\top w + b$ has margin $m_f(x) = (x^\top w + b)_+$ and verifies Conditions 2 and 3 with $\beta = \|w\|$, $L = 0$, and $\gamma = 0$. Also, thanks to Lemma 1, for any $k \in [d]$ and $t \in (0, \sqrt{k/d})$, a random k-dimensional subspace V of $\mathbb{R}^d$ is adversarially (α, δ)-viable with $\alpha = \sqrt{k/d} - t$ and $\delta = 2e^{-t^2 d/2}$. In Appendix 4.3, other lower-bounds established in Guo (2020) are recovered from Theorem 3 as special cases.

Hyper-Ellipsoids. We now generalize another result of Guo (2020), namely, Lemma 2.3 therein. Indeed, consider the case where $f(x) := (x^\top Bx - r^2)/2$, where B is a $d \times d$ positive semi-definite matrix and $r > 0$ is a scalar, so that the negative decision-region C of the classifier is the hyper-ellipsoid $f \leq 0$. In particular, C is a solid sphere of radius r when $B = I_d$. One computes $\nabla f(x) = Bx$, $\nabla^2 f(x) = B$, hence Conditions 2 and 3 are satisfied with $\gamma = 0$ and

$$L = \sup_{x \in \mathbb{R}^d} \|\nabla^2 f(x)\|_{op} = \|B\|_{op}, \quad (18)$$

$$\|\nabla f(x)\| = \|Bx\|, \text{ for all } x \in \mathbb{R}^d, \quad (19)$$

$$\beta = \inf_{x \in C'} \|\nabla f(x)\| = \inf_{x^\top Bx > r^2} \|Bx\| = r\sqrt{s_{min}}, \quad (20)$$

where s_{min} is the smallest singular / eigenvalue of B, and $\|B\|_{op}$ is the operator norm of B, i.e, its largest eigenvalue (since B is psd). Moreover, the margin of f at a any point $x \in \mathbb{R}^d$ is given by

$$m_f(x) = \frac{\max(f(x), 0)}{\|\nabla f(x)\|} = \frac{(x^\top Bx - r^2)_+}{2\|Bx\|}. \quad (21)$$

In particular, if $B = I_d$, then we deduce $L = 1$, $\beta = r$, Moreover, for any $x \in C'$, then the distance of x from C, i.e $d(x) = \|x\| - r$ and we have

$$m_f(x) = \frac{\|x\|^2 - r^2}{2\|x\|} = \frac{1}{2}(\|x\| - r)(1 + \frac{r}{\|x\|}). \quad (22)$$

Applying Theorem 3 with $B = I_d$ (corresponding to a solid sphere) then recovers exactly the bounds established in (Guo, 2020, Lemma 2.3) as a special case.

4.4 A Matching Upper-Bound under Convexity

We now show that the lower-bound given in Theorem 3 is tight by establishing a corresponding upper-bound for the case where C is convex (e.g., half-spaces, balls, hyper-ellipsoids, etc.).

Theorem 4. *Suppose f is convex differentiable, and let V be a subspace of $\mathbb{R}^d$ satisfying*

$$\|\Pi_V \eta(x)\| \leq \widetilde{\alpha}, \text{ for some } \widetilde{\alpha} \in [0, 1] \text{ and } \forall x \in C'. \quad (23)$$

Then, for any $\varepsilon \geq 0$, we have

$$\mathrm{FR}(V; \varepsilon) \leq \mathbb{P}_X(m_f(X) \leq \widetilde{\alpha}\varepsilon \mid X \in C'). \quad (24)$$

5 Locally Almost-Affine Decision-Boundary

We now consider the following smoothness condition for the classifier (1).

Condition 4 (Bounded oscillation of gradients)**.** *The exist $0 < R \leq \infty$ and $\theta \geq 0$ such that for all $x, \Delta x \in \mathbb{R}^d$ with $\|\Delta x\| \leq R$,*

$$\|\nabla f(x + \Delta x) - \nabla f(x)\| \leq \theta.$$

Examples of functions that satisfy this condition include half-spaces and wide feedforward ReLU neural nets with randomly initialized intermediate weights, where $\theta = o(1)$ w.h.p. over the intermediate weights, as will be seen in Section 5.2. The following is one of our main contributions.

Theorem 5. *Suppose Conditions 3 and 4 with parameters $\beta \in (0, \infty)$, $R \in (0, \infty]$ and $\theta \geq 0$ are in order. Let V be a possibly random adversarially (α, δ)-viable subspace of $\mathbb{R}^d$ with $\alpha > \theta/\beta$. Then, for any $0 \leq \varepsilon \leq R$, the average fooling rate of V is lower-bounded as follows*

$$\mathbb{E}_V \mathrm{FR}(V; \varepsilon) \geq \mathbb{P}_X(m_f(X) \leq \overline{\alpha}\varepsilon \mid X \in C') - \delta - \gamma, \quad (25)$$

where $\overline{\alpha} := \alpha - \theta/\beta > 0$.

Remark 2 (Tightness). *Theorem 5 is tight, as can be seen by considering the case where C is a half-space in which case $f(x) = x^\top w - b$, for some unit-vector $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$; take $V = \mathbb{R}w$. **N.B.:** $\nabla f(x) \equiv w$, and so Conditions 3 and 4 hold with $\alpha = \beta = 1$, $\theta = \gamma = 0$, and $R = \infty$.*

5.1 Sketch of Proof of Theorem 5

The core of the proof (detailed in the appendix) is similar to that of Theorem 3, but with (16) replaced by the following inequality which holds under Condition 4 for all $x \in \mathrm{supp}(P_X)$ and $\Delta x \in \mathbb{R}^d$ with $\|\Delta x\| \le R$

$$|f(x + \Delta x) - f(x) - \nabla f(x)^\top \Delta x| \le \theta \|\Delta x\|. \quad (26)$$

5.2 ReLU Neural Nets in Random Features Regime

Consider a feed-forward neural net with $M \ge 2$ layers with parameters matrices $W_1 \in \mathbb{R}^{d_0 \times d_1}, W_2 \in \mathbb{R}^{d_1 \times d_2}, \ldots, W_M = a \in \mathbb{R}^{d_{M-1} \times d_M}$, where $d_0 = d$ and $d_M := 1$. Each d_ℓ is the width of the ℓ layer, and the matrices $W_1, \ldots, W_{M-1}$ are the intermediate weights matrices, while $W_M = a$ is the output weights vector. For an input $x \in \mathbb{R}^d$, the output of the neural net is

$$\begin{aligned} f(x) = z_M &:= a^\top z_{M-1} \in \mathbb{R}, \text{ with } z_0 := x, \\ z_\ell &:= \sigma(W_\ell^\top z_{\ell-1}) \in \mathbb{R}^\ell, \forall \ell \in [M-1]. \end{aligned} \quad (27)$$

Here, σ is the *activation function*, and is applied entry-wise. We will focus on the case of ReLU neural nets, where $\sigma(t) \equiv (t)_+$. The matrices $W_1, \ldots, W_M$ are randomly initialized as follows: for all $\ell \in [M], i \in [d_\ell], j \in [d_{\ell-1}]$,

$$[W_\ell]_{i,j} \overset{iid}{\sim} N(0, 1/d_{\ell-1}). \quad (28)$$

The output weights vector $a \in \mathbb{R}^{d_{M-1}}$ can be arbitrary, for example: (1) random (as in Daniely and Shacham (2020); Bartlett et al. (2021)), or (2) optimized to fit training data, as in the so-called random features (RF) regime (Rahimi and Recht, 2008, 2009), with L_2-regularization on a. Let $d_{\min} := \min_{0 \le \ell \le M-1} d_\ell$ and $d_{\max} := \max_{0 \le \ell \le M-1} d_\ell$ be respectively, the minimum and maximum width of the layers. As in Bartlett et al. (2021), we will need the following technical condition.

Condition 5 (Genuinely wide, finite-width). *The neural net architecture verifies: (i) Bounded depth, i.e., only $M = \mathcal{O}(1)$ layers. (ii) Genuinely wide, i.e., $d_{\min} \gtrsim (\log d_{\max})^{40M}$ and $d_{\min} \to \infty$.*

We have the following corollary to Theorem 5.

Corollary 1. *Consider the case where the marginal distribution of the covariates X is supported on the sphere of radius $\sqrt{d}$ in $\mathbb{R}^d$, and f is the relu neural net defined in (27) with random intermediate weights $W_1, \ldots, W_{M-1}$ sampled according to (28). Suppose Conditions 5 is in order. Let V be a possibly random (α, δ)-viable subspace of $\mathbb{R}^d$, with $\alpha = \Omega(1)$. Then, for $0 \le \varepsilon \lesssim (\log d_{\max})^{40M}$, it holds w.h.p. over $W_1, \ldots, W_{M-1}$ that*

$$\mathbb{E}_V[\mathrm{FR}(V; \varepsilon)] \gtrsim \mathbb{P}_X(m_f(X) \le \varepsilon \mid X \in C') - \delta. \quad (29)$$

In particular, at initialization, we have $\mathbb{E}_V[\mathrm{FR}(V; \varepsilon)] \gtrsim 1 - \delta$ for all $\varepsilon \ge \varepsilon_0$, where ε_0 is an absolute constant.

The second part of the result implies that the subspace V contains adversarial perturbations of size $\sqrt{d}$ times smaller than the norm of a typical data point. Thus, it is a generalizes Daniely and Shacham (2020); Bartlett et al. (2021) to subspaces.

Proof of Corollary 1. The first part is obtained as a consequence of Theorem 5, by combining Lemma 2.2 and Lemma 2.8 of Bartlett et al. (2021) and Lemma 2 below.

The second part is because w.h.p over intermediate weights, it holds that

$$\begin{aligned} m_f(x) &\le |f(x)|/\|\nabla f(x)\| \\ &\lesssim \|a\| \|z_{L-1}(x)\|/\|a\| = \|z_{L-1}(x)\| \le \|x\|/\sqrt{d} = 1, \end{aligned}$$

where the last inequality is because $z_{L-1}(x)$ is $(\|x\|^2/d)$-subGaussian. □

Lemma 2. *Suppose P_X is supported on the sphere $\sqrt{d}\mathcal{S}_{d-1}$, and Condition 5 holds. Then, w.h.p. over the initialization of intermediate weights $W_1, \ldots, W_{M-1}$, the ReLU neural net f defined in (27) satisfies Conditions 3 and 4 with*

$$\gamma = 0, \; R = \frac{\sqrt{d_{min}}}{(\log d_{\max})^{80M}} = \Omega((\log d_{max})^{40M}),$$
$$\theta = \frac{\|a\|}{(\log d_{\max})^M}, \; \beta = \|a\|.$$

5.3 ReLU Neural Nets in Lazy Regime

At the moment, we are not able to extend our theoretical results to fully-trained neural nets. An exception is when the model is in the *lazy regime*, whereby the parameters of the network stay close to their value at definition. More, precisely

Definition 4 (Lazy regime). *The neural net (27) is said to be in the lazy regime if*

$$\sup_{j \in [d_\ell]} \frac{\|W_{\ell,j} - W^0_{\ell,j}\|_2}{\|W^0_{\ell,j}\|_2} \lesssim \frac{1}{\sqrt{d_\ell}} \text{ for all } \ell \in [M-1], \quad (30)$$

where W^0_ℓ is the initialization the ℓth layer.

Note that the lazy regime as defined above subsumes both relu neural nets at initialization and in the random features regime (studied in Section 5.2). Now, in Wang et al. (2022), it was shown that if $M = 2$ (i.e two-layer ReLU neural net),

then there exists absolute positive constants c_1, c_2, c_3, and c_4 that that: if the neural net is in the lazy regime, then w.h.p over the initialization, the following hold simultaneously for all $x \in \sqrt{d}\mathcal{S}_{d-1}$ and $\Delta x \in \mathbb{R}^d$ with $\|\Delta x\| \leq c_1$,

$$|f(x)| \leq c_2,\ \|\nabla f(x)\| \geq c_3,\ |\nabla f(x+\Delta x) - \nabla f(x)| \leq c_4.$$

See Lemma B.5, Lemma B.7, and Lemma B.9 (resp.) of Wang et al. (2022). We deduce that in the lazy regime, w.h.p over initialization, Conditions 3 and 4 hold with $R = c_1$ and $\beta = c_3$, $\theta = c_4$, and with $\gamma = 0$. On the same event, we also deduce the following margin bound

$$m_f(x) = \frac{(f(x))_+}{\|\nabla f(x)\|} \leq \frac{|f(x)|}{\|\nabla f(x)\|} \leq \frac{c_2}{c_3} =: c_5, \qquad (31)$$

for all $x \in \sqrt{d}\mathcal{S}_{d-1}$. Combining with Theorem 5, we obtain the following important corollary.

Corollary 2. *Suppose f defined in (27) is a two-layer network ($M = 2$) which is in the lazy regime. Also suppose the marginal distribution of the features X is supported on the sphere $\sqrt{d}\mathcal{S}_{d-1}$. If V is an adversarially (α, δ)-viable subspace of $\mathbb{R}^d$, then for any $0 \leq \varepsilon \leq R = c_1$ then w.h.p over the initial weights, the average fooling rate of V is lower-bounded as in (25), with $\beta = c_3$, $\theta = c_4$, and $\gamma = 0$.*

In particular, if $\varepsilon \in [c_5, c_1]$, then $\mathbb{E}_V \mathrm{FR}(V;\varepsilon) \geq 1 - \delta$.

6 Some Consequences of Our Results

Let us now outline some consequences for practical classifiers (neural networks). First, we recall the general form of our results. Given a possibly random adversarially (α, δ)-viable subspace V of $\mathbb{R}^d$, we have established in Theorem 3 and Theorem 5 lower-bounds on the fooling rate of the form

$$\mathbb{E}_V \mathrm{FR}(V;\varepsilon) \gtrsim \mathbb{P}(m_f(X) \leq \overline{\alpha}\varepsilon \mid X \in C') - \delta - \gamma. \quad (32)$$

Here, the scalar $\overline{\alpha} \in (0, 1]$ depends on α, β and the smoothness of f as in Condition 3. Importantly, the generic bound (32) explicitly highlights the dependence of the fooling rate on the pointwise margin of the classifier and on the alignment of the given subspace with the gradients of the f .

The L_2-norm N of a typical data point is of order $\sqrt{d}$, while the margin $m_f(X)$ is typically of order $O(1)$, as (i) observed empirically observed in Jiang et al. (2019) general trained neural networks (ii) formally proved in Daniely and Shacham (2020); Bartlett et al. (2021) for the case of relu networks at initialization and more recently, and in Wang et al. (2022) for the case of *lazy regime* where the intermediate parameters of the network stay close to their initial values throughout training (see (31)). Also, as observed in Moosavi-Dezfooli et al. (2017), the singular values of the gradient covariance matrix Σ_η are typically long-tailed. Thus, combining with Theorem 1, our results predict that for sufficiently large $k \ll d$, the subspace spanned by the top k singular-vectors of Σ_η has a nonzero fooling rate with attack budget $\varepsilon \asymp 1/\widetilde{\alpha} = O(1)$ which is $\sqrt{d}/\varepsilon = \Omega(\sqrt{d})$ times smaller than N, the L_2-norm of a typical data point, for relu neural networks in the lazy regime.

7 Empirical Verification

Our results are empirically verified in Figure 2 (random subspace attacks) and Figure 3 (singular subspace attack). Full details of the experimental setup and code for reproducing the results are provided in Appendix A.

Random Subspace Attacks. In Figure 2 (first and second row), the distribution P_X of the features is $N(0, I_d)$, and the training labels are given from a simple linear model: $y_i = x_{ij}$. For MNIST data (LeCun and Cortes, 2010) (third row), we construct a binary classification problem by restricting it to the digits 0 and 8. As in Guo et al. (2018), we run PGD Madry et al. (2017) attacks on a randomly chosen subspace V (of different dimensions) of the feature space $\mathbb{R}^d$, and report the fooling rates (solid lines) and compare them with our lower-bounds (32). As we can see from the figure, in all the cases, the lower-bounds are verified.

Eigen-Subspace Attacks. In Figure 3, we consider the same experimental setting in Figure 2. We use $n = 1000$ random examples $x_1, \ldots, x_n$, and compute the empirical covariance matrix of the gradient directions, i.e $\widehat{\Sigma}_\eta := \frac{1}{n-1}\sum_{i=1}^n (\eta_i - \overline{\eta})(\eta_i - \overline{\eta})^\top$, where $\eta_i := \eta(x_i)$, with $\overline{\eta} := (1/n)\sum_{i=1}^n \eta_i$. As in Khrulkov and Oseledets (2018), we extract the top eigenvector of $\widehat{\Sigma}_\eta$ and use it as a universal perturbation vector for a separate test set. In the leftmost subplot, we show a histogram of eigenvalues. Notice how the largest eigenvalue for each model is much larger than the other eigenvalues. Thanks to Theorem 1, this means that the principal eigenvector v spans an adversarially viable subspace. This is confirmed in the 2nd, 3rd, and 4th subplots where we see that the fooling rate rises rapidly as a function of the attack budget ε. We see from the figure that our predicted lower-bounds are satisfied in all cases.

Additional Experimental Results. In Appendix A, we provide experiments that empirically confirm our results on more complex models and datasets (like Resnet on CIFAR10), and also for adversarially trained models.

8 Concluding Remarks

In this work, we have conducted a rigorous analysis of the phenomenon of low-dimensional adversarial perturbations and derived tight lower-bounds for the fooling rate along arbitrary adversarial subspaces based on the geometry of the target decision-region, and the alignment between the

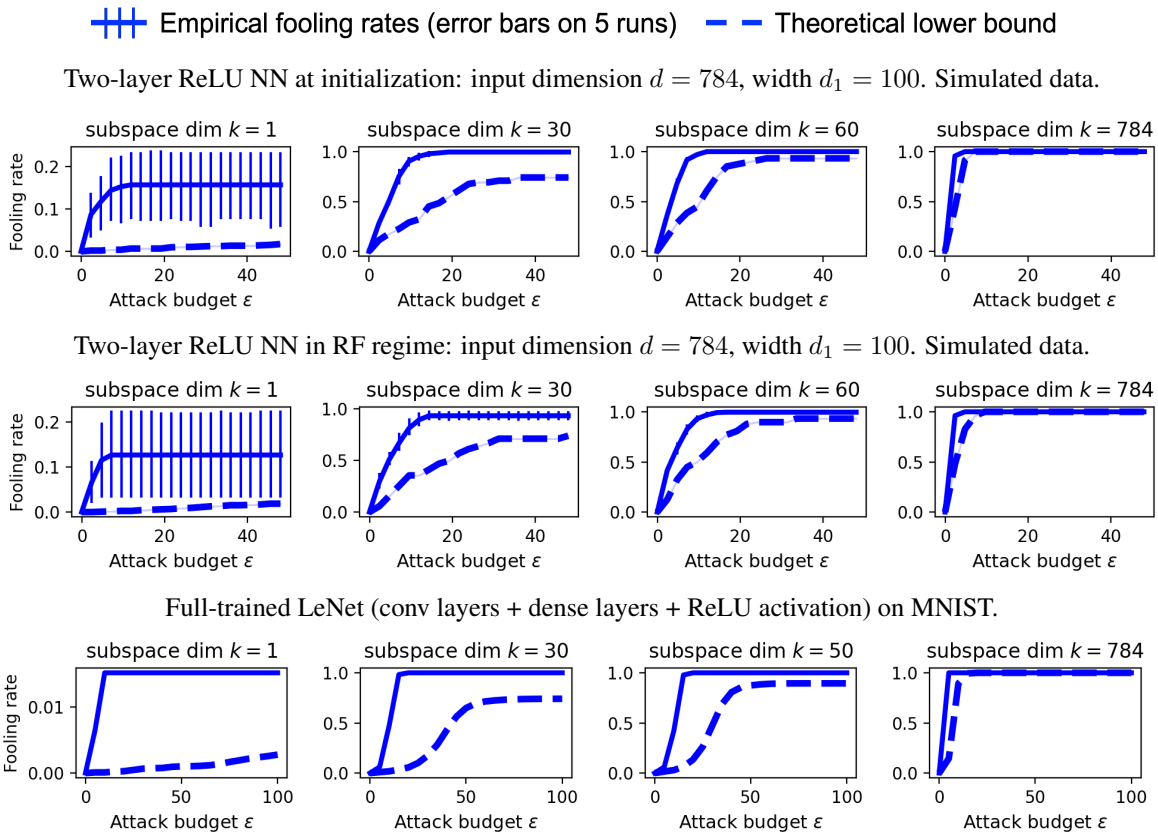


Figure 2: (Random subspace attack) Empirical confirmation of our results. Broken lines correspond to our theoretical lower-bounds (32), for different neural network regimes. k is the dimension of the random subspace from which the perturbations are constructed. In the first two rows, d_1 is the width of the network. Solid curves correspond to empirically computed fooling rates, with error-bars accounting for randomness in the initialization of the network, over 5 independent runs. Our theoretical lower bounds are confirmed in all cases. See Appendix A for details.

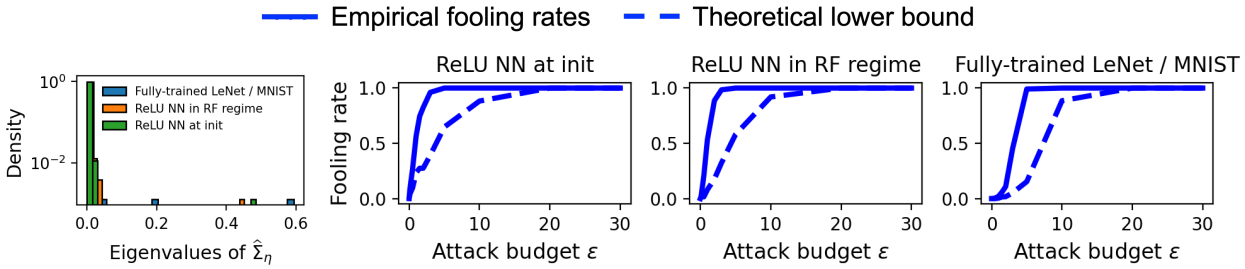


Figure 3: (Eigen-subspace attack). Same experimental setting in Figure 2. **Leftmost plot:** Showing a histogram of the eigenvalues of empirical covariance matrix $\widehat{\Sigma}_\eta$ of gradient directions (computed on 1000 examples). Notice how the largest eigenvalue for each model is much larger than the other eigenvalues. **Second to fourth (rightmost) plot:** Notice how the fooling rate rises rapidly. Further details are provided in Appendix A.

subspace and the gradients of the model, i.e., the adversarial viability of the subspace (Definition 3). Our work provides rigorous foundations for explaining intriguing empirical observations from the literature on the subject (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018; Yin et al., 2019; Guo et al., 2018). For the case of compact decision regions we have shown the existence of UAPs. We believe our work will further generate fruitful research in this area.

Finally, a non-trivial extension of our work would be the case of multi-class problems. It would also be interesting to extend our treatment of neural networks (Section 5.2) to general case, (i.e beyond the lazy regime). This would likely require the development of new theoretical tools.

Bibliography

Azé, D. and Corvellec, J.-N. (2017). Nonlinear error bounds via a change of function. *Journal of Optimization Theory and Applications*, 172. (Cited on 5)

Bartlett, P., Bubeck, S., and Cherapanamjeri, Y. (2021). Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34. (Cited on 5, 7, 8)

Brendel, W., Rauber, J., and Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*. (Cited on 1)

Bubeck, S., Cherapanamjeri, Y., Gidel, G., and des Combes, R. T. (2021). A single gradient step finds adversarial examples on random two-layers neural networks. In *Advances in Neural Information Processing Systems*. (Cited on 5)

Chen, J., Jordan, M. I., and Wainwright, M. J. (2020). Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE. (Cited on 1)

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26. (Cited on 1)

Daniely, A. and Shacham, H. (2020). Most relu networks suffer from \ell^2 adversarial perturbations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6629–6636. Curran Associates, Inc. (Cited on 5, 7, 8)

Dohmatob, E. (2019). Generalized no free lunch theorem for adversarial robustness. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*. PMLR. (Cited on 2, 3)

Fawzi, A., Fawzi, H., and Fawzi, O. (2018). Adversarial vulnerability for any classifier. *CoRR*, abs/1802.08686. (Cited on 3)

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. J. (2018). Adversarial spheres. *CoRR*, abs/1801.02774. (Cited on 2)

Guo, C. (2020). *Phd thesis: Threats and Countermeasures in Machine Learning Applications*. Cornell University. (Cited on 2, 3, 4, 6)

Guo, C., Frank, J. S., and Weinberger, K. Q. (2018). Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*. (Cited on 1, 3, 8, 9, 12)

Guo, C., Gardner, J., You, Y., Wilson, A. G., and Weinberger, K. (2019). Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR. (Cited on 1, 2, 3)

Huang, Z. and Zhang, T. (2019). Black-box adversarial attack with transferable model-based embedding. *arXiv preprint arXiv:1911.07140*. (Cited on 1, 3)

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. (2018). Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR. (Cited on 1)

Jiang, Y., Krishnan, D., Mobahi, H., and Bengio, S. (2019). Predicting the generalization gap in deep networks with margin distributions. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net. (Cited on 8)

Khrulkov, V. and Oseledets, I. (2018). Art of singular vectors and universal adversarial perturbations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8562–8570. (Cited on 1, 2, 3, 5, 8, 9, 12)

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. (Cited on 8, 12)

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*. (Cited on 8, 12)

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. (2018). The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *CoRR*, abs/1809.03063. (Cited on 2)

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. (2019). The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4536–4543. (Cited on 3)

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94. (Cited on 1, 2, 3, 4, 5, 8, 9, 12)

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P., and Soatto, S. (2018). Analysis of universal adversarial perturbations. abs/1705.09554. (Cited on 2, 4)

Rahimi, A. and Recht, B. (2008). Uniform approximation of functions with random bases. (Cited on 7)

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. (Cited on 7)

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. (2018a). Are adversarial examples inevitable? *CoRR*, abs/1809.02104. (Cited on 2)

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. (2018b). Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*. (Cited on 3)

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. (Cited on 1)

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2019). Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, volume abs/1805.12152. (Cited on 2)

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. (2019). Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749. (Cited on 1, 3)

Wang, Y., Ullah, E., Mianjy, P., and Arora, R. (2022). Adversarial robustness is at odds with lazy training. NeurIPS. (Cited on 7, 8)

Yan, Z., Guo, Y., and Zhang, C. (2019). Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *arXiv preprint arXiv:1906.04392*. (Cited on 1, 3)

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. (2019). A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*. (Cited on 1, 9)

A More Information About the Experiments

A.1 Details of Experimental Setup

Our theoretical results are empirically verified in Figure 2 (random subspace attacks) and Figure 3 (singular subspace attack). We now provide experimental details on these figures.

Given a binary classifier $h_f : x \mapsto \text{sign}(f(x))$ on $\mathbb{R}^d$ (e.g a neural net), with negative decision-region $C := \{x \in \mathbb{R}^d \mid h_f(x) = -1\}$. For a subspace $V \subseteq \mathbb{R}^d$ and a (Euclidean) attack budget ε, Refer to Definition 2 for the fooling rate of V on the classifier h_f.

Figure 2: Results for Random Subspace Attacks. In Figure 2 (first and second row), the distribution P_X of the features is $N(0, I_d)$, and the training labels are given from a simple linear model: $y_i = x_{ij}$. For classical LeNet convolutational neural network trained on MNIST data (LeCun and Cortes, 2010) (third row), we construct a binary classification dataset of $n = 2 \times 10\text{K} = 20\text{K}$ examples by restricting it to the digits 0 and 8. As in Guo et al. (2018), we run PGD Madry et al. (2017) attacks on a randomly chosen subspace V (of different dimensions $k \leq d$) of the feature space $\mathbb{R}^d$, and report the fooling rates (solid lines) and compare them with our proposed lower-bounds (32), from Theorem 5 with $R = \infty$ and $\theta = 0$ (these extremal values work for our experiments). As can be seen from the figure, in all the cases, the lower-bounds (broken lines) are verified.

Figure 3: Results for Attacks Based on Eigen-subspaces of Gradients. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be iid samples from the conditional distribution $\mathbb{P}_{X|X\in C'}$, the distribution of the data conditioned on the positive decision-region of the classifier, and let J be the $n \times d$ matrix with ith row given by $\eta(x_i) := \nabla f(x_i)/\|\nabla f(x_i)\| \in \mathcal{S}_{d-1}$. Moosavi-Dezfooli et al. (2017); Khrulkov and Oseledets (2018) have provided strong empirical evidence that the subspace spanned by the first top eigenvectors of the matrix of $\widehat{\Sigma}_\eta := J^\top J/n$ contains successful adversarial perturbations. In fact, the one-dimensional subspace spanned by the top eigenvector of $\widehat{\Sigma}_\eta$ was shown in Khrulkov and Oseledets (2018) to achieve state-of-the-art performance, on a variety of models and datasets.

Theorem 1 provides a rigorous justification for the success of these SVD-based heuristics used in Moosavi-Dezfooli et al. (2017); Khrulkov and Oseledets (2018) to compute UAPs. This is empirically verified in Figure 3, where we consider the same experimental setting (dataset and model) in Figure 2. We use $n = 1000$ random examples $x_1, \ldots, x_n$, and compute the empirical covariance matrix $\widehat{\Sigma}_\eta := (n-1)^{-1} \sum_{i=1}^n (\eta_i - \overline{\eta})(\eta_i - \overline{\eta})^\top$, of the gradient directions $\eta_i := \eta(x_i)$, with $\overline{\eta} := (1/n) \sum_{i=1}^n \eta_i$. As in Khrulkov and Oseledets (2018), we extract the top eigenvector of $\widehat{\Sigma}_\eta$ and use it as a universal perturbation vector for a separate test set. In the leftmost subplot, we show a histogram of eigenvalues. Notice how the largest eigenvalue for each model is much larger than the other eigenvalues. Thanks to Theorem 1, this means that the principal eigenvector v spans an adversarially viable subspace. This is confirmed in the 2nd, 3rd, and 4th subplots where we see that fooling rate rises rapidly as a function of the attack budget ε. We see from the figure that our predicted lower-bounds shown in broken lines (computed analogously to the case of Figure 2 described above) are satisfied in all cases.

Here the dataset is constructed by transforming MNIST into 10 one-versus-all binary classification problems.

Remark 3. *We ignore issues concerning the consistency of approximating the principal eigenvector Σ_η with that of $\widehat{\Sigma}_\eta$, used in practice (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018).*

Remark 4. *The gap in Figures 2 and 3 between experiments (solid curves) and our theoretical results (broken ones) is due to the fact that our established lower-bounds (32), though sufficient to explain the success of low-dimensional adversarial perturbations, might be too conservative for obtaining exact quantitative estimates for the fooling rate in the case of random adversarial subspaces on neural nets, because we only use first-order (see Conditions 2, 3, and 4) information on the neural net f. However, in the specific scenario where the target decision-region is a half-space or hyper-ellipsoid, this gap disappears because the aforementioned first-order information is sufficient in such cases, and our estimates for fooling rate are exact.*

A.2 Additional Experimental Results

At the moment, we are not able to extend our theoretical results to fully-trained NNs. An exception is when the model is in the lazy regime, as shown in Section 5.3. A rigorous study of the general case is left for future work. That not withstanding, we empirically observe that our results remain valid both on normally and adversarially trained (AT) NNs (see Figure 4),

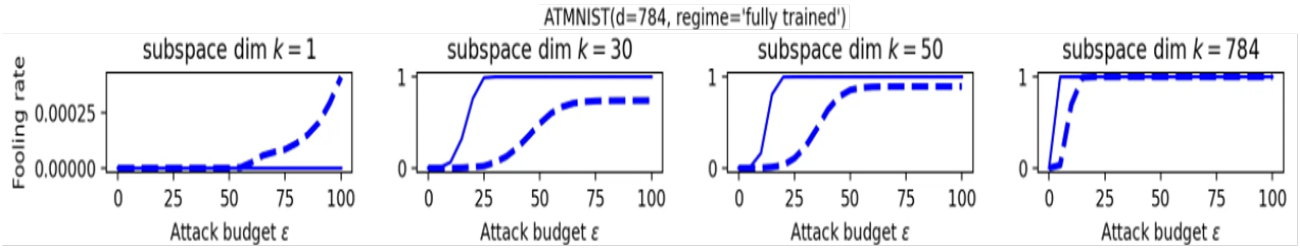


Figure 4: **Results for adversarially trained model.** We consider a LeNet convolutional neural network on MNIST dataset, learned via adversarial training.

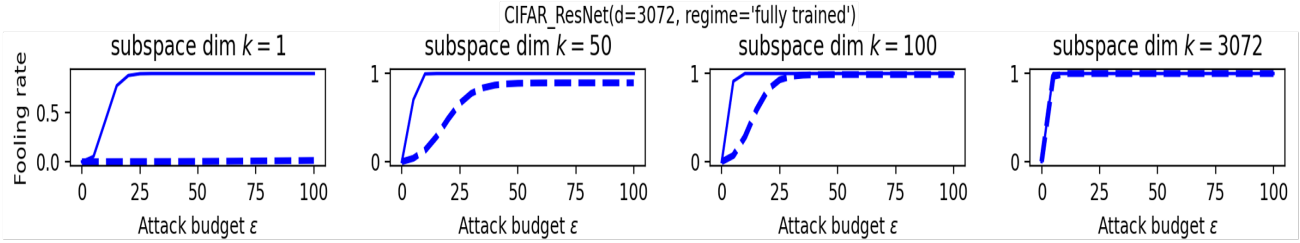


Figure 5: **Results on larger models.** Here we consider Resnet on CIFAR10 dataset.

and on more complicated NNs / datasets like Resnet on CIFAR10 (see Figure 5). Comparing with the last row of Figure 2, notice how AT slightly helps to slightly decrease the fooling rate.

B Proof of Theorem 1: Adversarial Viability of Eigenspaces of Unit-Normals

Theorem 1. *For any $k \in [d]$, let $s_k \in (0,1]$ be the sum of first the k eigenvalues of Σ_η. Then, for any $\alpha \in (0, \sqrt{s_k})$, the (deterministic) subspace $V_{\mathrm{eigen},k}$ of $\mathbb{R}^d$ corresponding to the top k eigendirections of Σ_η is adversarially $(\alpha, (1-s_k)/(1-\alpha^2))$-viable.*

Proof of Theorem 1. Let $\Sigma_\eta = USU^\top$ be the SVD of Σ_η, where S is a diagonal matrix containing the nonzero eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ of Σ_η, $r \in [d]$ is the rank of Σ_η, and U is a $d \times r$ matrix with orthonormal columns. Then, the orthogonal projector for the subspace $V := V_{\mathrm{eigen},k}$ is given explicitly by $\Pi_V = U_{\leq k} U_{\leq k}^\top$, where $U_{\leq k}$ is the $d \times \min(k,r)$ orthogonal matrix corresponding to the first $\min(k,r)$ columns of U. Consider the r.v $Z := \|\Pi_V \eta(X)\|$. By a standard formula for the expectation of a quadratic form, one computes

$$\begin{aligned} \mathbb{E}\left[Z^2 \mid X \in C'\right] &= \mathbb{E}[\eta(X)^\top \Pi_V \eta(X) \mid X \in C'] \\ &= \mathrm{tr}(\Pi_V \Sigma_\eta) = \mathrm{tr}(U_{\leq k} U_{\leq k}^\top \Sigma_\eta) \\ &= \mathrm{tr}(U_{\leq k}^\top \Sigma_\eta U_{\leq k}) = \sum_{i=1}^{\min(k,r)} \lambda_i =: s_k. \end{aligned} \tag{33}$$

On the other hand, conditioned on $X \in C'$ we have $0 \leq Z \leq \|\eta(X)\|$. Thus, for any $\alpha \in (0, \sqrt{s_k})$, we have

$$X \in C' \implies \mathbb{1}(Z \geq \alpha) \geq (Z^2 - \alpha^2)/(1-\alpha^2), \tag{34}$$

with equality on the event $Z^2 \in \{\alpha^2, 1\}$. The claim then follows upon taking expectations on both sides of the above inequality conditioned on the event $X \in C'$. □

C Lower-Bound under Smoothness Assumptions

C.1 Auxiliary Lemmas

Lemma 3. *For any $\rho, r > 0$ and $b \in \mathbb{R}^d$, we have the identity*

$$\sup_{z \in \rho B_n} b^\top z - \frac{1}{2r}\|z\|^2 = \begin{cases} r\|b\|^2/2, & \text{if } \|b\| \le \rho/r, \\ \rho\|b\| - \rho^2/(2r), & \text{otherwise.} \end{cases} \tag{35}$$

Proof. Since the quadratic function $z \mapsto (1/2)\|z\|^2$ is unchanged upon taking the *Fenchel-Legendre transform*, we have

$$\begin{aligned}
\sup_{z \in \rho B_d} b^\top z - \frac{1}{2r}\|z\|^2 &= \sup_{\|z\| \le \rho} b^\top z - \frac{1}{r}\left(\sup_{u \in \mathbb{R}^d} z^\top u - \frac{1}{2}\|u\|^2\right) \\
&\overset{(*)}{=} \inf_{u \in \mathbb{R}^d}\left(\frac{1}{2r}\|u\|^2 + \sup_{\|z\| \le \rho} z^\top (b - u/r)\right) \\
&= \inf_{u \in \mathbb{R}^d}\left(\frac{1}{2r}\|u\|^2 + \rho\|b - u/r\|\right) \\
&= \inf_{v \in \mathbb{R}^d}\left(\frac{r}{2}\|v - b\|^2 + \rho\|v\|\right), \text{ by change of variable } v := b - u/r \\
&= \rho \inf_{v \in \mathbb{R}^d}\left(\frac{1}{2\rho/r}\|v - b\|^2 + \|v\|\right), \text{ by factoring out } \rho \\
&\overset{(**)}{=} \rho \begin{cases} \|b\|^2/(2\rho/r), & \text{if } \|b\| \le \rho/r, \\ \|b\| - \rho/(2r), & \text{else} \end{cases} \\
&= \begin{cases} r\|b\|^2/2, & \text{if } \|b\| \le \rho/r, \\ \rho\|b\| - \rho^2/(2r), & \text{else,} \end{cases}
\end{aligned}$$

where $(*)$ uses *Sion's Minimax Theorem*, and in $(**)$ we have recognized a rescaled *Moreau envelope* of the Euclidean norm, which is the Huber function evaluated at $\|b\|$. □

We will also need the following auxiliary lemma.

Lemma 4. *For any $r, \rho > 0$ and $b \in \mathbb{R}^d$, we have the identity*

$$\sup_{z \in \rho B_n} b^\top z - \frac{1}{r}\|z\| = \rho(\|b\| - 1/r)_+. \tag{36}$$

Proof. By direct computation, we have

$$\begin{aligned}
\sup_{\|z\| \le \rho} b^\top z - \frac{1}{r}\|z\| &= \sup_{\|z\| \le \rho} b^\top z - \sup_{\|u\| \le 1} z^\top u/r \\
&= \inf_{\|u\| \le 1} \sup_{\|z\| \le \rho} z^\top (b - u/r) \\
&= \rho \inf_{\|u\| \le 1} \|b - u/r\| \\
&= \rho(\|b\| - 1/r)_+,
\end{aligned}$$

we in the last step, we have recognized the well-known block *soft-thresholding* operator. □

Finally, we will need the following lemma.

Lemma 5. *Suppose R_1, R_2, R_3 are random variables and $\phi : \mathbb{R} \to [-\infty, \infty]$ is a possibly random nondecreasing function. If $\mathbb{P}(R_2 \ge R_3) \ge 1 - \delta$*

$$\mathbb{P}(R_1 \le \phi(R_2)) \ge \mathbb{P}(R_1 \ge \phi(R_3)) - \delta. \tag{37}$$

Proof. Indeed, consider the events $E_1 := \{R_1 \leq \phi(R_3)\}$, $E_2 := \{R_3 \leq R_2\}$, $E_3 := E_1 \cap E_2$ and $E_4 := \{R_1 \leq \phi(R_2)\}$. It is clear that $E_3 \subseteq E_4$. One then easily computes

$$\begin{aligned}\mathbb{P}(R_1 \leq \phi(R_2)) = \mathbb{P}(E_4) &\geq \mathbb{P}(E_3) = \mathbb{P}(E_1 \cap E_2) \\ &= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cup E_2) \\ &\geq \mathbb{P}(E_1) + \mathbb{P}(E_2) - 1 \\ &\geq \mathbb{P}(E_1) - \delta \\ &= \mathbb{P}(R_1 \leq \phi(R_3)) - \delta,\end{aligned}$$

as claimed. □

C.2 Proof of Theorem 3: Lipschitz Decision-Boundary

We are now ready to prove Theorem 3. First, we restate it for convenience

Theorem 3. *Suppose Condition 2 holds. Let V be a possibly random adversarially (α, δ)-viable subspace of $\mathbb{R}^d$. Then,*

(A) For any $\varepsilon \geq 0$, the average fooling rate of V is lower-bounded as follows

$$\mathbb{E}_V[\mathrm{FR}(V;\varepsilon)] \geq \mathbb{P}_X\left(m_f(X) \leq \widetilde{\alpha}(X) \mid X \in C'\right) - \delta, \tag{14}$$

where $\widetilde{\alpha}(X) := \min\left(\alpha\varepsilon/2, \alpha^2\|\nabla f(X)\|/(2L)\right)$. (B) If in addition Condition 3 is in order, then for any $\varepsilon \in [0, \alpha\beta/L]$,

$$\begin{aligned}\mathbb{E}_V \mathrm{FR}(V;\varepsilon) \geq \mathbb{P}_X(m_f(X) \leq \alpha\varepsilon/2 \mid X \in C') \\ - \delta - \gamma.\end{aligned} \tag{15}$$

Proof of Theorem 3. Let $x \in C' := \mathbb{R}^d \setminus C$ and set $v(x) := \Pi_V \nabla f(x)/\|\Pi_V \nabla f(x)\| \in \mathcal{S}_{d-1} \cap V$. Define $p_V(x) := \|\Pi_V \nabla f(x)\|$, the L_2-norm of the orthogonal projection of the gradient vector $\nabla f(x)$ onto the subspace V. It is clear that $\nabla f(x)^\top v(x) = \|\Pi_V \nabla f(x)\| = p_V(x)$. Let $d_V(x) \in (0, \infty]$ be the distance of x from C along the subspace V (see (48)). By definition, $d_V(x)$ is no larger than the distance between x and the point where the line $x + \mathbb{R}v(x) := \{x + sv(x) \mid s \in \mathbb{R}\}$ first meets C (if it meets it at all!). Thus, with the convention $\inf \emptyset = \infty$, we have

$$\begin{aligned}d_V(x) &\leq \inf_{s\in\mathbb{R}} |s| \text{ subject to } x + sv(x) \in C \\ &= \inf_{s\in\mathbb{R}} |s| \text{ subject to } f(x + sv(x)) \leq 0 \\ &\leq \inf_{s\in\mathbb{R}} |s| \text{ subject to } f(x) + s\nabla f(x)^\top v(x) + Ls^2/2 \leq 0 \\ &= \inf_{s\in\mathbb{R}} |s| \text{ subject to } f(x) + p_V(x)s + Ls^2/2 \leq 0,\end{aligned} \tag{38}$$

where we have invoked the RHS of (16) with $x' = x + sv(x)$ to arrive at the third line.

$$f(x) \geq \sup_{|s|<d_V(x)} -p_V(x)s - Ls^2/2 = \begin{cases} p_V(x)^2/(2L), & \text{if } p_V(x) \leq Ld_V(x), \\ p_V(x)d_V(x) - Ld_V(x)^2/2, & \text{otherwise,}\end{cases} \tag{39}$$

where the second step is an application of Lemma 3 with $n = 1$, $b = -p_V(x)$, $r = 1/L$ and $\rho = d_V(x)$. Now, if $f(x) < p_V(x)^2/(2L)$, we deduce from (39) that $d_V(x) < p_V(x)/L$ and $f(x) \geq p_V(x)d_V(x) - Ld_V(x)^2/2$ (see Figure 6 for geometric intuition), and so

$$\begin{aligned}d_V(x) &\leq p_V(x)/L - \sqrt{(p_V(x)/L)^2 - 2f(x)/L} \\ &= \frac{2f(x)}{p_V(x) + \sqrt{p_V(x)^2 - 2f(x)L}} \\ &\leq \frac{2f(x)}{p_V(x)} = \frac{2m_f(x)}{\alpha_V(x)},\end{aligned} \tag{40}$$

where $\alpha_V(x) = p_V(x)/\|\nabla f(x)\| = \|\Pi_V \nabla f(x)\|/\|\nabla f(x)\| = \|\Pi_V \eta(x)\|$. Now, because $C_V^\varepsilon = \{x \in \mathbb{R}^d \mid d_V(x) \leq \varepsilon\}$, we deduce that

$$\left\{x \in C' \mid m_f(x) \leq \min\left(\frac{\alpha_V(x)\varepsilon}{2}, \frac{\alpha_V(x)^2\|\nabla f(x)\|}{2L}\right)\right\} \subseteq C_V^\varepsilon \setminus C. \tag{41}$$

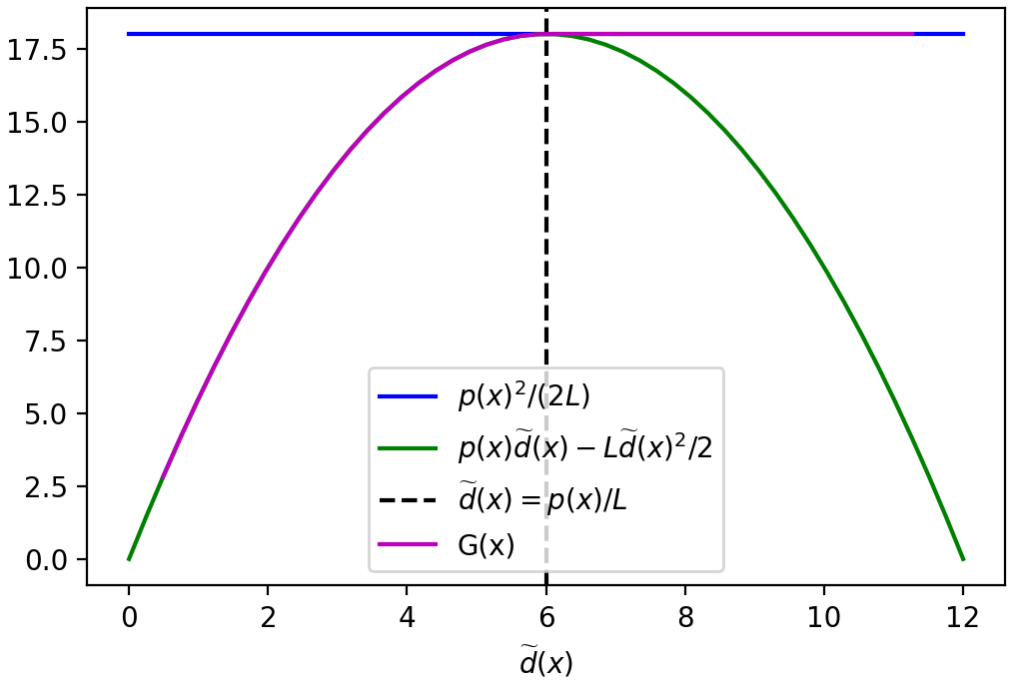


Figure 6: Graphical illustration of the RHS of (39), denote here as $G(x)$. In this illustration, $p(x) = p_V(x)$ and L are fixed to 5 and 1 respectively. Here, $\widetilde{d}(x)$ is shorthand for $d_V(x)$, the distance of x from C along the subspace V.

Now, define $s_V(x) := \alpha_V(x)^2\|\nabla f(x)\|/(2L)$ and $s(x) := \alpha^2\|\nabla f(x)\|/(2L)$. Since the subspace V is an adversarial (α, δ)-viable by hypothesis, it follows from Definition 3 that

$$\begin{aligned}\mathbb{P}_{X,V}(\min(\alpha_V(X)\epsilon/2, s_V(X)) \geq \min(\alpha\epsilon/2, s(X)) \mid X \in C') \\ \geq \mathbb{P}_{X,V}(\|\Pi_V \eta(X)\| \geq \alpha \mid X \in C') \geq 1 - \delta.\end{aligned} \tag{42}$$

The *Fubini-Tonelli Theorem* then gives,

$$\begin{aligned}\mathrm{FR}(V;\varepsilon) &:= \mathbb{E}_V \mathbb{P}_X(X \in C_V^\varepsilon \mid X \in C') = \mathbb{E}_X \mathbb{P}_V(X \in C_V^\varepsilon \mid X \in C') \\ &\geq \mathbb{E}_X \mathbb{P}_V(m_f(X) \leq \min(\alpha_V(X)\varepsilon/2, s_V(X)) \mid X \in C') \\ &\geq \mathbb{E}_X \mathbb{P}_V(m_f(X) \leq \min(\alpha\varepsilon/2, s(X)) \mid X \in C') - \delta,\end{aligned}$$

where the last step is thanks to Lemma 5 with $R_1 = m_f(X)$, $R_2 = \min(\alpha_V(x)\epsilon/2, s_V(X))$, $R_3 = \min(\alpha\epsilon/2, s(X))$, and $\phi = Id$, and recalling (42). This proves the first part of the theorem.

For the second part, Condition 3 is in order and so we have $\mathbb{P}(\|\nabla f(X)\| \geq \beta \mid X \in C') \geq 1 - \gamma$. On the other hand, if $0 \leq \epsilon \leq \alpha\beta/L$, then conditioned on the event $\|\nabla f(x)\| \geq \beta$, we have $\min(\alpha\epsilon/2, s(X)) \geq \min(\alpha\epsilon/2, \alpha^2\beta/(2L)) = \alpha\epsilon/2$, and the result follows from the first part and Lemma 5. □

C.3 Proof of Theorem 5: Locally Affine Decision-Boundaries

Theorem 5. *Suppose Conditions 3 and 4 with parameters $\beta \in (0, \infty)$, $R \in (0, \infty]$ and $\theta \geq 0$ are in order. Let V be a possibly random adversarially (α, δ)-viable subspace of $\mathbb{R}^d$ with $\alpha > \theta/\beta$. Then, for any $0 \leq \varepsilon \leq R$, the average fooling rate of V is lower-bounded as follows*

$$\mathbb{E}_V \mathrm{FR}(V;\varepsilon) \geq \mathbb{P}_X(m_f(X) \leq \overline{\alpha}\varepsilon \mid X \in C') - \delta - \gamma, \tag{25}$$

where $\overline{\alpha} := \alpha - \theta/\beta > 0$.

Proof of Theorem 5. Under Condition 4, it is easy to establish the classical inequality

$$-\theta\|x' - x\| \leq f(x') - f(x) - \nabla f(x)^\top (x' - x) \leq \theta\|x' - x\|, \text{ for all } \|x' - x\| \leq R. \tag{43}$$

Now, let $x \in C' := \mathbb{R}^d \setminus C$ and let $d_V(x)$ be the distance of x from V along the subspace V. Let $v(x)$, $p_V(x)$, $\alpha_V(x)$, $s_V(x)$, and $s(x)$ be as defined in the proof of Theorem 3. By an argument analogous to the beginning of the proof of Theorem 3 but with (43) used in place of (16) and the restriction that $|s| \leq R$ so that (43) is valid for every x' on the line $x + \mathbb{R}v(x)$, it is straightforward to establish that

$$\begin{aligned} d_V(x) &\leq \inf_{s\in\mathbb{R}} |s| \text{ subject to } x + sv(x) \in C,\ |s| \leq R \\ &\leq \inf_{s\in\mathbb{R}} |s| \text{ subject to } f(x) + p_V(x)s + \theta|s| \leq 0,\ |s| \leq R \\ &\leq \inf_{s\in\mathbb{R}} |s| \text{ subject to } f(x) + p_V(x)s + \theta|s| \leq 0,\ |s| \leq R. \end{aligned} \tag{44}$$

We deduce that

$$f(x) \geq \sup_{|s|<\min(d_V(x),R)} -p_V(x)s - \theta|s| = \min(d_V(x), R) \cdot (p_V(x) - \theta)_+, \tag{45}$$

where the equality is thanks to Lemma 4 applied with $n = 1$, $b = -p_V(x)$, $r = 1/\theta$, and $\rho = \min(d_V(x), R)$. Thus, we deduce from (45) that

$$\min(d_V(x), R) \leq \frac{f(x)}{(p_V(x) - \theta)_+} = c_V(x) m_f(x), \tag{46}$$

with $c_V(x) := \|\nabla f(x)\|/(\alpha_V(x)\|\nabla f(x)\| - \theta)_+$. One the event $\alpha_V(x) \geq \alpha > \theta/\beta$, we have $1/c_V(x) \geq \overline{\alpha} := \alpha - \theta/\beta$. Thus, if $m_f(x) \leq \overline{\alpha}\varepsilon$ and $0 \leq \varepsilon < R$, then $d_V(x) \leq \varepsilon$. That is, if $0 \leq \epsilon < R$

$$\{x \in C' \mid m_f(x) \leq \overline{\alpha}\varepsilon\} \subseteq C_V^\varepsilon \setminus C. \tag{47}$$

The rest of the proof is analogous to the end of the proof of the first part of Theorem 3 (starting from the set-inclusion (41)), and is thus omitted. □

D Proof of Theorem 4: A Matching Upper-Bound under Convexity

Theorem 4. *Suppose f is convex differentiable, and let V be a subspace of $\mathbb{R}^d$ satisfying*

$$\|\Pi_V \eta(x)\| \leq \widetilde{\alpha}, \text{ for some } \widetilde{\alpha} \in [0, 1] \text{ and } \forall x \in C'. \tag{23}$$

Then, for any $\varepsilon \geq 0$, we have

$$\mathrm{FR}(V; \varepsilon) \leq \mathbb{P}_X(m_f(X) \leq \widetilde{\alpha}\varepsilon \mid X \in C'). \tag{24}$$

Proof. Let $d(x) \in [0, \infty)$ be the distance of x from C and let $d_V(x) \in [0, \infty]$ be the distance of x from C along the subspace V, i.e.,

$$\begin{aligned} d(x) &:= \inf_{v\in\mathbb{R}^d} \|v\| \text{ subject to } x + v \in C, \\ d_V(x) &:= \inf_{v\in V} \|v\| \text{ subject to } x + v \in C, \end{aligned} \tag{48}$$

with the convention that $\inf \emptyset = \infty$. By definition of the (ε, V)-expansion C_V^ε of C (refer to (3), we have

$$C_V^\varepsilon = \{x \in \mathbb{R}^d \mid d_V(x) \leq \varepsilon\}. \tag{49}$$

Also, it is clear that $d_V(x) \geq d(x)$, attained when $V = \mathbb{R}^d$. By definition of $d_V(x)$, it is clear that $x - d_V(x)v \in C$, where $v = \Pi_V \nabla f(x)/\|\Pi_V \nabla f(x)\|$. Observe that $\nabla f(x)^\top v = \|\Pi_V \nabla f(x)\|$. Now, thanks to the convexity of f, we have

$$f(x') \geq f(x) + \nabla f(x)^\top (x' - x), \tag{50}$$

for all $x' \in \mathbb{R}^d$. Thus,

$$\begin{aligned} x - d_V(x)v \in C &\implies f(x - d_V(x)v) \leq 0 \\ &\implies f(x) - d_V(x)\nabla f(x)^\top v \leq 0 \text{ thanks to (50)} \\ &\implies m_f(x) \leq \frac{d_V(x)\nabla f(x)^\top v}{\|\nabla f(x)\|} \leq \frac{d_V(x)\|\Pi_V \nabla f(x)\|}{\|\nabla f(x)\|} \leq \widetilde{\alpha} d_V(x). \end{aligned}$$

We deduce that $\{x \in C' \mid m_f(x) \leq \widetilde{\alpha}\varepsilon\} \supseteq \{x \in C' \mid d_V(x) \leq \varepsilon\} =: C_V^\varepsilon \setminus C$, and the result follows. □