Adversarial Robustness is at Odds with Lazy Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent works show that random neural networks are vulnerable against adversarial attacks [Daniely and Schacham, 2020] and that such attacks can be easily found using a single step of gradient descent [Bubeck et al., 2021]. In this work, we take it one step further and show that a single gradient step can find adversarial examples for networks trained in the so-called lazy regime. This regime is interesting because even though the neural network weights remain close to the initialization, there exist networks with small generalization error, which can be found efficiently using first-order methods. Our work challenges the model of the lazy regime, the only regime in which neural networks are provably efficiently learnable. We show that the networks trained in this regime, even though they enjoy good theoretical computational guarantees, remain vulnerable to adversarial examples. In doing so, we resolve an open question posed by the work of Bubeck et al. [2021], who show a similar result for random neural networks. To the best of our knowledge, this is the first work to prove that such well-generalizable neural networks are still vulnerable to adversarial attacks.

1 Introduction

Despite the tremendous success of deep learning, recent works have demonstrated that neural networks are extremely susceptible to attacks. One such vulnerability is due to arbitrary adversarial corruption of data at the time of prediction, commonly referred to as *inference time attack* [Biggio et al., 2013, Szegedy et al., 2013]. Such attacks are catastrophically powerful, especially in settings where adversaries can directly access the model parameters. By adding crafted imperceptible perturbations on the natural input, such attacks are easily realized across various domains, including in computer vision where an autonomous driving car may misclassify traffic signs [Sitawarin et al., 2018], in natural language processing where automatic speech recognition misinterprets the sentence's meaning [Schönherr et al., 2018], among many others.

The potential threat of adversarial examples has lead to serious concerns regarding the security and trustworthiness of neural network-based models that are increasingly being deployed in real-world systems. It is crucial, therefore, to understand why trained neural networks classify clean data with high accuracy yet remain extraordinarily fragile due to strategically induced perturbations.

One of the earliest works to describe adversarial examples was that of Szegedy et al. [2013], motivated by the need for networks that not only generalize well but are also robust to small perturbations of its input. This was followed by a body of works that finds adversarial examples using various methods [Goodfellow et al., 2014, Papernot et al., 2016, Madry et al., 2017, Carlini and Wagner, 2017a,b]. Although lots of defense algorithms have been designed, most of them are quickly defeated by stronger attacks [Carlini and Wagner, 2017a, Carlini et al., 2019, Tramer et al., 2020]. Much of the prior work has focused on this empirical "arms race" between adversarial defenses and attacks; yet, a deep theoretical understanding of why adversarial examples exist has been somewhat limited.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Recently, a line of work theoretically constructs adversarial attacks for random neural networks. Inspired by the work of Shamir et al. [2019], Daniely and Schacham [2020] prove that small ℓ_2 norm adversarial perturbations can be found by multi-step gradient descent method for random ReLU networks with small width – each layer has vanishing width compared to the previous layer. Bubeck et al. [2021] generalize this result to two-layer randomly initialized networks with relatively large network width and show that a single step of gradient descent suffices to find adversarial examples. Bartlett et al. [2021] further generalize this result to random multilayer ReLU networks. However, all of the works above focus on random neural networks and fail to explain why adversarial examples exist for neural networks trained using stochastic gradient descent. In other words, why neural networks that are guaranteed to generalize well also suffer the same fate.

In this paper, we make progress toward addressing that question. In particular, we show that adversarial examples can be found easily using a single gradient step for neural networks trained using first-order methods. Specifically, we focus on over-parametrized two layer ReLU networks which have been studied extensively in recent years due to their favorable computational aspects. Several works have shown that under mild over-parametrization (size of the network is polynomial in the size of the training sample), the weights of two-layer ReLU networks stay close to initialization throughout the training of the network using gradient descent based methods [Allen-Zhu et al., 2019, Ji and Telgarsky, 2019, Arora et al., 2019]. In such a "lazy regime" of training, the networks are locally linear which allows us to give computational guarantees for minimizing the training loss (aka, optimization) as well as bound the test loss (aka, generalization). In fact, the lazy regime training of two-layer ReLU networks is the only instance of computationally efficient learning of neural networks. The purpose of our work is to show the deficiency of the lazy regime even though it is the dominant framework in the theory of deep learning.

Our key contributions are as follows. We investigate the robustness of over-parametrized neural networks against inference time attack in the lazy training regime. Building on the ideas of Bubeck et al. [2021], we show that neural networks that have small generalization error still suffer from adversarial examples, which can be found using a single gradient step.

In particular, we show that an adversarial perturbation of size $O(\frac{\|\mathrm{x}\|}{\sqrt{d}})$ in the direction of the gradient suffices to flip the prediction of two-layer ReLU networks trained in the lazy regime – these are networks with all weights at a distance of $O(\frac{1}{\sqrt{m}})$ from their initialization, where m is the width of the network. To the best of our knowledge, this is the first work which shows that networks with small generalization error are still vulnerable to adversarial attacks.

Further, we validate our theory empirically. We confirm that a perturbation of size $O(1/\sqrt{d})$ suffices to generate a strong adversarial example for a trained two-layer ReLU networks.

The rest of the paper is organized as follows. In Section 2, we introduce the preliminaries, give the problem setup, and discuss the related work. We present our main result in Section 3, and give a a proof sketch in Section 4. In Section 5 we provide empirical support for our theory, and conclude with a discussion in Section 6.

2 Preliminaries

Notation Throughout the paper, we denote scalars, vectors and matrices, respectively, with lowercase italics, lowercase bold and uppercase bold Roman letters, e.g. u, u, and U. We use $[m]$ to denote the set $\{1, 2, \ldots, m\}$. We use $\|\cdot\|$ or $\|\cdot\|_2$ exchangeably for ℓ_2 norm. Given a matrix $\mathrm{U} = [\mathrm{u}_1, \ldots, \mathrm{u}_m] \in \mathbb{R}^{d\times m}$, we define $\|\mathrm{U}\|_{2,\infty} = \max_{s\in[m]} \|\mathrm{u}_s\|$. We use $\mathcal{B}_2(\mathrm{u}, R)$ to denote the ℓ_2 ball centered at u of radius R. For any $\mathrm{U} \in \mathbb{R}^{d\times m}$, we use $\mathcal{B}_{2,\infty}(\mathrm{U}, R) = \{\mathrm{U}' \in \mathbb{R}^{d\times m} \mid \|\mathrm{U}' - \mathrm{U}\|_{2,\infty} \leq R\}$ to denote the $\ell_{2,\infty}$ ball centered at U of radius R. For any function $f : \mathbb{R}^d \to \mathbb{R}$, ∇f denotes the gradient vector. We define the standard normal distribution as $\mathcal{N}(0, 1)$, and the standard multivariate normal distribution as $\mathcal{N}(0, \mathrm{I}_d)$. We use $\mathbb{S}^{d-1}$ to denote the unit sphere in d dimensions. We use the standard O-notation (O and Ω).

2.1 Problem Setup

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ denote the input space and the label space, respectively. In this paper, we focus on the binary classification setting where $\mathcal{Y} = \{-1, +1\}$. We assume that the data (x, y) is drawn

from an unknown joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. For a function $f_{\mathrm{w}} : \mathcal{X} \to \mathcal{Y}$ parameterized by w in some parameter space $\mathcal{W}$, the *generalization error* captures the probability that f_{w} makes a mistake on a sample drawn from $\mathcal{D}$:

$$\text{generalization error} := P_{(\mathrm{x},y)\sim\mathcal{D}}(y f_{\mathrm{w}}(\mathrm{x}) \leq 0).$$

For a fixed $\mathrm{x} \in \mathcal{X}$, we consider norm-bounded adversarial perturbations given by $\mathcal{B}_2(\mathrm{x}, R)$, for some fixed perturbation budget R. The robust loss on x is defined as

$$\ell_R(\mathrm{w}; \mathrm{x}, y) = \mathbb{1}\left(\max_{\mathrm{x}' \in \mathcal{B}_2(\mathrm{x},R)} y f_{\mathrm{w}}(\mathrm{x}') \leq 0)\right).$$

The *robust error* then captures the probability that there exists an adversarial perturbation on samples drawn from $\mathcal{D}$ such that f_{w} makes a mistake on the perturbed point

$$L_R(\mathrm{w}; \mathrm{x}, y) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}[\ell_R(\mathrm{w}, \mathrm{x}, y)].$$

In this work, we focus on two-layer neural networks with ReLU activation, parameterized by a pair of weight matrices (a, W), computing the following function:

$$f(\mathrm{x}; \mathrm{a}, \mathrm{W}) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(\mathrm{w}_s^\top \mathrm{x}).$$

Here, m corresponds to the number of hidden nodes, i.e., the width of the network, $\sigma(\mathrm{x}) = \max\{0, \mathrm{x}\}$ is the ReLU activation function, and $\mathrm{W} = [\mathrm{w}_1, \ldots, \mathrm{w}_m] \in \mathbb{R}^{d\times m}$ and $\mathrm{a} = [a_1, \ldots, a_m] \in \mathbb{R}^m$ denote the top and the bottom layer weight matrices, respectively.

In this work, we study the robustness of neural networks which are close to their initialization. In particular, we are interested in the *lazy regime*, defined as follows:

Definition 2.1 (Lazy Regime)**.** We initialize the top and bottom layer weights as $a_s \sim \mathrm{unif}(\{-1, +1\})$ and $\mathrm{w}_{s,0} \sim \mathcal{N}(0, \mathrm{I}_d), \forall s \in [m]$. Let $\mathrm{W}_0 = [\mathrm{w}_{s,1}, \ldots, \mathrm{w}_{s,m}] \in \mathbb{R}^{d\times m}$. *Lazy regime* is the set of all networks parametrized by (a, W), such that $\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, C_0/\sqrt{m})$, for some constant C_0.

2.2 Related Work

Adversarial Examples. The field of adversarially robust machine learning has received significant attention starting with the seminal work of Szegedy et al. [2013]. Most existing works focus on proposing methods to generate adversarial examples, for example, the fast gradient sign method (FGSM) [Goodfellow et al., 2014], the projected gradient descent (PGD) method [Madry et al., 2017], the Carlini & Wagner attack [Carlini and Wagner, 2017b], to name a few. FGSM, in particular, involves a single step of gradient descent to generate an adversarial example, which is what we consider here.

Most related to our work are those on proving the existence of adversarial examples on random neural networks. This line of work initiated with the work of Shamir et al. [2019], where authors propose an algorithm to generate bounded ℓ_0 norm adversarial perturbation with guarantees for arbitrary deep networks. Shortly afterwards, Daniely and Schacham [2020] showed that multi-step gradient descent can find adversarial examples for random ReLU networks with small width (i.e. $m = o(d)$). More recently, work of Bubeck et al. [2021] shows that a single gradient descent update finds adversarial examples for sufficiently wide randomly initialized ReLU networks. This was later extended to multi-layer networks with the work of Bartlett et al. [2021]. A very recent work of Vardi et al. [2022] shows that gradient flow induces a bias towards non-robust networks. Our focus and techniques here are different – we show that independent of the choice of the training algorithm, all neural networks trained in the lazy regime remain vulnerable to adversarial attacks.

Over-parametrized Neural Network. Our investigation into the existence of adversarial examples in the lazy regime is motivated by recent advances in deep learning theory. In particular, a series of recent works establish generalization error bounds for first order optimization methods under the assumption that the weights of the network do not move much from their initialization [Jacot et al., 2018, Allen-Zhu et al., 2019, Arora et al., 2019, Gao et al., 2019, Chen et al., 2019, Ji and Telgarsky, 2019, Cao and Gu, 2020]. In particular, Chizat et al. [2018] recognize that "lazy training"

phenomenon is due to a choice of scaling that makes the model behave as its linearization around the initialization. Interestingly, linearity has been hypothesized as key reason for existence of adversarial examples in neural networks Goodfellow et al. [2014]. Furthermore, this was used to prove existence of adversarial attacks for random neural networks Bubeck et al. [2021], Bartlett et al. [2021] and also provide the basis of our analysis in this work.

3 Main Results

In this section, we present our main result. We show that under the lazy regime assumption, a single gradient step on f suffices to find an adversarial example to flip the prediction, given the network width is sufficiently wide but not excessively wide.

Theorem 3.1 (Main Result)**.** Let $\mathrm{w}_{s,0} \sim N(0, \mathrm{I}_d)$, $\mathrm{a} \sim \mathrm{unif}\{1, +1\}$ and $\gamma \in (0, 1)$. For any given x, with probability at least $1 - \gamma$, the following holds simultaneously for all $\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, C_0/\sqrt{m})$

$$\mathrm{sign}(f(\mathrm{x}; \mathrm{a}, \mathrm{W})) \neq \mathrm{sign}(f(\mathrm{x} + \delta; \mathrm{a}, \mathrm{W}))$$

where $\delta = \eta \nabla_{\mathrm{x}} f(\mathrm{x}; \mathrm{a}, \mathrm{W})$, provided that the following conditions are satisfied:

1. **Step Size:** $|\eta| = \dfrac{C_2}{\|\nabla f(x; \mathrm{a}, \mathrm{W})\|^2}$,
2. **Width requirement:** $\left\{d^{2.4}, C_3 \log\left(\frac{1}{\gamma}\right)\right\} \leq m \leq \min\left\{C_4 \exp(d^{0.24}), C_5 \gamma \sqrt{\exp\left(\frac{d}{\log(\frac{1}{\gamma})^2}\right)}\right\}$,

where $C_0, C_1, C_2, C_3, C_4, C_5$ are constants independent of width m and dimension d.

Several remarks are in order.

We note that the setting in Theorem 3.1 is motivated by the popular gradient descent based attack model which has been shown to be fairly effective in practice. While in practice, we perform multiple steps of projected gradient descent to generate an adversarial perturbation, here we show that a single step of gradient suffices with a small step size of $O(1/d)$. This attests to the claim that over-parameterized neural networks trained in the lazy regime remain quite vulnerable to adversarial examples. We confirm our findings empirically in Section 5.

Our analysis suggests that an attack size of $O(1/\sqrt{d})$ suffices, i.e., for any x a perturbation δ of size $\|\delta\| = O(1/\sqrt{d})$ can flip the sign of $f(\mathrm{x})$. This is a relatively small budget for adversarial perturbations, especially for for high-dimensional data. Indeed, in practice, we find that a noise budget of this size allows for adversarial attacks on neural networks trained on real datasets such as MNIST and CIFAR-10 Madry et al. [2017]. Interestingly, we find in our experiments, that $O(1/\sqrt{d})$ is the "right" perturbation budget in the sense that there is a sharp drop in robust accuracy. More details are provided in Section 5.

Our investigation focuses on the settings wherein the weights of the trained neural network stay close to the initialization. Further, given the initialization scheme we consider, the $O(1/\sqrt{m})$ deviation of the incoming weight vector at any neuron of the trained network from its initialization precisely captures the neural tangent kernel (NTK) regime studied in several prior works Ji and Telgarsky [2019], Du et al. [2018]. This $O(1/\sqrt{m})$ deviation bound is also the largest radius that we can allow in our analysis; see Section 4 for further details. In fact, our results would also hold for adversarial training given the same initialization and the bound on the deviation of weights is met. Curiously, our empirical results exhibit a phase transition around the perturbation budget of $O(1/\sqrt{m})$ – we see in Section 5 that robust accuracy increases as we allow the network weights to deviate more.

There is an interesting interplay between the network width and the input dimensionality. Our results require that the network is sufficiently wide (i.e., polynomially large in d), but not excessively wide (is sub-exponential in d). The upper bound is large enough to allow for the over-parameterization required in the NTK setting, as long as the input dimensionality is sufficiently large.

The following corollary about the robust error of networks in the lazy regime is immediate from the main theorem.

Corollary 3.2. Let $\mathrm{w}_{s,0} \sim N(0, \mathrm{I}_d)$, $\mathrm{a} \sim \mathrm{unif}\{1, +1\}$, $R = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ and $\gamma \in (0, 1)$. In the parameter settings of Theorem 3.1, for any data distribution $\mathcal{D}$, supported on the sphere $\mathbb{S}^{d-1}$, with probability at least $1 - 10\gamma$ (over the draw of W_0 and a), for all W s.t. $\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, C_0/\sqrt{m})$, the robust error $L_R(\mathrm{W}) \geq 0.9$.

To give some intuition about the result, we first paraphrase the reasoning from Bubeck et al. [2021], Bartlett et al. [2021] for random neural networks with weights W_0. We consider a linear model $f(\mathrm{w}; \mathrm{x}) = \frac{1}{\sqrt{d}} \sum_{s=1}^{d} \mathrm{w}_{s,0}\mathrm{x}_s$. Since the initial weights are independent ($\mathcal{N}(0, 1)$), from standard concentration arguments, for large d, w.h.p $|f(\mathrm{w}_0; \mathrm{x})| \leq O(1)$. In contrast, $\nabla_{\mathrm{x}} f(\mathrm{w}_0; \mathrm{x}) = \mathrm{w}$, so $\|\nabla_{\mathrm{x}} f(\mathrm{w}_0; \mathrm{x})\| = \|\mathrm{w}\|$, which is roughly $\Theta(\sqrt{d})$. Hence, for large enough d, a gradient step suffices to change the sign of f.

The above reasoning can be simply extended to weights that are "close" to random weights. In particular, for arbitrary weights W close to the initialization, consider the linear model: $f(\mathrm{w}; \mathrm{x}) = \frac{1}{\sqrt{d}} \sum_{s=1}^{d} \mathrm{w}_s \mathrm{x}_s$. Since $\|\mathrm{w}_s - \mathrm{w}_{s,0}\| = O(1)$, thus even in the worst case function still behaves as a constant translation of the initial prediction model $f(\mathrm{w}_0; \mathrm{x})$, and thus is constant. Similarly, $\|\nabla f(\mathrm{w}, \mathrm{x})\| = \Omega(\sqrt{d})$.

The work of Bubeck et al. [2021] extends the above observation to ReLU activation exploiting the fact that the function, even though non-linear, is still locally linear. Note that, in a different context, this local linearity property is central to the Neural Tangent Kernel regime. In particular, the key underlying idea there is based on the linear approximation to the prediction function. Furthermore, with appropriate initialization, a linear approximation is not only good enough model close to the initialization, but also enables tractable analysis. We collapse this "good" linearization property under the lazy training assumption since that suffices for our purposes.

Beyond Lazy regime. Our result can be easily extended to a broader class of neural networks. Define a cone $C = \{\mathrm{W} : \exists r > 0, r\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, \frac{C_0}{\sqrt{m}})\}$. Note that for binary classification, scaling the network weight with a positive factor would not change the prediction. As a result, for neural network that goes beyond lazy regime, Theorem 3.1 still holds if the network belongs to the cone C.

4 Proof Sketch

In this section, we provide an outline of the proof of Theorem 3.1. Recall that our goal is to show that for any given $\mathrm{x} \in \mathbb{R}^d$, we can find an adversarial perturbation $\delta \in \mathbb{R}^d$ such that $\|\delta\| \ll \|\mathrm{x}\|$, and $\mathrm{sign}(f(\mathrm{x}; \mathrm{a}, \mathrm{W})) \neq \mathrm{sign}(f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W}))$. Assume without loss of generality that $f(\mathrm{x}; \mathrm{a}, \mathrm{W}) > 0$, then $\mathrm{sign}(\eta) = -1$. From the fundamental theorem of calculus, we have $f(\mathrm{x}+\delta) = f(\mathrm{x}) + \int_0^1 f'(\mathrm{x}+t\delta)dt$. Using simple algebraic manipulations (see full proof in Appendix B), we have the following,

$$
\begin{aligned}
f(\mathrm{x}+\eta\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}); \mathrm{a}, \mathrm{W}) \quad \leq \quad & \overbrace{f(\mathrm{x}; \mathrm{a}, \mathrm{W})}^{O(1)} - |\eta| \overbrace{\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|}^{\Omega(\sqrt{d})} \\
& + |\eta| \overbrace{\sup_{\substack{\delta \in \mathbb{R}^d: \\ \|\delta\| \leq \eta \|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|}} \|\nabla f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|}^{o(\sqrt{d})}
\end{aligned}
$$

The goal is now to control each of the three terms as shown in the display above. Note that, at a high-level, the above bounds suffices to flip the label, thereby establishing the main result.

Given any x and $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$, we need to control:

1. The function value of neural network $f(\mathrm{x}; \mathrm{a}, \mathrm{W})$
2. The gradient of neural network $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$
3. The difference of gradients $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W})\|$ for $\|\delta\| \leq O(1/\sqrt{d})$

4.1 Bounding the function value

Lemma 4.1. (Informal) For any x, with probability at least $1-\gamma$, $|f(\mathrm{x};\mathrm{a},\mathrm{W})| \le 2\sqrt{\log(2/\gamma)} + C_0$ holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$, provided that $m \ge \Omega(\log(2/\gamma))$.

Proof Sketch: From triangle inequality,

$$|f(\mathrm{x};\mathrm{a},\mathrm{W})| \le |f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| + |f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)|$$

From Bubeck et al. [2021], we have $|f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| = O(1)$. The other term $|f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| \le O(1)$ simply follows from Lipschitzness of ReLU. Note that $O(1/\sqrt{m})$ is the largest deviation on $\|\mathrm{w}_s - \mathrm{w}_{s,0}\|$ that we can allow to get the constant bound here.

We next move to the second and third terms. We note that while bounding the first term simply follows from the Lipschitzness of the network, as we show in the following sections, the second and third terms require a finer analysis, due to the non-smoothness of ReLU.

4.2 Bounding the gradient

Lemma 4.2. (Informal) For any x, with probability at least $1-\gamma$, $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \ge \frac{1}{4}\sqrt{d}$ holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, \frac{C_0}{\sqrt{m}})$, provided that $m \ge \Omega(\log(1/\gamma))$ and $d \ge \Omega(\frac{\log(m/\gamma)}{m})$.

Proof Sketch: Let $\mathrm{P} := \mathrm{I}_d - \mathrm{x}\mathrm{x}^\top$ be the projection matrix onto orthogonal complement of x. From triangle inequality,

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \ge \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \ge \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| - \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|.$$

The first term $\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \ge \Omega(\sqrt{d})$, which follows from Bubeck et al. [2021] with minor changes arising from different scaling of the initialization.

We further decompose the second term,

$$\begin{aligned}\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \;\le\; & \left\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathrm{P}(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \mathrm{w}_{s,0}\sigma'(\mathrm{w}_s^\top\mathrm{x}))\right\| \\ & + \left\|\sup_{\mathrm{W}\in\mathcal{B}_{2,\infty}(\mathrm{W}_0,\frac{C_0}{\sqrt{m}})} \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathrm{P}\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\right\|\end{aligned}$$

The first term is $O(1)$ simply from the fact that $\|\mathrm{w}_s - \mathrm{w}_0\| \le \frac{C_0}{\sqrt{m}}$. To bound the second term, the key arguments are as follows: by definition of P, the term $\mathrm{P}\mathrm{w}_{s,0}$ is independent of $\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})$, therefore, as in Bubeck et al. [2021], we can bound both these terms independently. Note that $a_s\mathrm{P}\mathrm{w}_{s,0} \sim \mathcal{N}(0, \mathrm{I}_{d-1})$. Therefore, the second term is equal in distribution to the following random variable:

$$\sup_{\mathrm{W}\in\mathcal{B}_2(\mathrm{W}_0,\frac{C_0}{\sqrt{m}})} \frac{\|\mathrm{z}\|}{\sqrt{m}}\sqrt{\textstyle\sum_{s=1}^{m}(\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))^2}$$

where $\mathrm{z} \sim \mathcal{N}(0, \mathrm{I}_{d-1})$.

The term $|\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})| = 1$ only when the s-th neuron has different signs for incoming weights $\mathrm{w}_{s,0}$ and w_s. So, at a high-level, we want to bound the number of neurons changing sign between weights w_0 and w. To do this, we draw insights from results in the lazy training regime that bound the number of such neurons in terms of deviations in weights. In particular, the following lemma shows that with high probability, for all $\mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, \frac{C_0}{\sqrt{m}})$, at most $O(\sqrt{m})$ neurons can have a different sign from their initialization.

Lemma 4.3. Let $S_v := \left\{s \middle| \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right), \mathbb{1}[\langle \mathrm{w}_s, \mathrm{x}\rangle > 0] \ne \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle > 0]\right\}$. Then with probability at least $1-\gamma$, we have $|S_v| \le C_0\sqrt{m} + \sqrt{\frac{m\log(1/\gamma)}{2}}$.

The key idea is that we can absorb the union bound in the definition of the set above, without compromising any increase in the size of the set. Putting all these pieces together yields the claimed bound in the lemma.

4.3 Bounding the difference of gradients

Lemma 4.4. For any x, with probability at least $1-\gamma$, for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$, $\sup_{r\in\mathbb{S}^{d-1},\delta\in\mathbb{R}^d} \|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\| \leq c\sqrt{d}$, provided that $\|\delta\| \leq O(1/\sqrt{d})$, and $d^{2.4} \leq m \leq \min\left\{O(\exp(d^{0.24})), O(\gamma\sqrt{\exp\left(\frac{d}{\log(1/\gamma)^2}\right)})\right\}$

Proof Sketch: It follows from the definition of norm that

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\|$$

The strategy is to bound the right-hand side for fixed r and δ, and then use an ε-net argument to bound the supremum. To that end, define the following quantities:

$$X_{s,0} = a_s \mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))), X_s = a_s \mathrm{w}_s^\top r(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))).$$

Then, the norm of difference of gradients is upper bounded by

$$\tfrac{1}{\sqrt{m}}\sum_{s=1}^m X_{s,0} + \tfrac{1}{\sqrt{m}}\sum_{s=1}^m (X_s - X_{s,0}).$$

The first term $\frac{1}{\sqrt{m}}\sum_{s=1}^m X_{s,0} = O((R\sqrt{\log(d)})^{1/4})$ which follows from Bubeck et al. [2021] with minor changes due to the different scale of the initialization. However, the second term requires a more careful analysis. In particular, the difference $X_s - X_{s,0}$ includes deviations of both variables x and W. Therefore, we need to expand the difference in terms of each variable and carefully bound the resulting terms. Similar to the ideas presented in the previous section, here we need to bound the size of the set of neurons that can change sign under perturbations on both x and W. It turns out that the size of this set is in the same order as the one in Lemma 4.3, as long as the perturbation in x is $O(1)$.

5 Experiments

The primary goal of this section is to provide empirical support for our theoretical findings. In particular, we want to verify how close the practice is to our theoretical bounds in terms of the size of the perturbation needed and the interplay of robustness with training in the lazy regime. For instance, are trained models indeed vulnerable to small perturbations of size $O(1/\sqrt{d})$ as suggested in Theorem 3.1? Is the assumption of training in the lazy regime merely for analytical convenience or do we see a markedly different behavior if we allow the iterates to deviate more than $O(1/\sqrt{m})$ from initialization?

Datasets & Model specification. We utilize the MNIST dataset for our empirical study. MNIST is a dataset of 28×28 greyscale handwritten digits LeCun et al. [1998]. We extract examples corresponding to images of the digits '0' and '1', resulting in 12665 training examples and 2115 test examples. To study the behaviour of different quantities as a function of the input feature dimensionality, we downsample the images by different factors to simulate data with d ranging between 25 and 784 (the original dimension). We use two-layer ReLU networks with a fixed-top layer to match our theory and use the initialization described in Section 2.

5.1 Lazy regime for standard SGD

Setting. For each $d \in \{5^2, 7^2, 10^2, 14^2, 20^2, 25^2, 28^2\}$, we set the width of the network $m \in \{1e3, 1e4, 1e5\}$. We train the network using standard SGD on the logistic loss using a learning rate 0.1. We denote the weight matrix at the tth iterate of SGD as W_t and the incoming weight vector into the sth hidden node at iteration t as $\mathrm{w}_{s,t}$. The training terminates if the iterates exit the lazy training regime, i.e., if $\sup_{1\leq s\leq m}\|\mathrm{w}_{s,t} - \mathrm{w}_{s,0}\| > \frac{C_0}{\sqrt{m}}$. We experiment with $C_0 \in \{10, 20, 30, 40\}$. We note that regardless of the above stopping criterion, each of the trained models achieves a test error of less than 1%. In other words, the models our training algorithm returns are guaranteed to satisfy the lazy regime assumption and generalize well.

Given the trained neural network with parameters (W, a), for each test example x, we find the smallest step size η such that the perturbation $\delta = \eta\nabla_{\mathrm{W}} f(\mathrm{x})$ flips the sign of the prediction, i.e., $f_{\mathrm{W}}(\mathrm{x}) f_{\mathrm{W}}(\mathrm{x}+\delta) < 0$. For each experimental setting (i.e, for each value of d, C_0, and m), we report results averaged over 5 independent random runs.

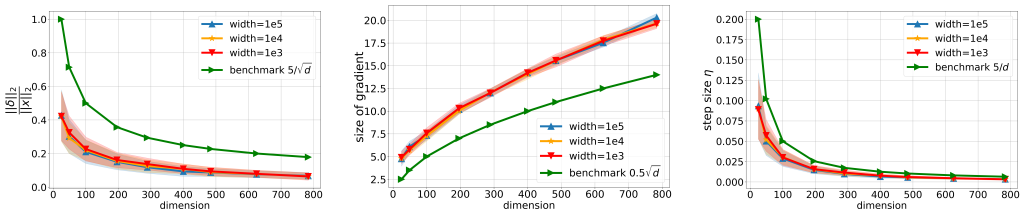

Figure 1: Minimal size of the perturbation vector needed to flip the predicted label (**left**), the norm of the gradient $\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})$ for W trained in the lazy regime (**middle**), and the corresponding step size (**right**), as a function of the input dimension d for a fixed value of $C_0 = 10$ and different values of network widths (m).

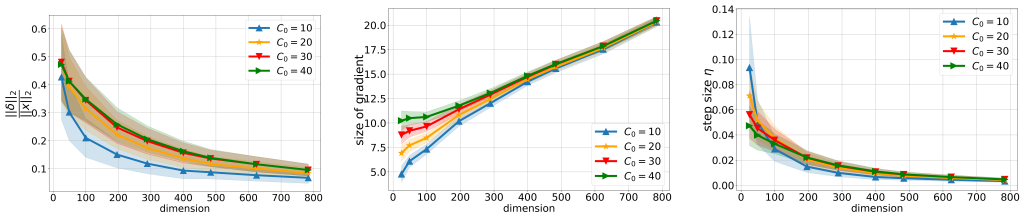

Figure 2: Minimal size of the perturbation vector needed to flip the predicted label (**left**), the norm of the gradient $\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})$ for W trained in the lazy regime (**middle**), and the corresponding step size (**right**), as a function of the input dimension d for a fixed value of network width of $m = 10^5$ and different values of C_0.

Results. The key quantities in our analysis are the step size η, the norm of the gradient $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$, and the size of the perturbation vector which given as the product of the step size and the norm of the gradient. In Figure 1, we plot these quantities as a function of input dimension d, for a fixed C_0, for various values of width $m \in \{1e3, 1e4, 1e5\}$ (red, orange, and blue curves, respectively). The green curves represent the corresponding theoretical bound. As predicted by Theorem 3.1, the gradient norm behaves roughly as $\Omega(\sqrt{d})$. Furthermore, the minimum step-size required to flip the label, and the minimum size of the perturbation vector (normalized by the input norm) behave as $O(1/d)$ and $O(1/\sqrt{d})$, respectively.

In Figure 2, we explore the behaviour of the key quantities as we allow for larger deviations in the lazy regime, by allowing for larger values of C_0, for a fixed network width of $m = 10^5$. We see that as C_0 increases, both the size of the gradient and the minimum step size required to flip the label increase, thereby increasing the size of the perturbation vector needed to change the sign of the prediction. This suggests that allowing the trained networks to deviate more from the initialization potentially leads to models that are more robust (against the attack model based on a single step of gradient descent).

Remarkably, the minimum perturbation size, closely follows the $O(1/\sqrt{d})$ behaviour predicted by the theorem. This begets further investigation into the required perturbation size, which is the focus of the next subsection.

5.2 Lazy regime for adversarial training

We begin by noting that our theoretical results are not limited to any particular learning algorithm. In fact, the results hold simultaneously for all networks in the lazy regime. In this section, we further verify our theoretical results for networks that are adversarially trained in the lazy regime.

Setting. We downsample images as in the previous section to simulate data with $d \in \{5^2, 6^2, 7^2, \ldots, 28^2\}$. For this set of experiments, we also normalize the data, so that $\|\mathrm{x}\| = 1$, to closely match the setting of our theorem. We use projected gradient descent (PGD) Madry et al. [2017] for adversarial training. To ensure the updates remain in the lazy regime, after each gradient step, we project the weights onto $\mathcal{B}_{2,\infty}(\mathrm{W}_0, V)$; see Algo. 1 in the appendix for details. We experiment with values of V and the size of $\|\delta\|$ in the intervals $0.05 \leq V \leq 1$ and $0.05 \leq \|\delta\| \leq 1$, respectively. These ranges are wide enough to include the $O(1/\sqrt{d})$ and $O(1/\sqrt{m})$ theoretical

bounds that we aim to verify here. We note that the projection step is not commonly used in adversarial training; however, in our experiments, it does not hurt the test accuracy of the trained models, perhaps due to the simplicity of the classification task. At test time, we generate the adversarial examples using the PGD attack instead of a single gradient attack as in Section 5.1. We report the robust test accuracy averaged over 5 independent random runs of the experiment.

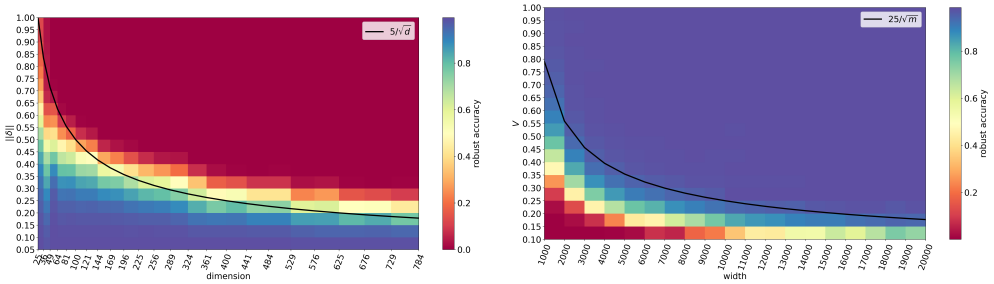

Figure 3: **(Left):**Robust accuracy as a function of perturbation size $\|\delta\|$ and dimension d, for a fixed network width of $m = 1000$ and a fixed value of maximal deviation of weight vectors of $V = \frac{C_0}{\sqrt{m}}$ with $C_0 = 10$. **(Right):**Robust accuracy as a function of width m and maximal deviation of weight vectors, V, for a fixed dimension of $d = 784$ and fixed perturbation size $\|\delta\| = 0.2$.

Results. In Figure 3 left one, we fix the network width to $m = 1000$ and set $V = \frac{C_0}{\sqrt{m}} = \frac{10}{\sqrt{1000}} = 0.316$, to check how the dimension d and the perturbation budget affect the robust accuracy. We observe a sharp drop in robust accuracy about the $5/\sqrt{d}$ threshold for the perturbation budget $\|\delta\|$, as predicted by Theorem 3.1. We can see that for perturbation sizes slightly larger than $5/\sqrt{d}$, the robust accuracy drops to 0, while the trained networks are robust against smaller perturbations.

In Figure 3 right one, we fix the input dimensionality to $d = 28^2$, and the perturbation size to $\|\delta\| = 0.2$, to check how the network width m and the maximal deviation in weights, given by V, affects the robust accuracy. In Figure 3 right one, we observe a phase transition in the robust accuracy at value of V around $O(1/\sqrt{m})$, as required by Theorem 3.1. In particular, for any fixed width m, even a slight deviation from the $25/\sqrt{m}$ radius allows for training networks that are completely robust against the PGD attacks. With a smaller radius, however, the robust accuracy quickly drops to 0. This experiment suggests that adversarial robustness may require going beyond the lazy regime.

Why MNIST? We remark that our empirical demonstration on a simple dataset makes a stronger case for our theoretical result. This is because the perturbation size of $O(1/\sqrt{d})$ prescribed by our theoretical result is **worst-case**, regardless of the dataset. So, in an easy setting like binary-MNIST, we would expect our bound to be pessimistic and a smaller perturbation to be sufficient. On the contrary, we see that even MNIST exhibits a perturbation-dimension relationship that closely follows our theory. This is empirical evidence of the tightness of our bounds.

6 Conclusion

In this paper, we study the robustness of two-layer neural networks trained in the lazy regime against inference time attack. In particular, we show that the class of networks for which the weights are close to the initialization after training, are still vulnerable to adversarial attacks based on a single step of gradient descent with a perturbation of size $O(\frac{\|\mathrm{x}\|}{\sqrt{d}})$.

Our works suggests several research directions for future work. First,the focus here is on two-layer networks, so a natural question is to ask whether the same holds for multi-layer networks. Second, while we show that a single step of gradient descent based attack is sufficient for our purposes, we expect to get stronger results if we consider stronger attacks, for instance, with gradient descent based attack that is run to convergence. Third, our experiments highlight an intriguing relationship between the width, the input dimension, and the robust accuracy. So, a natural avenue is to understand these dependencies more carefully. These include, figuring out the dependence on the input dimensionality of the minimal perturbation required to generate an adversarial example. Similarly, how does the robust error behave as a function of the maximal distance of the weight vectors from the initialization.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim vSrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Rémi Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. *arXiv preprint arXiv:2104.03863*, 2021.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Amit Daniely and Hadas Schacham. Most relu networks suffer from ℓ^2 adversarial perturbations. *arXiv preprint arXiv:2010.14927*, 2020.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32:13029–13040, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019.

Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Prateek Mittal, and Mung Chiang. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. *arXiv preprint arXiv:1801.02780*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *arXiv preprint arXiv:2202.04347*, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Checklist

A. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

B. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix B.

C. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1, Figure 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

D. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

E. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional experiments

Figure 4 and 5 plot the same phenomena as Figure 1 and 2. The only difference is that they strictly follows the setting described in Section 2 where the training data and test data has been normalized to $\|\mathrm{x}\| = 1$, and there's no bias term when initializing the neural network linear layer to consist of our problem setup. For each experimental setting (i.e, for each value of d, C_0, and m), we report results averaged over 5 independent random runs.

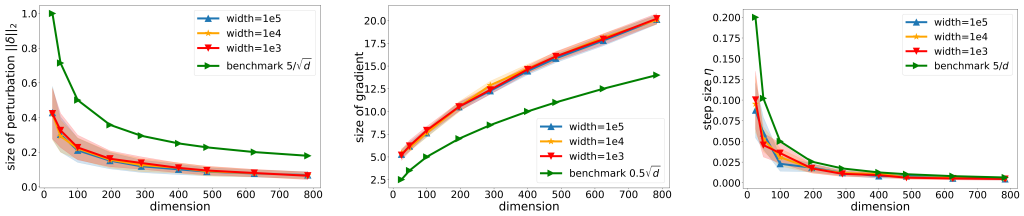

Figure 4: Normalize data $\|\mathrm{x}\| = 1$. Smallest size of perturbation to switch the prediction (**left**), norm of the gradient after training the neural networks (**middle**), smallest step size (**right**), as a function of input dimension d for fix $C_0 = 10$ with different width m.

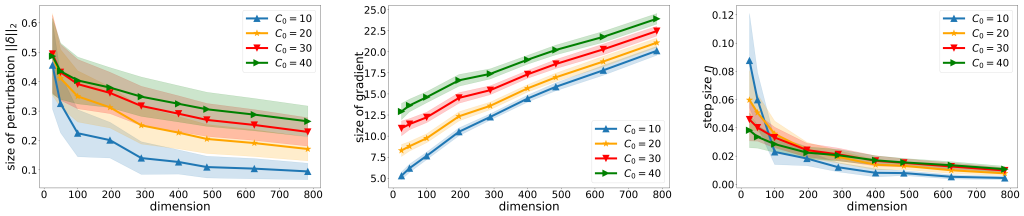

Figure 5: Normalize data $\|\mathrm{x}\| = 1$. Smallest size of perturbation to switch the prediction (**left**), norm of the gradient after training the neural networks (**middle**), smallest step size (**right**), as a function of input dimension d for fix $m = 10^5$ with different C_0.

Figure 6 is the histogram of smallest size of perturbation to switch the prediction, the norm of the gradient after training the neural networks, and the step size when $d = 784, m = 10^5, C_0 = 10$. The histogram exhibits a Gaussian distribution, and the step size to flip the prediction is small. (η concentrates around 0.045.)

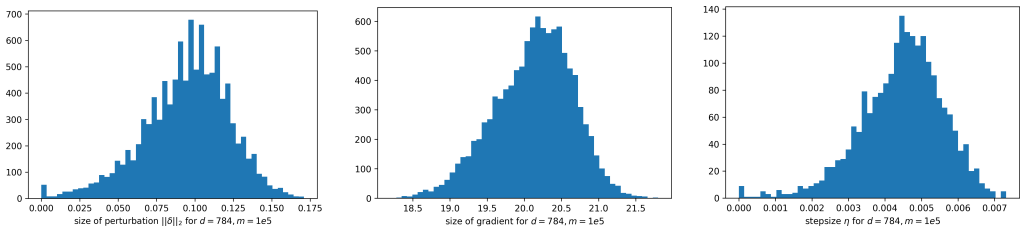

Figure 6: Histogram of smallest size of perturbation to switch the prediction, the norm of the gradient after training the neural networks, and the step size when $d = 784, m = 10^5, C_0 = 10$.

In Algorithm 1, we describe our projected adversarial training algorithm in details. For generating adversarial example with budget R at each round, we choose learning rate $\alpha = 2.5 \times R/100$ with $T_2 = 100$. This is a common choice, which is first introduced in [Madry et al., 2017]. For updating the weight matrix, we choose learning rate $\beta = 0.01$. We stop when the robust training accuracy is not increasing.

Algorithm 1 Projected Adversarial Training

Training samples $(\mathrm{X}, \mathrm{y}) = \{\mathrm{x}_i, y_i\}_{i=1}^n$. Initialize $\mathrm{w}_{s,0} \sim N(0, \mathrm{I}_d), a_s \sim \mathrm{unif}\{\pm 1\}, \forall s \in [m]$ with fixed a. Epochs T_1, T_2. Learning rate α, β, perturbation budget per sample R. Logistic loss ℓ. Batch size bs. $\mathcal{P}_\Delta$ is the projection operator onto the set Δ.
for $t = 1, \ldots, T_1$ **do**
  for $k = 1, \ldots, T_2$ **do**
    $\tilde{\mathrm{X}} = \mathrm{X} + \alpha \sum_{i=1}^n \nabla_{\mathrm{x}_i} \ell(y_i f(\mathrm{x}_i; \mathrm{a}, \mathrm{W}_{t-1}))$
    $\tilde{\mathrm{X}} = \mathcal{P}_{\mathcal{B}_{2,\infty}(\mathrm{X}, R)}(\tilde{\mathrm{X}})$ {generate adversarial examples}
  end for
  for $\lfloor \frac{n}{bs} \rfloor$ rounds **do**
    Sample a mini-batch of size bs from $(\tilde{\mathrm{X}}, \mathrm{y})$ as $(\tilde{\mathrm{x}}_{i_j}, y_{i_j})_{j=1}^{bs}$.
    $\mathrm{W}_t = \mathrm{W}_{t-1} - \beta \nabla_{\mathrm{W}_{t-1}} \sum_{j=1}^{bs} \ell(y_{i_j} f(\tilde{\mathrm{x}}_{i_j}; \mathrm{a}, \mathrm{W}_{t-1}))$
  end for
  $\mathrm{W}_t = \mathcal{P}_{\mathcal{B}_{2,\infty}(\mathrm{W}_0, \frac{C_0}{\sqrt{m}})}(\mathrm{W}_t)$ {Project the weight matrix to satisfy lazy regime.}
end for
return: W_{T_1}

B Proof of theorems

Proof of Theorem 3.1. From Lemma B.5, B.7, B.9, we have that

$$\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\| \geq C_1' \sqrt{d},$$
$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \leq C_2',$$
$$\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x} + \delta; \mathrm{a}, \mathrm{W})\| \leq C_3' \sqrt{d} \text{ for } \|\delta\| \leq o(\frac{1}{\sqrt{d}})$$

We simplify the notation as $f(\mathrm{x}) = f(\mathrm{x}; \mathrm{a}, \mathrm{W})$. Define $\tilde{\eta} = \eta \|\nabla f(\mathrm{x})\|^2$. Without loss of generality, assume $f(\mathrm{x}) > 0$, let

$$\tilde{\eta} = -\frac{2f(\mathrm{x})}{1 - \sup_{\|\delta\| \leq \frac{\tilde{\eta}}{\|\nabla f(\mathrm{x})\|}} \frac{\|\nabla f(\mathrm{x}+\delta)\| - \|f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|}},$$

we have for the point $\mathrm{x} + \tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}$,

$$\begin{aligned}
& f(\mathrm{x} + \tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) \\
&= f(\mathrm{x}) + \int_0^1 f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2})' dt && \text{(Fundamental theorem of calculus)} \\
&= f(\mathrm{x}) + \int_0^1 \tilde{\eta} \frac{\nabla f(\mathrm{x})^\top}{\|\nabla f(\mathrm{x})\|^2} \nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) dt \\
&= f(\mathrm{x}) + \tilde{\eta} + \int_0^1 \tilde{\eta} \frac{\nabla f(\mathrm{x})^\top}{\|\nabla f(\mathrm{x})\|} \frac{(\nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) - \nabla f(\mathrm{x}))}{\|\nabla f(\mathrm{x})\|} dt \\
&\leq f(\mathrm{x}) + \tilde{\eta} + |\tilde{\eta}| \int_0^1 \frac{\|\nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) - \nabla f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|} dt && \text{(Cauchy–Schwarz)} \\
&\leq f(\mathrm{x}) + \tilde{\eta} + |\tilde{\eta}| \sup_{\delta \in \mathbb{R}^d : \|\delta\| \leq \frac{\tilde{\eta}}{\|\nabla f(\mathrm{x})\|}} \frac{\|\nabla f(\mathrm{x} + \delta) - \nabla f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|} \\
&= -f(\mathrm{x}) < 0
\end{aligned}$$

Note $|\tilde{\eta}| = \Theta(1)$ and thus $\eta \leq O(\frac{1}{d})$, $\|\eta \nabla f(\mathrm{x})\| = \|\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(x)\|^2}\| \leq O(\frac{1}{\sqrt{d}})$.

□

In order to proof the main theorem, we start by giving some standard theorems, which will be used in later proof.

Theorem B.1 (Berstein's inequality)**.** Let $z_1, \ldots, z_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that

$$\sum_{i=1}^{n} \mathbb{E}[z_i^2] \leq v \text{ and } \sum_{i=1}^{n} \mathbb{E}[|z_i|^q] \leq \frac{q!}{2} v c^{q-2} \text{ for all integers } q \geq 3.$$

If $S = \sum_{i=1}^{n}(z_i - \mathbb{E}z_i)$, then for all $t > 0$,

$$\mathrm{P}(S \geq \sqrt{2vt} + ct) \leq e^{-t}.$$

We will also use repeatly that

$$\mathbb{E}_{z\sim\mathcal{N}(0,1)}[|z|^q] \leq (q-1)!! \leq \frac{q!}{2}, \tag{1}$$

as well as the following concentration of χ^2 random variables: let $z_1, \ldots, z_m$ be i.i.d. standard Gaussians, then with probability at least $1-\gamma$, one has

$$\left| \sum_{s=1}^{m} z_s^2 - m \right| \leq 4\sqrt{m \log(2/\gamma)} \tag{2}$$

Theorem B.2. (Chernoff's inequality) If $z_1, z_2, \ldots, z_N$ are independent Bernoulli random variables with parameters μ_i. Let $S_N = \sum_i z_i$ and $p = \mathbb{E}S_N = \sum_i \mu_i$, then for $t > p$,

$$P(S_N \geq t) \leq \exp(-p)\left(\frac{ep}{t}\right)^t$$

Theorem B.3. (Theorem 2.26 in Wainwright [2019]) Let $\mathrm{z} = (z_1, z_2, \ldots, z_n)$ be a vector of i.i.d standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t. the Euclidean norm. Then we have

$$P(|f(\mathrm{z}) - \mathbb{E}f(\mathrm{z})| \geq t) \leq 2e^{-\frac{t^2}{2L^2}} \text{ for all } t \geq 0$$

In Lemma B.4, we let S_v denote the set of neurons that change sign between weights w_0 and w for the x, S_v' as the set of neurons that change sign between weights w_0 and w for the $\mathrm{x}+\delta$. We show that the size of S_v and S_v' is small as long as the weights stay close to initialization.

Lemma B.4. Define the following

$$S_v := \left\{ s \middle| \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathbb{1}[\langle \mathrm{w}_s, \mathrm{x}\rangle > 0] \neq \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle > 0] \right\}$$
$$S_v' := \left\{ s \middle| \exists \delta, \|\delta\| \leq R \leq 0.5, \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathbb{1}[\langle \mathrm{w}_s, \mathrm{x}+\delta\rangle > 0] \neq \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle > 0] \right\}$$

Then with probability at least $1-\gamma$, the following hold:

$$|S_v| \leq \left| \left\{ s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V \right\} \right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V\right] \leq Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}$$
$$|S_v'| \leq \left| \left\{ s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \leq V \right\} \right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \leq V\right] \leq \left(Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}\right)(1+R)$$

Proof of Lemma B.4. Define $\mathrm{v}_s = \mathrm{w}_s - \mathrm{w}_{s,0}$. Note that $s \in S_v$ implies $\forall 1 \leq s \leq m$,

$$\left|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle\right| \leq \sup_{\|\mathrm{v}_s\| \leq V} \left|\langle \|\mathrm{w}_s - \mathrm{w}_{s,0}\|, \mathrm{x}\rangle\right| \leq \sup_{\|\mathrm{v}_s\| \leq V} \|\mathrm{v}_s\|_2 \|\mathrm{x}\|_2 \leq V$$

Thus we have

$$\mathbb{E}\left[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V\right]\right] = P\big(|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V\big) \leq \frac{2V}{\|\mathrm{x}\|_2^2\sqrt{2\pi}} \leq V,$$

where the expectation is with respect to the randomness in initialization, and the inequality holds since $\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle$ is a Gaussian r.v. with variance 1. By Hoeffding inequality, with probability at least $1-\gamma$,

$$\begin{aligned}
&\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\right] \\
&\le \mathbb{E}\left[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\right]\right] + \sqrt{\frac{\log(1/\gamma)}{2m}} && \text{(Hoeffding inequality)} \\
&\le V + \sqrt{\frac{\log(1/\gamma)}{2m}}
\end{aligned}$$

As a result, with probability at least $1-\gamma$, we arrive at the following upperbound on the size of S_v:

$$|S_v| \le \left|\left\{s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\right\}\right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\right] \le Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}$$

Same way we can bound the size of S_v'. $s \in S_v'$ implies $\forall 1 \le s \le m$,

$$\left|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle\right| \le \sup_{\|\mathrm{v}_s\|\le V, \|\delta\|\le R} \left|\langle \mathrm{v}_s, \mathrm{x}+\delta\rangle\right| \le \sup_{\|\mathrm{v}_s\|\le V, \|\delta\|\le R} \|\mathrm{v}_s\|_2 \|\mathrm{x}+\delta\|_2 \le V(1+R)$$

Thus we have

$$\mathbb{E}\left[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\right]\right] = P\big(|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\big) \le V(1+R),$$

where the expectation is with respect to the randomness in initialization, and the inequality holds since $\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle$ is a Gaussian r.v. with variance 1. By Hoeffding inequality, with probability at least $1-\gamma$,

$$\begin{aligned}
&\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\right] \\
&\le \mathbb{E}\left[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\right]\right] + \sqrt{\frac{\log(1/\gamma)}{2m}}(1+R) && \text{(Hoeffding inequality)} \\
&\le V(1+R) + \sqrt{\frac{\log(1/\gamma)}{2m}}(1+R)
\end{aligned}$$

As a result, with probability at least $1-\gamma$, we arrive at the following upperbound on the size of S_v:

$$|S_v| \le \left|\left\{s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V\right\}\right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V\right] \le \left(Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}\right)(1+R)$$

□

Now we begin our proof. Essentially we want to lower bound $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$, upper bound $|f(\mathrm{x}; \mathrm{a}, \mathrm{W})|$ and upper bound $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W})\|$. Lemma B.5 gives the upper bound of $|f(\mathrm{x}; \mathrm{a}, \mathrm{W})|$ as $O(1)$.

Lemma B.5. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \le \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}} + C_0$$

Particularly, there exists $C > 0$ such that for $m \ge C\log(2/\gamma)$, we have

$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \le 2\sqrt{\log(2/\gamma)} + C_0$$

Proof of Lemma B.5. From Bubeck et al. [2021] we know that $|f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| \leq \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}}$. Now we consider bound $|f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)|$,

$$
\begin{aligned}
|f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| &\leq |\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\sigma(\mathrm{w}_s^\top \mathrm{x}) - \sigma(\mathrm{w}_{s,0}^\top \mathrm{x}))| \\
&\leq \frac{1}{\sqrt{m}}\sum_{s=1}^{m} |\sigma(\mathrm{w}_s^\top \mathrm{x}) - \sigma(\mathrm{w}_{s,0}^\top \mathrm{x})| && \text{(Triangle Inequality)} \\
&\leq \sqrt{m}|\mathrm{w}_s^\top \mathrm{x} - \mathrm{w}_{s,0}^\top \mathrm{x}| && (\sigma(\cdot) \text{ is 1-Lipschitz.}) \\
&\leq \sqrt{m}\|\mathrm{w}_s - \mathrm{w}_{s,0}\|\|\mathrm{x}\| && (\|\mathrm{w}_s - \mathrm{w}_{s,0}\| \leq \tfrac{C_0}{\sqrt{m}}) \\
&\leq C_0
\end{aligned}
$$

Thus,

$$
|f(\mathrm{x};\mathrm{a},\mathrm{W})| = |f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| + |f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| \leq \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}} + C_0
$$

□

In Lemma B.6, we want to calculate the upper bound of the probability that the neuron flips sign due to (1) perturbation δ; (2) different perturbation δ, δ' on the data x at weight w_s.

Lemma B.6. For any δ such that $\|\delta\| \leq R \leq \frac{1}{2}$,

$$
P\Big(\mathrm{sign}(\mathrm{w}_{s,0}^\top \mathrm{x}) \neq \mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))\Big) \leq R\sqrt{2\log(d)} + \frac{1}{d} \tag{3}
$$

$$
\begin{aligned}
&P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) \neq \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\Big) \\
&\qquad \leq 2\varepsilon\Big(\sqrt{d} + 2\sqrt{d\log(2/\varepsilon)}\Big) + (1+R+\varepsilon)V
\end{aligned} \tag{4}
$$

Proof of Lemma B.6. Equation (3) directly follows from Bubeck et al. [2021]. For equation (4), we have

$$
\begin{aligned}
&P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) \neq \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\Big) \\
&\leq P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\delta'-\delta)| \geq t\Big) \\
&\quad + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\mathrm{x}+\delta)| \leq t\Big) && \text{(Holds for any threshold } t.) \\
&\leq P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \|\mathrm{w}_s\| \geq t/\varepsilon\Big) + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\mathrm{x}+\delta)| \leq t\Big) \\
&\leq P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \|\mathrm{w}_{s,0}\| \geq t/\varepsilon - \|\mathrm{w}_s - \mathrm{w}_{s,0}\|\Big) \\
&\quad + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_{s,0}^\top(x+\delta)| \leq t + |(\mathrm{w}_s - \mathrm{w}_{s,0})^\top(\mathrm{x}+\delta)|\Big) \\
&\qquad\qquad \text{(Triangle Inequality, pick } t = \varepsilon\Big(\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + V\Big)) \\
&\leq P\Big(\|\mathrm{w}_{s,0}\| \geq \sqrt{d + 4\sqrt{d\log(2/\varepsilon)}}\Big) + P\Big(|\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)| \leq \varepsilon\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + (1+R+\varepsilon)V\Big) \\
&\qquad\qquad (\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta) \sim \mathcal{N}(0,\sigma^2) \text{ with } \sigma^2 \geq \tfrac{1}{2} \text{ since } \|\delta\| \leq \tfrac{1}{2}) \\
&\leq \varepsilon + \varepsilon\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + (1+R+\varepsilon)V && \text{(Using equation (2).)} \\
&= 2\varepsilon(\sqrt{d} + 2\sqrt{d\log(2/\varepsilon)}) + (1+R+\varepsilon)V
\end{aligned}
$$

$\square$

Lemma B.7 gives lower bound on $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \Omega(\sqrt{d})$.

Lemma B.7. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \left(\frac{1}{2} - (\sqrt{\frac{2\log(4/\gamma)}{m}} + \frac{\log(4/\gamma)}{m})\right)^{1/2}\left(d - 5\sqrt{d\log(8/\gamma)}\right)^{1/2} - C_0 - \sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})}\left(d + 4\sqrt{d\log(8m/\gamma)}\right)^{1/2}$$

Particularly, there exists $C > 0$ such that for $m \geq C\log(4/\gamma)$ and $d \geq C\frac{\log(8m/\gamma)}{m}$, we have

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \frac{1}{4}\sqrt{d}$$

Proof of Lemma B.7. Follow the process of Bubeck et al. [2021], let $\mathrm{P} = \mathrm{I}_d - \mathrm{x}\mathrm{x}^\top$ be the projection on the orthogonal complement of the span of x. We have $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \|P\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\|$. Thus we have,

$$\begin{aligned}\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| &\geq \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \\ &= \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0) + \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \\ &\geq \|P\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| - \|P\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|\end{aligned} \tag{5}$$

Since $a_s\mathrm{P}\mathrm{w}_{s,0}$ is distributed as $\mathcal{N}(0, \mathrm{I}_{d-1})$, denote $\mathrm{z} := a_s\mathrm{P}\mathrm{w}_{s,0}$, we have

$$\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}) \overset{(d)}{=} \left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})^2}\right)\mathrm{z} \text{ where } \mathrm{z} \sim \mathcal{N}(0, \mathrm{I}_{d-1})$$

where $\overset{(d)}{=}$ means equal in distribution.

Using (2), we have with probability at least $1-\gamma$

$$\|\mathrm{z}\|^2 \geq d - 1 - 4\sqrt{d\log(2/\gamma)} \geq d - 5\sqrt{d\log(2/\gamma)} \qquad (d \geq 1, \gamma < 2/e)$$

Apply Bernstein's inequality with $v = m, c = 1$, we have with probability at least $1-\gamma$,

$$\begin{aligned}\frac{1}{m}\sum_{s=1}^{m}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})^2 &\geq \mathbb{E}_{X\sim\mathcal{N}(0,1)}[|\sigma'(X)|^2] - \left(\sqrt{\frac{2\log(1/\gamma)}{m}} + \frac{\log(1/\gamma)}{m}\right) \\ &= \frac{1}{2} - \left(\sqrt{\frac{2\log(1/\gamma)}{m}} + \frac{\log(1/\gamma)}{m}\right)\end{aligned}$$

Thus we can get

$$\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \geq \left(\frac{1}{2} - (\sqrt{\frac{2\log(2/\gamma)}{m}} + \frac{\log(2/\gamma)}{m})\right)^{1/2}\left(d - 5\sqrt{d\log(4/\gamma)}\right)^{1/2}$$

Now we start upper bound $\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|$. Note that

$$
\begin{aligned}
&\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \\
&= \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_s^\top\mathrm{x}))\| + \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_s^\top\mathrm{x}))\| \\
&\qquad + \sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}} \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le \sqrt{m}\|\mathrm{w}_s-\mathrm{w}_{s,0}\| + \sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}} \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le C_0 + \sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}} \|\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \qquad (6)
\end{aligned}
$$

Now we start bounding the second term of equation (6). We have that with probability at least $1-\gamma$,

$$
\begin{aligned}
&\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})) \\
&\overset{(d)}{=} \sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}} \left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))^2}\right)\mathrm{z} \quad (\mathrm{z} := a_s\mathrm{P}\mathrm{w}_{s,0}\sim N(0,\mathrm{I}_{d-1})) \\
&= \left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))^2}\right)\mathrm{z} \\
&= \sqrt{\frac{1}{m}|S_v|}\cdot\mathrm{z} \qquad \text{(By the definition of } S_v \text{ in Lemma B.4 with } V=\tfrac{C_0}{\sqrt{m}}) \\
&\le \sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(1/\gamma)}{2}})}\cdot\mathrm{z} \qquad (7)
\end{aligned}
$$

Using equation (2), we see that with probability at least $1-\gamma$, one has for all $1\le s\le m$,

$$\|\mathrm{z}\| = \|a_s\mathrm{P}\mathrm{w}_{s,0}\| \le \sqrt{d+4\sqrt{d\log(2m/\gamma)}}$$

Thus follow from equation (6), we have

$$\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \le C_0+\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(2/\gamma)}{2}})}\sqrt{d+4\sqrt{d\log(4m/\gamma)}}$$

And as a result, with probability at least $1-\gamma$

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \left(\frac{1}{2} - (\sqrt{\frac{2\log(4/\gamma)}{m}} + \frac{\log(4/\gamma)}{m})\right)^{1/2}\left(d - 5\sqrt{d\log(8/\gamma)}\right)^{1/2}$$
$$- C_0 - \sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d + 4\sqrt{d\log(8m/\gamma)}} \geq \Omega(\sqrt{d})$$

□

Now we start bounding $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\|$. In order to prove Lemma B.9, we will leverage the fact that for any $h \in \mathbb{R}^d$, we have $\|h\| = \sup_{r\in\mathbb{S}^{d-1}} r\cdot h$. We first start a lemma with fixed r, δ to prove Lemma B.8, then use the ε-net argument to prove Lemma B.9.

Lemma B.8. Fix $r \in \mathbb{S}^{d-1}$, and $\delta \in \mathbb{R}^d$ such that $\|\delta\| \leq R \leq \frac{C_1}{\sqrt{d}}$, C_1 is a constant. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$$
$$\leq 2\sqrt{\log(4/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{\log(4/\gamma)}{m}}\right) + C_0$$
$$+ 3\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d+4\sqrt{d\log(8m/\gamma)}}+5(C_0+\sqrt{\log(4/\gamma)})\cdot\sqrt{\log(4m/\gamma)}$$

Proof of Lemma B.8. Let $X_{s,0} = a_s(\mathrm{w}_{s,0}\cdot r)(\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))), X_s = a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$. We have $\mathbb{E}[X_{s,0}] = 0$, and equation (1) gives us $\mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^{2q}] \leq \frac{(2q)!}{2}$. Thus, we have

$$\begin{aligned}
\mathbb{E}[|X_{s,0}|^q] &\leq \mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^q|\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))|^q] \\
&\leq \sqrt{\mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^{2q}]\cdot P(\mathrm{sign}(\mathrm{w}_{s,0}^\top \mathrm{x}) \neq \mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))} \qquad \text{(Using equation (3))} \\
&\leq \sqrt{\frac{(2q)!}{2}}\cdot\sqrt{R\sqrt{2\log(d)} + \frac{1}{d}} \\
&\leq \frac{q!}{2}2^q\cdot\sqrt{2R\sqrt{2\log(d)}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[|X_{s,0}|^2] &\leq \sqrt{\mathbb{E}[(\mathrm{w}_{s,0}\cdot r)^4]\cdot P(\mathrm{sign}(\mathrm{w}_{s,0}^\top \mathrm{x}) \neq \mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))} \\
&\leq 2\sqrt{R\sqrt{2\log(d)}} \qquad (\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|X|^4] \leq 3!! \leq 4) \\
&\leq 4\sqrt{R\sqrt{\log(d)}}
\end{aligned}$$

Using Theorem B.1 with $v = 4m\sqrt{R\sqrt{\log(d)}}, c = 2$, we have

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{s=1}^{m} X_{s,0} &\leq \sqrt{8\sqrt{R\sqrt{\log(d)}}\log(1/\gamma)} + \frac{2\log(1/\gamma)}{\sqrt{m}} \\
&= 2\sqrt{\log(1/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{\log(1/\gamma)}{m}}\right)
\end{aligned}$$

Now we start bounding $|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}(X_s - X_{s,0})|$. In fact, here we no longer fix r, δ and directly bound $|\sup_{r\in\mathbb{S}^{d-1},\delta\in\mathbb{R}^d,\|\delta\|\leq\frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}}\sum_{s=1}^{m}(X_s - X_{s,0})|$.

$$
\begin{aligned}
&\Big| \sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m (X_s - X_{s,0})\Big| \\
&= \Big| \sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\Big(\mathrm{w}_s^\top r(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big)\Big| \\
&\le \Big| \sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\Big(\mathrm{w}_s^\top r(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))\Big)\Big| \\
&\quad + \Big| \sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\Big(\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big)\Big| \\
&\hfill \text{(Triangle Inequality)} \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^m \|\mathrm{w}_s-\mathrm{w}_{s,0}\| |\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))| + \Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| \\
&\quad + \Big\| \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big\| \\
&\hfill \text{(Cauchy-Schwarz, triangle Inequality)} \\
&\le C_0 + \Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| \qquad (8) \\
&\quad + \Big\| \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big\| \qquad (9)
\end{aligned}
$$

Define $\mathrm{P} = \mathrm{I}_d - \mathrm{x}\mathrm{x}^\top$ as the projection on the orthogonal complement of the span of x. For the second term of equation (9), we have

$$
\begin{aligned}
&\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| \\
&\le \Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{P}\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| + \Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{x}\mathrm{x}^\top \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| \\
&\le \Big\| \sup_{\substack{\forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le \frac{C_0}{\sqrt{m}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| \\
&\quad + \frac{1}{\sqrt{m}}\sum_{s=1}^m \|\mathrm{x}\mathrm{x}^\top \mathrm{w}_{s,0}\| \cdot |\sigma'(\mathrm{w}_s^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x})| \\
&\le \Big\| \sup_{\substack{\forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le \frac{C_0}{\sqrt{m}}}} \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top \mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top \mathrm{x}))\Big\| + \frac{1}{\sqrt{m}}|S_v| \max_{1\le s\le m} \|\mathrm{x}\mathrm{x}^\top \mathrm{w}_{s,0}\| \\
&\hfill \text{(By the definition of } S_v \text{ in Lemma B.4, Hölder's inequality)} \\
&\le \sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d + 4\sqrt{d\log(8m/\gamma)}} + (C_0 + \sqrt{\log(4/\gamma)}) \cdot \sqrt{\log(4m/\gamma)} \\
&\hfill \text{(Equation (7), } \mathrm{x}^\top \mathrm{w}_{s,0} \sim N(0,1))
\end{aligned}
$$

where the last line holds because with probability at least $1-\gamma$, one has for all $1 \le s \le m$, $|\mathrm{x}^\top \mathrm{w}_{s,0}| \le \sqrt{\log(m/\gamma)}$.

Define $\mathrm{P}' = \mathrm{I}_d - \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2}$ as the projection on the orthogonal complement of the span of $\mathrm{x}+\delta$. We bound the third term of equation (9) the same way as follows.

$$
\begin{aligned}
&\Bigg\| \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Bigg\| \\
&\le \Bigg\| \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s P' \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Bigg\| \\
&\quad + \Bigg\| \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Bigg\| \\
&\le \Bigg\| \sup_{\substack{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}} \\ \forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d, \|\mathrm{v}_s\|\le \frac{C_0}{\sqrt{m}}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s P' \mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Bigg\| \\
&\quad + \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m |\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))| \Bigg\| \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0}\Bigg\| \\
&\le \sqrt{\sup_{\substack{\forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d, \\ \|\mathrm{v}_s\|\le \frac{C_0}{\sqrt{m}}}} \frac{1}{m} \sum_{s=1}^m (\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))} \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \|\mathrm{z}(\delta)\| \\
&\qquad\qquad \text{(Define } \mathrm{z}(\delta) := a_s \mathrm{P}'\mathrm{w}_{s,0} \sim N(0, \mathrm{I}_{d-1})) \\
&\quad + \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} |S_v'| \Bigg\| \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0}\Bigg\| \qquad \text{(By definition of } S_v' \text{ in Lemma B.4)} \\
&\le \sqrt{\frac{1}{m}|S_v'|} \cdot \sup_{\delta\in\mathbb{R}^d, \|\delta\|\le \frac{C_1}{\sqrt{d}}} \|\mathrm{z}(\delta)\| \\
&\quad + \frac{1}{\sqrt{m}} |S_v'| \cdot 2 \max_{1\le s\le m} \left| \mathrm{x}^\top \mathrm{w}_{s,0} + \delta^\top \mathrm{w}_{s,0} \right| \qquad (\|\mathrm{z}(\delta)\| \le \|\mathrm{w}_{s,0}\|) \\
&\le \sqrt{\frac{1+R}{\sqrt{m}}\left(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}\right)} \sqrt{d + 4\sqrt{d\log(8m/\gamma)}} \\
&\quad + (C_0 + \sqrt{\log(4/\gamma)})(1+R) \cdot 2 \left(\sqrt{\log(4m/\gamma)} + \frac{\sqrt{d + 4\sqrt{d\log(8m/\gamma)}}}{\sqrt{d}} \right) \\
&\qquad\qquad (R \le \frac{C_1}{\sqrt{d}}, d > \log(m/\gamma)\log(1/\gamma)) \\
&\le \sqrt{\frac{2}{\sqrt{m}}\left(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}\right)} \sqrt{d + 4\sqrt{d\log(8m/\gamma)}} + 4(C_0 + \sqrt{\log(4/\gamma)}) \left(\sqrt{\log(4m/\gamma)} + 3 \right)
\end{aligned}
$$

As a result, we get

$$
\begin{aligned}
\frac{1}{\sqrt{m}} \sum_{s=1}^m X_s &\le \frac{1}{\sqrt{m}} \sum_{s=1}^m X_{s,0} + \left|\frac{1}{\sqrt{m}} \sum_{s=1}^m (X_s - X_{s,0})\right| \\
&\le 2\sqrt{\log(4/\gamma)} \left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{\log(4/\gamma)}{m}} \right) + C_0 \\
&\quad + 3\sqrt{\frac{1}{\sqrt{m}}\left(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}\right)} \sqrt{d + 4\sqrt{d\log(8m/\gamma)}}
\end{aligned}
$$

$$+ 5(C_0 + \sqrt{\log(4/\gamma)})\left(\sqrt{\log(4m/\gamma)} + 3\right)$$

□

Lemma B.9. Let $\|\delta\| \le R \le \frac{C_1}{\sqrt{d}}, \bar{C} = \max\{1, C_0\}, m \ge d^{2.4}$. Then, for any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d:\|\delta\|\le R}} \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))) \le 9\left(C_1 d^2\log^2(md)\sqrt{\frac{\log(d)}{d}}\right)^{1/4} + \frac{15d\log(md)}{\sqrt{m}} + 327\bar{C}C_0 d^{0.25}$$

provided that $d \ge \log(4m/\gamma)^2\log(4/\gamma)^2$. Particularly, for any $c > 0$, there exists C, C' such that if $\log^2(md)\sqrt{\frac{\log(d)}{d}} \le C$, $\frac{d\log(md)}{\sqrt{m}} \le C'\sqrt{d}$, then we have,

$$\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le R}} \|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\| \le c\sqrt{d}$$

The above can realize when $m \le O(\exp(d^{0.24}))$

Proof. Define $\Phi(r,\delta) = \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top\mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$, and N_ε an ε-net for $\Omega = \{(r,\delta), \|r\| = 1, \|\delta\| \le R\}$. In Lemma B.8 we bound $\Phi(r,\delta)$ for a fixed r and δ, here we bounded it uniformly over Ω. We know that $|N_\varepsilon| \le (10/\varepsilon)^{2d}$. Using Lemma B.8, we obtain with probability at least $1-\gamma$,

$$\begin{aligned}
\sup_{(r,\delta)\in\Omega}\Phi(r,\delta) &\le \sup_{(r,\delta)\in N_\varepsilon}\Phi(r,\delta) + \sup_{(r,\delta),(r',\delta')\in\Omega:\|r-r'\|+\|\delta-\delta'\|\le\varepsilon}|\Phi(r,\delta)-\Phi(r',\delta')| \\
&\le 2\sqrt{2d\log(10/\varepsilon)+\log(4/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{2d\log(10/\varepsilon)+\log(4/\gamma)}{m}}\right) \\
&\quad + C_0 + 3\sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d + 4\sqrt{d\log(8m/\gamma)}} \\
&\quad + 5(C_0+\sqrt{\log(4/\gamma)})\cdot\left(\sqrt{\log(4m/\gamma)}+3\right) + \sup_{\substack{(r,\delta),(r',\delta')\in\Omega:\\ \|r-r'\|+\|\delta-\delta'\|\le\varepsilon}}|\Phi(r,\delta)-\Phi(r',\delta')|
\end{aligned} \tag{10}$$

For r, r', one has

$$|\Phi(r,\delta) - \Phi(r',\delta)| \le \frac{\|r-r'\|}{\sqrt{m}}\sum_{s=1}^m\|\mathrm{w}_s\| \le \frac{\|r-r'\|}{\sqrt{m}}\sum_{s=1}^m(\|\mathrm{w}_{s,0}\| + \frac{C_0}{\sqrt{m}})$$

Using (2), we know with probability at least $1-\gamma$, one has for all $s \in [m]$,

$$\|\mathrm{w}_{s,0}\|^2 \le d + 4\sqrt{d\log(m/\gamma)},$$

so that in this event we have,

$$|\Phi(r,\delta) - \Phi(r',\delta)| \le \|r-r'\|\left(\sqrt{md + 4m\sqrt{d\log(m/\gamma)}} + C_0\right)$$

On the other hand, for δ, δ', we write

$$\begin{aligned}
&|\Phi(r,\delta) - \Phi(r,\delta')| \\
&\quad \le \frac{1}{\sqrt{m}}\left|\sum_{s=1}^m \mathbb{1}\{\mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) > \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\}a_s\mathrm{w}_s\cdot r\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{\sqrt{m}} \left| \sum_{s=1}^{m} \mathbb{1}\{\text{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) < \text{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\} a_s \mathrm{w}_s \cdot r \right| \\
\leq \frac{1}{\sqrt{m}} &\left\| \sup_{\substack{\forall s \in [m], \\ \mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \leq V}} \sum_{s=1}^{m} \mathbb{1}\{\text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) > \text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\} a_s \mathrm{w}_s \right\| \\
&+ \frac{1}{\sqrt{m}} \left\| \sup_{\substack{\forall s \in [m], \\ \mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \leq V}} \sum_{s=1}^{m} \mathbb{1}\{\text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) < \text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\} a_s \mathrm{w}_s \right\|
\end{aligned}
$$

Letting $X_s(\delta) = \mathbb{1}\{\exists \delta' : \|\delta - \delta'\| \leq \varepsilon \text{ and } \exists \mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \leq V, \text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) \neq \text{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\}$, we now control with exponentially high probability $\sum_{s=1}^{m} X_s(\delta)$. By (4) in Lemma B.6, we know that $X_s(\delta)$ is a Bernoulli of parameter at most $p = 2\varepsilon(\sqrt{d} + 2\sqrt{d\log(2/\varepsilon)}) + (1+R+\varepsilon)V$. Using Theorem B.2 and apply a union bound, we have

$$
P\left(\exists (r,\delta) \in N_\varepsilon : \sum_{s=1}^{m} X_s(\delta) \geq k\right) \leq \left(\frac{10}{\varepsilon}\right)^{2d} \exp(-pm) \left(\frac{epm}{k}\right)^k := g(p)
$$

We would like $g(p) \leq \gamma$. Since $\sqrt{\varepsilon}(1 + 2\sqrt{\log(2/\varepsilon)}) \leq 4\varepsilon^{3/8}$ holds for $\varepsilon \in (0,1)$, we keep upper bound $p \leq 8\sqrt{d}\varepsilon^{7/8} + 3V$. Note that $p \leq \frac{k}{m}$ to make sure $g(p)$ increases as p increases. Choose $\varepsilon = m^{-4/7}d^{-4/7}$, $V = \frac{C_0}{\sqrt{m}}$, $\bar{C} = \max\{C_0, 1\}$, $k = 44\bar{C}\sqrt{m}$, with $m \geq 58, m \geq d^{2.4}$, we have

$$
\begin{aligned}
\log g(8\sqrt{d}\varepsilon^{7/8} + \frac{3C_0}{\sqrt{m}}) &= -pm + 2d\log\left(\frac{10}{\varepsilon}\right) + 44\bar{C}\sqrt{m}\log\left(\frac{epm}{44\bar{C}\sqrt{m}}\right) \\
&\qquad\qquad \text{(Here } p = 8\sqrt{d}\varepsilon^{7/8} + \tfrac{3C_0}{\sqrt{m}} \leq \tfrac{11\bar{C}}{\sqrt{m}}\text{)} \\
&= -(8\sqrt{d}\varepsilon^{7/8} + \frac{3C_0}{\sqrt{m}})m + 2d\log(10m^{4/7}d^{4/7}) + 44\bar{C}\sqrt{m}\log\left(\frac{11e\bar{C}\sqrt{m}}{44\bar{C}\sqrt{m}}\right) \\
&\leq -8\sqrt{m} + 2.3d\log(md) \qquad (m \geq 58) \\
&\leq -8\sqrt{m} + 7\sqrt{m} \qquad (m \geq d^{2.4} \implies \sqrt{m} \geq 0.33d\log(md)) \\
&\leq -\sqrt{m} \\
&\leq \log(\gamma)
\end{aligned}
$$

The last line holds for $m \geq d^{2.4}$ and $d \geq \log(4m/\gamma)^2 \geq \log(4/\gamma)^2$.

By the concentration of Lipschitz functions of Gaussians (Theorem B.3) and a union bound, we have

$$
P\left(\exists S \in [m], |S| \leq 44\bar{C}\sqrt{m}, \left\|\frac{1}{\sqrt{m}}\sum_{s\in S} a_s \mathrm{w}_{s,0}\right\| \geq \sqrt{\frac{|S|}{m}}(\sqrt{d}+t)\right) \leq m^{44\bar{C}\sqrt{m}} e^{-\frac{t^2}{2}}
$$

By setting $t = \sqrt{88\bar{C}\sqrt{m}\log(m) + 2\log(\frac{8}{\gamma})}$, we get that with probability at least $1 - \gamma/8$,

$$
\forall S \subset [m], |S| \leq 44\bar{C}\sqrt{m} : \left\|\frac{1}{\sqrt{m}}\sum_{s\in S} a_s \mathrm{w}_{s,0}\right\| \leq \sqrt{\frac{44\bar{C}}{\sqrt{m}}}\sqrt{88\bar{C}\sqrt{m}\log(m) + 2\log(\frac{8}{\gamma})} + \sqrt{\frac{44\bar{C}d}{\sqrt{m}}}
$$

$$
\begin{aligned}
\forall S \subset [m], |S| \le 44\bar{C}\sqrt{m} : & \left\| \sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{w}_s \right\| \\
& \le \left\| \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{w}_{s,0} \right\| + \left\| \sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{v}_s \right\| \\
& \le \sqrt{\frac{44\bar{C}}{\sqrt{m}}} \sqrt{88\bar{C}\sqrt{m}\log(m) + 2\log(\frac{8}{\gamma})} + \sqrt{\frac{44\bar{C}d}{\sqrt{m}}} + \frac{44\bar{C}C_0}{\sqrt{m}} \\
& \le 113\bar{C}C_0 \sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}
\end{aligned}
$$

Note that for all $(r, \delta) \in N$, $\|\delta - \delta'\| \le \varepsilon$,

$$
\sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \mathbb{1}\{\mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top (\mathrm{x} + \delta)) < \mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top (\mathrm{x} + \delta'))\} \le X_s(\delta)
$$

$$
\sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \mathbb{1}\{\mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top (\mathrm{x} + \delta)) > \mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top (\mathrm{x} + \delta'))\} \le X_s(\delta)
$$

With probability at least $1 - \gamma$, we have for all δ, δ', r, r' with $\|\delta - \delta'\| \le m^{-4/7} d^{-4/7}$ and $\|r - r'\| \le m^{-4/7} d^{-4/7}$,

$$
\begin{aligned}
|\Phi(r, \delta) - \Phi(r', \delta)| &\le m^{-4/7} d^{-4/7} \left(\sqrt{md + 4m\sqrt{d\log(m/\gamma)}} + C_0 \right) \\
|\Phi(r, \delta) - \Phi(r, \delta')| &\le 226\bar{C}C_0 \sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}
\end{aligned}
$$

Combining this with (10) we obtain with probability at least $1 - \gamma$,

$$
\begin{aligned}
& \sup_{(r,\delta) \in \Omega} \Phi(r, \delta) \\
& \le 2\sqrt{2d\log(10m^{4/7}d^{4/7}) + \log(4/\gamma)} \left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{2d\log(10m^{4/7}d^{4/7}) + \log(4/\gamma)}{m}} \right) \\
& \quad + C_0 + 3\sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})} \sqrt{d + 4\sqrt{d\log(8m/\gamma)}} \\
& \quad + 5(C_0 + \sqrt{\log(4/\gamma)}) \cdot \left(\sqrt{\log(4m/\gamma)} + 3 \right) + m^{-4/7} d^{-4/7} \left(\sqrt{md + 4m\sqrt{d\log(m/\gamma)}} + C_0 \right) \\
& \quad + 226\bar{C}C_0 \sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})} \\
& \le 6\sqrt{d\log(md)} \left((4R\sqrt{\log(d)})^{1/4} + \sqrt{\frac{6d\log(md)}{m}} \right) + 50d^{0.25} + 227\bar{C}C_0 \sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})} \\
& \le 9\left(C_1 d^2 \log^2(md) \sqrt{\frac{\log(d)}{d}} \right)^{1/4} + \frac{15d\log(md)}{\sqrt{m}} + 327\bar{C}C_0 d^{0.25} \qquad (R \le \tfrac{C_1}{\sqrt{d}}) \\
& = o(\sqrt{d}) \ \ \text{(holds when } \log^2(md)\sqrt{\tfrac{\log(d)}{d}} = o(1), \tfrac{d\log(md)}{\sqrt{m}} = o(\sqrt{d}) \implies m \le O(\exp(d^{0.24}))\text{)}
\end{aligned}
$$

□

Proof of Corollary 3.2. From Theorem 3.1, we have that for any point $\mathrm{x} \in \mathbb{S}^{d-1}$, with probability at least $1-\gamma$, there exists an adversarial example for the neural network, with parameters (W, a), at x for perturbation size R. Thus, with probability at least $1-\gamma$, the robust error of W at x with perturbation R is one: $\ell_R(\mathrm{W}; \mathrm{x}, y) = 1$. This gives us that $\mathbb{E}_{\mathrm{W}}[\sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)] \geq 1-\gamma$. From Markov's inequality,

$$
\begin{aligned}
\mathbb{P}\left[\sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W}) < 0.9\right] &= \mathbb{P}\left[1 - \sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W}) > 0.1\right] \\
&\leq 90\mathbb{E}_{\mathrm{W}}\left[1 - \sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W})\right] \\
&= 10\mathbb{E}_x\mathbb{E}_{\mathrm{W}}\left[1 - \sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)\right] \\
&\leq 10 \sup_x \mathbb{E}_{\mathrm{W}}\left[1 - \sup_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)\right] \\
&\leq 10\gamma
\end{aligned}
$$

where in the second equality we swap the expectations using Fubini's theorem. □