Math. Stat. Learn. 6 (2023), 143–200
DOI 10.4171/MSL/41

Adversarial examples in random neural networks with general activations

Andrea Montanari and Yuchen Wu

Abstract. A substantial body of empirical work documents the lack of robustness in deep learning models to adversarial examples. Recent theoretical work proved that adversarial examples are ubiquitous in two-layers networks with sub-exponential width and ReLU or smooth activations, and multi-layer ReLU networks with sub-exponential width. We present a result of the same type, with no restriction on width and for general locally Lipschitz continuous activations.

More precisely, given a neural network $f(\cdot;\boldsymbol{\theta})$ with random weights $\boldsymbol{\theta}$, and feature vector $\boldsymbol{x}$, we show that an adversarial example $\boldsymbol{x}'$ can be found with high probability along the direction of the gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta})$. Our proof is based on a Gaussian conditioning technique. Instead of proving that f is approximately linear in a neighborhood of $\boldsymbol{x}$, we characterize the joint distribution of $f(\boldsymbol{x};\boldsymbol{\theta})$ and $f(\boldsymbol{x}';\boldsymbol{\theta})$ for $\boldsymbol{x}' = \boldsymbol{x} - s(\boldsymbol{x})\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta})$, where $s(\boldsymbol{x}) = \operatorname{sign}(f(\boldsymbol{x};\boldsymbol{\theta}))\cdot s_d$ for some positive step size s_d.

1. Introduction

The output of a neural network at test time can be significantly changed by an imperceptible but carefully chosen perturbation of its input. Such perturbed inputs are referred to as *adversarial examples*. In the context of deep learning, the existence of adversarial examples was first discovered experimentally in [23]. A rapidly expanding literature developed algorithms to produce adversarial examples [7, 11, 13, 19, 28], as well as techniques to increase model robustness [6, 14, 20, 21, 24, 27].

Throughout this paper, we will focus on the standard supervised learning setting, whereby a data sample takes the form $(\boldsymbol{x}, y)$, with $\boldsymbol{x} \in \mathbb{R}^d$ a covariates vector and $y \in \mathbb{R}$ the corresponding label. A model is a function $f(\cdot;\boldsymbol{\theta})\colon \mathbb{R}^d \to \mathbb{R}$ parametrized by weights $\boldsymbol{\theta} \in \mathbb{R}^p$. In this setting, given a test point $\boldsymbol{x} \in \mathbb{R}^d$, an adversary constructs $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x}^{\text{adv}}(\boldsymbol{x};\boldsymbol{\theta}) \in \mathbb{R}^d$. The adversary is successful if, with high probability

$$\operatorname{sign}\big(f(\boldsymbol{x}^{\text{adv}};\boldsymbol{\theta})\big) = -\operatorname{sign}\big(f(\boldsymbol{x};\boldsymbol{\theta})\big), \quad \|\boldsymbol{x}^{\text{adv}} - \boldsymbol{x}\| \ll \|\boldsymbol{x}\|.$$

2020 Mathematics Subject Classification. Primary 68T07; Secondary 68Q32.
Keywords. Adversarial example, neural network.

In this paper we will interpret 'with high probability' as with probability converging to one as $d \to \infty$ with respect to a certain distribution over the random weights $\boldsymbol{\theta}$, for any fixed $\boldsymbol{x} \in \mathbb{R}^d$ (with normalization $\|\boldsymbol{x}\| = \sqrt{d}$). Results in the literature differ in the choice of the point $\boldsymbol{x}$ (e.g., random according to the test distribution, or an arbitrary training point, or a fixed $\boldsymbol{x}$ as in the present paper), and the norm $\|\cdot\|$ (empirical work often adopts ℓ_∞ norm, but we will follow earlier theoretical papers and use ℓ_2 norm.). We refer to the next sections for formal statements.

Among the earlier hypotheses about the origins and ubiquity of adversarial examples was the idea, put forward in [11], that they are related to the fact that $f(\cdot\,;\boldsymbol{\theta})$ is approximately linear (better, affine) over large regions of the input space. This hypothesis has several consequences that match empirical observations at a qualitative level:

(1) Prevalence of adversarial examples. Indeed, if $f(\boldsymbol{x};\boldsymbol{\theta}) \approx a(\boldsymbol{\theta}) + \langle \boldsymbol{b}(\boldsymbol{\theta}), \boldsymbol{x}\rangle$, then

$$f(\boldsymbol{x}^{\text{adv}};\boldsymbol{\theta}) - f(\boldsymbol{x};\boldsymbol{\theta}) \approx \langle \boldsymbol{b}(\boldsymbol{\theta}), \boldsymbol{x}^{\text{adv}} - \boldsymbol{x}\rangle. \tag{1.1}$$

Assuming without loss of generality $a(\boldsymbol{\theta}) = 0$, we have $|f(\boldsymbol{x};\boldsymbol{\theta})| = \Theta(\|\boldsymbol{b}(\boldsymbol{\theta})\|_2)$ for most $\|\boldsymbol{x}\|_2 = \sqrt{d}$. By choosing

$$\boldsymbol{x}^{\text{adv}} - \boldsymbol{x} = \pm \boldsymbol{b}(\boldsymbol{\theta})/\|\boldsymbol{b}(\boldsymbol{\theta})\|_2,$$

we obtain that $|f(\boldsymbol{x}^{\text{adv}};\boldsymbol{\theta}) - f(\boldsymbol{x};\boldsymbol{\theta})|$ is of order $|f(\boldsymbol{x};\boldsymbol{\theta})|$, while $\|\boldsymbol{x}^{\text{adv}} - \boldsymbol{x}\|_2 = 1 \ll \|\boldsymbol{x}\|_2$. With appropriate choice of sign and step size, such perturbation also flips the sign of f.

(2) Adversarial examples can be found by efficient algorithms. Indeed, the above argument suggests to take

$$\boldsymbol{x}^{\text{adv}}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{x} - s(\boldsymbol{x})\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta}), \tag{1.2}$$

for a suitable $s(\boldsymbol{x})$. This approach was successfully implemented in [11], who referred[1] to it as the 'fast gradient sign method' (FGSM).

The main result of this paper is a proof that this procedure indeed produces adversarial examples when $f(\boldsymbol{x};\boldsymbol{\theta})$ is a two-layer or multi-layer fully connected neural network with random weights. This can be interpreted as the function implemented by the network at initialization.

Several groups obtained theoretical results on the existence of adversarial examples. One basic remark is that, if the distribution of the covariates $\boldsymbol{x}$ satisfies an isoperimetry property, and $\mathbb{P}_{\boldsymbol{x}}(f(\boldsymbol{x};\boldsymbol{\theta}) > 0) \in [\delta, 1-\delta]$ for some constant $\delta > 0$, then a

[1]The original proposal attempted to minimize $\|\boldsymbol{x}^{\text{adv}} - \boldsymbol{x}\|_\infty$, and consequently takes $\boldsymbol{x}^{\text{adv}} - \boldsymbol{x} \propto \text{sign}(\nabla_{\boldsymbol{x}} f)$.

random $\boldsymbol{x}$ will be close to the decision boundary (and to an adversarial example) with high probability. This is the case, for instance, when $\boldsymbol{x}$ is a uniformly random vector on a high-dimensional sphere, or a standard Gaussian vector. Increasingly sophisticated incarnations of this argument were given in [9, 10, 22].

The isoperimetry argument clarifies why adversarial examples are ubiquitous, but does not explain why they can be found so easily, for instance via FGSM. In the other direction, [5] proved that learning robust classifiers can be computationally hard.

A somewhat different point of view was developed in [12], which proposed that non-robustness is related to the presence of non-robust features in the data. These functions $h(\boldsymbol{x})$ of the data are used by a normally trained classifier (minimizing the training error), but can be significantly changed by an imperceptible perturbation of $\boldsymbol{x}$. By itself, this is not incompatible with the 'approximate linearity' hypothesis described above. However, [12] emphasized the existence of robust features alongside non-robust ones.

Our work is most closely related to a recent sequence of papers analyzing the brittleness of fully connected neural networks to the FGSM-style attack (1.2) [2, 4, 8]. In particular, [8] showed that random ReLU networks are vulnerable if the width of each layer is small relative to the width of the previous layer. For the case of two-layer networks, this result was improved in [4] which considered either smooth activations and width subexponential in the input dimension $m = \exp(o(d))$, or ReLU activations, and width $m \leq \exp(d^{0.24})$. Finally, [2] generalized the latter analysis to multi-layer networks with maximal width $m \leq \exp(d^c)$ for some small c.

We also point to the recent paper [26] which studied trained two-layer ReLU networks (under the assumption that gradient flow converges to a network that perfectly classifies the training set). These are shown to be non-robust (in a stronger sense than above) for sample size $n \leq \sqrt{d}$.

In this paper, we prove that the FGSM-like attack (1.2) indeed finds adversarial examples for neural networks with random Gaussian weights. We present the following novel contributions, with respect to earlier work:

Arbitrary width. Our results apply to an arbitrary diverging width, without upper bounds on the growth rate. This question is posed as an open problem in [4] and is not merely academic. A large body of literature connects wide random neural networks to Gaussian processes and kernel methods, see [1, 3, 15, 18] for a few pointers. For instance [16, 17] prove that the generalization properties of two-layer networks linearized around their initialization approach the one of the associated infinite-width kernel as soon as the number of parameters becomes larger than the number of samples.

Within this context, the upper bounds on width assumed in [2, 4] are somewhat puzzling. A priori, they could suggest that exponentially wide networks are more robust than sub-exponentially wide, although their generalization properties are similar. Here we prove that this is not the case. In particular, our results apply to Gaussian processes as well.

General activation. Both for two-layer and multi-layer networks, our proof applies for a general class of activation functions $\sigma(x)$. We only require $\sigma'(x)$ to exist almost everywhere, continuous, and bounded by a polynomial.

While most activation function of practical use are more regular than this (e.g., Lipschitz continuous), this generalization clarifies that the approximate linearity property (1.1) is not a naive consequence of the smoothness of the activation functions.

Weak linearity condition. Our proofs are based on a weaker notion of linearity than [2, 4, 8]. Namely, instead of proving that $f(\,\cdot\,;\boldsymbol{\theta})$ is approximately linear in a neighborhood of $\boldsymbol{x}$, we only prove that it is approximately linear along the direction of interest $\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta})$.

Gaussian conditioning. Establishing approximate linearity along a specific direction poses an obvious mathematical challenge: The direction $\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{\theta})$ is correlated with the function $f(\,\cdot\,;\boldsymbol{\theta})$ itself. We deal with this difficulty by introducing a Gaussian conditioning technique that was not used before in this context, and we believe can be useful to study other attacks.

For clarity of exposition, we will treat separately the two-layer and multi-layer cases. This allows the reader to understand the proof strategy in a simpler example, before diving into the notational intricacies of multi-layer networks. In the case of two-layer networks, we prove two theorems. The first one is stated in Section 2, and establishes that the attack succeeds with probability converging to one, but does not provide explicit probability bounds. On the other hand, this theorem holds for very general activation functions. We present the proof of these results in Section 3. We then state a complementary result in Section 4 which gives explicit non-asymptotic probability bounds, but limited to Lipschitz activations. The result for multi-layer network is stated in Section 5 with proofs in Section 6. Several technical lemmas are deferred to the appendices.

1.1. Notations

We generally use lowercase letters for scalars, lowercase bold for vectors and uppercase bold for matrices. The ordinary scalar product of vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is denoted by $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$, and we let $\|\boldsymbol{u}\|_2 := \langle \boldsymbol{u}, \boldsymbol{u}\rangle^{1/2}$. For $n \in \mathbb{N}_+$, we let $[n] = \{1, 2, \dots, n\}$. We denote by p-lim convergence in probability. For random variables X, Y, we denote by $X \perp Y$ if X and Y are independent of each other. We denote by p-lim convergence in probability.

2. Random two-layer networks

We begin by studying the following random two-layer neural network:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i \sigma(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x}). \tag{2.1}$$

Here, $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a fixed activation function, $\boldsymbol{\theta} = \{(a_i, \boldsymbol{w}_i)\}_{i\le m}$, the weight vectors $\boldsymbol{w}_i \in \mathbb{R}^d$ are i.i.d. generated from $\mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$, and $(a_i)_{i\le m} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/m)$. We denote by $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ the random weight matrix with the i-th row equal to $\boldsymbol{w}_i$, $\boldsymbol{a} \in \mathbb{R}^m$ the vector with the i-th coordinate equal to a_i. We assume that $\boldsymbol{W}$ is independent of $\boldsymbol{a}$. In what follows, we shall typically drop the argument $\boldsymbol{\theta}$ from f, and write $f(\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta})$.

Let $\tau := \operatorname{sign}(f(\boldsymbol{x}))$, and $s_d \in \mathbb{R}^+$ be the step size which depends on (m, d). We define

$$\boldsymbol{x}^s := \boldsymbol{x} - \tau s_d \nabla f(\boldsymbol{x}). \tag{2.2}$$

Our main result on two-layer networks establishes that there exists a sequence of step sizes $\{s_d\}_{d\ge 1}$, such that for $d, m = m(d) \to \infty$, $\|\boldsymbol{x} - \boldsymbol{x}^s\|_2/\|\boldsymbol{x}\|_2 \overset{\mathbb{P}}{\to} 0$, while with high probability $\operatorname{sign}(f(\boldsymbol{x})) \neq \operatorname{sign}(f(\boldsymbol{x}^s))$.

Theorem 2.1. *Let $\boldsymbol{x} \in \mathbb{R}^d$ be a deterministic vector with $\|\boldsymbol{x}\|_2 = \sqrt{d}$. Assume that $\sigma(x)$ is not a constant, σ is continuous, almost everywhere differentiable, σ' is almost everywhere continuous, and $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$, where $k \in \mathbb{N}_+$ is a fixed positive integer and $C_\sigma > 0$ is a constant depending only on σ.*

Then the following hold:

(1) *There exists a constant $C > 0$ depending only on σ, such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^s\|_2}{\|\boldsymbol{x}\|_2} \le \frac{C s_d}{\sqrt{d}}(1 + d^{-1/2}\log(1/\delta))(1 + (m\delta)^{-1/2}).$$

(2) *Let $\{\xi_d\}_{d\in\mathbb{N}_+} \subseteq \mathbb{R}^+$ be an increasing sequence such that $\xi_d \to \infty$ as $d \to \infty$. Then there exists $\{s_d\}_{d\in\mathbb{N}_+} \subseteq \mathbb{R}^+$, such that $s_d \le \xi_d$ and the following hold:*

$$\operatorname{p-lim}_{m,d\to\infty} \frac{\|\boldsymbol{x} - \boldsymbol{x}^s\|_2}{\|\boldsymbol{x}\|_2} = 0,$$
$$\lim_{m,d\to\infty} \mathbb{P}(\operatorname{sign}(f(\boldsymbol{x})) \neq \operatorname{sign}(f(\boldsymbol{x}^s))) = 1.$$

Remark 2.1. Note that this theorem provides a completely quantitative non-asymptotic upper bound on the size of the perturbation $\|\boldsymbol{x} - \boldsymbol{x}^s\|_2$. On the other hand, it does not provide convergence rates for the success probability $\mathbb{P}(\operatorname{sign}(f(\boldsymbol{x})) \neq \operatorname{sign}(f(\boldsymbol{x}^s)))$.

This can be traced to the use of a non-quantitative uniform central limit theorem in our proof (see below). In Section 4 we obtain an explicit rate by a more careful handling of that step, at the price of assuming Lipschitz activations.

Remark 2.2. The scale of the input is chosen so that the output after the first layer $\sigma(\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x})$ are of order one when $\boldsymbol{w}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$. As a consequence, the output after the second layer $f(\boldsymbol{x};\boldsymbol{\theta})$ is also of order one for $a_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/m)$.

The proof of the theorem applies with barely any change to all inputs satisfying $c\sqrt{d} \le \|\boldsymbol{x}\|_2 \le C\sqrt{d}$ for some positive constants c, C. Here, we choose not to state this general version to simplify notations.

We note that the scaling is unimportant when the activation function is positively homogeneous (that is to say, for $z > 0$, $\sigma(zx) = z\sigma(x)$). On the other hand, for general activations the sensitivity to input perturbations is necessarily dependent of the scale. For instance, if $\sigma(t) = (t-1)_+$ and $\|\boldsymbol{x}\|_2 \ll \sqrt{d}$ then, with high probability we have $|\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x}| \ll 1$ for all $i \le m$. In other words, the inputs to the hidden neurons lie in the region in which the activation vanishes, and therefore a small perturbation will not change the network output.

2.1. Proof technique

As mentioned in the introduction, our proofs are based on Gaussian calculus. Before delving into the actual calculation, it is perhaps useful to describe a simple calculation along the same lines. Let $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_2 = \sqrt{d}$ be fixed and the adversarial example be given by equation (2.2), with $f(\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta})$, the two-layer network of equation (2.1).

The gradient of f is given by

$$\nabla f(\boldsymbol{x}) = \boldsymbol{W}^\mathsf{T}\boldsymbol{D}_\sigma \boldsymbol{a}, \tag{2.3}$$

where $\boldsymbol{D}_\sigma \in \mathbb{R}^{m\times m}$ is a diagonal matrix: $\boldsymbol{D}_\sigma = \mathrm{diag}(\{\sigma'(\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x})\}_{i\le m})$.

Note that $\nabla f(\boldsymbol{x})$ is a random vector, because $\boldsymbol{W}, \boldsymbol{a}$ are random. It is clearly useful to understand the distribution of this random vector: this will tell us about the properties of the adversarial perturbation. It is elementary that, if $\boldsymbol{b}$ is a random vector independent of $\boldsymbol{W}$, then conditional on the norm $\|\boldsymbol{b}\|_2$, $\boldsymbol{W}^\mathsf{T}\boldsymbol{b}$ is a Gaussian random vector:

$$\boldsymbol{W}^\mathsf{T}\boldsymbol{b}\Big|_{\|\boldsymbol{b}\|_2} \sim \mathsf{N}(\boldsymbol{0}, (\|\boldsymbol{b}\|_2^2/d)\cdot \boldsymbol{I}_d).$$

Equivalently, we can express the same fact by saying that $\boldsymbol{W}^\mathsf{T}\boldsymbol{b} = \|\boldsymbol{b}\|_2\boldsymbol{z}/\sqrt{d}$, where $\boldsymbol{z}$ is a standard Gaussian vector that is independent of $\boldsymbol{b}$.

This fact might suggest that

$$\nabla f(\boldsymbol{x}) \approx \boldsymbol{z}\cdot \frac{1}{\sqrt{d}}\big\|\boldsymbol{D}_\sigma \boldsymbol{a}\big\|_2, \tag{2.4}$$

where $\boldsymbol{z}$ is a standard Gaussian vector that is independent of everything else. Since (as is easy to see) $\|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2$ concentrates, this would further imply that the gradient is approximately Gaussian with i.i.d. entries, whose variances we can easily compute.

However, the implication is not straightforward because in this case, $\boldsymbol{b} = \boldsymbol{D}_\sigma \boldsymbol{a}$ is not independent of $\boldsymbol{W}$ (because the diagonal elements of $\boldsymbol{D}_\sigma$ are $\sigma'(\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x})$ and are therefore a deterministic function of $\boldsymbol{W}$). As a consequence, equation (2.4) is at best an approximate equality, and quantifying the error requires an argument.

Luckily, the Gaussian distribution allows for a particularly elegant such argument. Let $\Pi_{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$ denote the orthogonal projector onto the linear space spanned by $\boldsymbol{x}$, $\Pi_{\boldsymbol{x}} = \boldsymbol{x}\boldsymbol{x}^\mathsf{T}/d$ and $\Pi_{\boldsymbol{x}}^\perp := \boldsymbol{I}_d - \Pi_{\boldsymbol{x}}$. Further, define $\boldsymbol{g} = \boldsymbol{W}\boldsymbol{x}$ (the input first-layer neurons). Then we have

$$\begin{aligned}\boldsymbol{W} &= \boldsymbol{W}\Pi_{\boldsymbol{x}} + \boldsymbol{W}\Pi_{\boldsymbol{x}}^\perp \\ &= \frac{1}{d}\boldsymbol{g}\boldsymbol{x}^\mathsf{T} + \boldsymbol{W}\Pi_{\boldsymbol{x}}^\perp.\end{aligned}$$

This decomposition has several useful properties: (i) $\boldsymbol{W}\Pi_{\boldsymbol{x}}^\perp$ is independent of $\boldsymbol{g}$ (because two orthogonal projections of a standard normal are independent); (ii) $\boldsymbol{D}_\sigma$ is a function uniquely of $\boldsymbol{g}$; (iii) the distribution of $\boldsymbol{g}$ is simple, namely $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_d)$.

Using this decomposition, we obtain

$$\begin{aligned}\nabla f(\boldsymbol{x}) &= \frac{1}{d}\boldsymbol{x}\boldsymbol{g}^\mathsf{T}\boldsymbol{D}_\sigma \boldsymbol{a} + \Pi_{\boldsymbol{x}}^\perp \boldsymbol{W}^\mathsf{T}\boldsymbol{D}_\sigma \boldsymbol{a} \\ &\overset{(*)}{=} \alpha_\parallel \boldsymbol{x} + \alpha_\perp \Pi_{\boldsymbol{x}}^\perp \boldsymbol{z},\end{aligned} \tag{2.5}$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ is independent of $\boldsymbol{g}$ and

$$\alpha_\parallel := \frac{1}{d}\sum_{i=1}^m g_i a_i \sigma'(g_i), \quad \alpha_\perp^2 := \frac{1}{d}\sum_{i=1}^m a_i^2 \sigma'(g_i)^2.$$

Note that the crucial step $(*)$ is correct because $\Pi_{\boldsymbol{x}}^\perp \boldsymbol{W}^\mathsf{T}$ is independent of $\boldsymbol{g}$ as discussed above. By the law of large numbers, both $\alpha_\parallel = O_P(1/d)$ and $\alpha_\perp$ concentrates around its expectation, which is of order $1/\sqrt{d}$. Therefore, equation (2.5) provides an exact version of equation (2.4).

The basic intuition in the decomposition (2.5) is quite simple. Even if $\boldsymbol{D}_\sigma$ depends on $\boldsymbol{W}$, it only depends on a low-dimensional projection of this matrix. We can condition on this projection, and resample the orthogonal component of $\boldsymbol{W}$ independently from it.

Our proofs push forward the same type of reasoning. Instead of computing the distribution of $\nabla f(\boldsymbol{x})$, we now have to compute the distribution of

$$f(\boldsymbol{x}^s) = f(\boldsymbol{x} - \tau s_d \nabla f(\boldsymbol{x})).$$

By conditioning on suitable projections of $\boldsymbol{W}$, we write this quantity (better, its difference from the first order Taylor approximation) in terms of a certain number of empirical averages similar to $\alpha_{\|}, \alpha_{\perp}$ above. Thanks to such representations, the proofs of our main theorems reduce to controlling these empirical averages and this can be achieved by standard empirical process theory.

3. Proof of Theorem 2.1

As discussed above, our proof strategy is based on conditioning on low-dimensional projections of $\boldsymbol{W}$. We state a Gaussian conditioning lemma that will be used repeatedly throughout the paper.

We say that $\boldsymbol{Y}$ depends on $\boldsymbol{X}$ only through $g(\boldsymbol{X})$ if and only if there exists a deterministic function h and a random vector $\boldsymbol{Z}$ that is independent of $\boldsymbol{X}$, such that $\boldsymbol{Y} = h(g(\boldsymbol{X}), \boldsymbol{Z})$.

Lemma 3.1. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ be a matrix with i.i.d. standard Gaussian entries, and $\boldsymbol{A}_1 \in \mathbb{R}^{k_1\times m}$, $\boldsymbol{A}_2 \in \mathbb{R}^{d\times k_2}$ be other random matrices. Let $\boldsymbol{Y} = h_1(\boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{X}\boldsymbol{A}_2, \boldsymbol{Z}_1)$ and $\boldsymbol{A}_2 = h_2(\boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{Z}_2)$ for deterministic functions h_1 and h_2. Further assume that $(\boldsymbol{X}, \boldsymbol{A}_1, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$ are mutually independent. Then there exists $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m\times d}$ which has the same distribution with $\boldsymbol{X}$ and is independent of $\boldsymbol{Y}$, such that*

$$\boldsymbol{X} = \Pi_{\boldsymbol{A}_1}^{\perp}\widetilde{\boldsymbol{X}}\Pi_{\boldsymbol{A}_2}^{\perp} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{A}_2} + \Pi_{\boldsymbol{A}_1}\boldsymbol{X}\Pi_{\boldsymbol{A}_2}^{\perp} + \Pi_{\boldsymbol{A}_1}\boldsymbol{X}\Pi_{\boldsymbol{A}_2},$$

where $\Pi_{\boldsymbol{A}_1} \in \mathbb{R}^{m\times m}$, $\Pi_{\boldsymbol{A}_2} \in \mathbb{R}^{d\times d}$ are the orthogonal projectors onto the subspaces spanned by the rows of $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, respectively. Further,

$$\Pi_{\boldsymbol{A}_1}^{\perp} := \boldsymbol{I}_m - \Pi_{\boldsymbol{A}_1}, \quad \Pi_{\boldsymbol{A}_2}^{\perp} := \boldsymbol{I}_d - \Pi_{\boldsymbol{A}_2}.$$

The proof of Lemma 3.1 is a straightforward application of the properties of Gaussian ensembles, and we defer it to Section A.1.

Proof of the first claim. We first prove claim (1) of the theorem. Recall that the gradient of f is given by equation (2.3). The next lemma provides non-asymptotic control over the Euclidean norm of $\nabla f(\boldsymbol{x})$.

Lemma 3.2. *Under the conditions of Theorem 2.1, there exists a constant $C > 0$ that depends only on σ, such that for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\|\nabla f(\boldsymbol{x})\|_2 \leq C(1 + d^{-1/2}\log(1/\delta))(1 + (m\delta)^{-1/2}).$$

The proof of Lemma 3.2 is deferred to Section A.2 in the appendix. Recall that $\|\boldsymbol{x}\|_2 = \sqrt{d}$, thus the first claim of the theorem follows directly from Lemma 3.2.

Proof of the second claim. We next invoke Lemma 3.1 to prove the second claim of Theorem 2.1. For the sake of simplicity, we define $\boldsymbol{g} := \boldsymbol{W}\boldsymbol{x}$, $\boldsymbol{g}^s := \boldsymbol{W}\boldsymbol{x}^s$, then $\boldsymbol{g} \stackrel{d}{=} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. By definition, we have

$$\begin{aligned} \boldsymbol{g}^s &= \boldsymbol{g} - \tau s_d \boldsymbol{W} \nabla f(\boldsymbol{x}) \\ &= \boldsymbol{g} - \tau s_d \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{D}_\sigma \boldsymbol{a}. \end{aligned} \tag{3.1}$$

Recall that $\Pi_{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$ is the orthogonal projector onto the linear subspace spanned by $\boldsymbol{x}$, and let $\Pi_{\boldsymbol{x}}^\perp := \boldsymbol{I}_d - \Pi_{\boldsymbol{x}}$. Using Lemma 3.1, we can decompose the weight matrix $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{g}\boldsymbol{x}^\top/d + \widetilde{\boldsymbol{W}}\Pi_{\boldsymbol{x}}^\perp$, where $\widetilde{\boldsymbol{W}}$ has the same marginal distribution as $\boldsymbol{W}$ and is independent of $(\boldsymbol{g}, \boldsymbol{a})$. We then substitute this result into equation (3.1), which gives

$$\begin{aligned} \boldsymbol{g}^s &= \boldsymbol{g} - \tau s_d(\boldsymbol{g}\boldsymbol{g}^\top/d + \widetilde{\boldsymbol{W}}\Pi_{\boldsymbol{x}}^\perp\widetilde{\boldsymbol{W}}^\top)\boldsymbol{D}_\sigma\boldsymbol{a} \\ &= \boldsymbol{g}(1 - \tau s_d \boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}/d) - \tau s_d \overline{\boldsymbol{W}}\,\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a} \\ &= \boldsymbol{g}(1 - \tau s_d \boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}/d) - \tau s_d \Pi_{\boldsymbol{D}_\sigma \boldsymbol{a}} \overline{\boldsymbol{W}}\,\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a} - \tau s_d \Pi_{\boldsymbol{D}_\sigma \boldsymbol{a}}^\perp \overline{\boldsymbol{W}}_c \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a} \\ &= \boldsymbol{g}(1 - \tau s_d \boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}/d) - \tau s_d \cdot \frac{\|\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2 - \langle \overline{\boldsymbol{W}}_c^\top \boldsymbol{D}_\sigma \boldsymbol{a}, \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\rangle}{\|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2} \boldsymbol{D}_\sigma \boldsymbol{a} \\ &\quad - \frac{s_d}{\sqrt{d}} \|\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2 \boldsymbol{u}, \end{aligned} \tag{3.2}$$

where $\overline{\boldsymbol{W}}, \overline{\boldsymbol{W}}_c \in \mathbb{R}^{m\times(d-1)}$ are matrices which have i.i.d. Gaussian entries with mean zero and variance $1/d$. Furthermore, $\overline{\boldsymbol{W}}$ is independent of $(\boldsymbol{g}, \boldsymbol{a})$ and $\overline{\boldsymbol{W}}_c$ is independent of $(\boldsymbol{g}, \boldsymbol{a}, \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a})$. Such independence is established by Lemma 3.1. In the last line above, $\boldsymbol{u} = \sqrt{d}\tau \overline{\boldsymbol{W}}_c \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a} \|\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^{-1} \in \mathbb{R}^m$, which by the property of the Gaussian distribution and the independence result we have just established has i.i.d. standard Gaussian entries and is independent of $(\boldsymbol{g}, \boldsymbol{a}, \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a})$.

We introduce the following notations for the sake of simplicity.

$$\begin{aligned} \mu &:= \frac{1}{d}\tau s_d \boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}, \\ \beta &:= \tau s_d \cdot \frac{\|\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2 - \langle \overline{\boldsymbol{W}}_c^\top \boldsymbol{D}_\sigma \boldsymbol{a}, \overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\rangle}{\sqrt{m}\|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2}, \\ \gamma &:= \frac{s_d}{\sqrt{d}} \|\overline{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2. \end{aligned} \tag{3.3}$$

The following lemma states that under the current setting, the above quantities are small in probability.

Lemma 3.3. *Under the conditions of Theorem 2.1, if we further assume that there exists a constant $S_0 > 0$, such that $s_d \to S_0$ as $d \to \infty$, then as d goes to infinity we have*

$$\mu = o_P(1), \quad \beta = o_P(1), \quad \gamma = o_P(1).$$

We postpone the proof of Lemma 3.3 to Section A.3, which constitutes merely of standard applications of concentration inequalities. In the following parts of the proof, we will assume $\{s_d\}_{d\in\mathbb{N}_+}$ satisfies the condition stated in Lemma 3.3.

Let $F:\mathbb{R}^m\to\mathbb{R}$ be a random function, such that $F(\boldsymbol{y})=\sum_{i=1}^m a_i\sigma(y_i)$. Then the quantity of interest $f(\boldsymbol{x}^s)-f(\boldsymbol{x})$ can be expressed as $F(\boldsymbol{g}^s)-F(\boldsymbol{g})$. Furthermore, by equations (3.2) and (3.3), we have

$$\begin{aligned}
&F(\boldsymbol{g}^s)-F(\boldsymbol{g})-\langle\nabla F(\boldsymbol{g}),\boldsymbol{g}^s-\boldsymbol{g}\rangle\\
&\quad=\sum_{i=1}^m\big\{a_i\big(\sigma((1-\mu)g_i-\beta\sqrt{m}a_i\sigma'(g_i)-\gamma u_i)-\sigma(g_i)\big)+\mu a_i\sigma'(g_i)g_i\\
&\qquad+\beta\sqrt{m}a_i^2\sigma'(g_i)^2+\gamma a_i u_i\sigma'(g_i)\big\}.
\end{aligned}$$

Then we proceed to show that $F(\boldsymbol{g}^s)$ can be well approximated by the corresponding first order Taylor expansion at $\boldsymbol{g}$. Namely, we will show that

$$|F(\boldsymbol{g}^s)-F(\boldsymbol{g})-\langle\nabla F(\boldsymbol{g}),\boldsymbol{g}^s-\boldsymbol{g}\rangle|=o_P(1).$$

Let $b_i=\sqrt{m}a_i$, then $b_i\overset{\text{i.i.d.}}{\sim}\mathsf{N}(0,1)$ for $i\in[m]$. For $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)\in\mathbb{R}^3$, we define

$$h_{\boldsymbol{\theta}}(b,g,u):=b\sigma((1-\theta_1)g-\theta_2 b\sigma'(g)-\theta_3 u)-b\sigma(g).$$

Notice that

$$\begin{aligned}
F(\boldsymbol{g}_s)&=\frac{1}{\sqrt{m}}\sum_{i=1}^m h_{(\mu,\beta,\gamma)}(b_i,g_i,u_i),\\
F(\boldsymbol{g})&=\frac{1}{\sqrt{m}}\sum_{i=1}^m h_{(0,0,0)}(b_i,g_i,u_i).
\end{aligned}$$

Given these expressions, it is a natural reflex to apply the central limit theorem to study $F(\boldsymbol{g}_s)-F(\boldsymbol{g})$. However, the fact that (μ,β,γ) are random and depend on $(\boldsymbol{a},\boldsymbol{g},\boldsymbol{u})$ raises doubts about such application. To fix this issue, we resort to the uniform central limit theorem to present a valid result.

For $\boldsymbol{\theta}\in\mathbb{R}^3$, we define the empirical process $\mathbb{G}_m$ evaluated at $\boldsymbol{\theta}$ as

$$\mathbb{G}_m(\boldsymbol{\theta}):=\frac{1}{\sqrt{m}}\sum_{i=1}^m\big(h_{\boldsymbol{\theta}}(b_i,g_i,u_i)-\mathbb{E}[h_{\boldsymbol{\theta}}(b_i,g_i,u_i)]\big),\tag{3.4}$$

where the expectation is taken over $\{(b_i,g_i,u_i)\}_{i\le m}\overset{\text{i.i.d.}}{\sim}\mathsf{N}(\boldsymbol{0},\boldsymbol{I}_3)$. For $\boldsymbol{\theta},\overline{\boldsymbol{\theta}}\in\mathbb{R}^3$, we define the covariance function $c:\mathbb{R}^3\times\mathbb{R}^3\to\mathbb{R}$ as

$$c(\boldsymbol{\theta},\overline{\boldsymbol{\theta}}):=\mathbb{E}[h_{\boldsymbol{\theta}}(b,g,u)h_{\overline{\boldsymbol{\theta}}}(b,g,u)]-\mathbb{E}[h_{\boldsymbol{\theta}}(b,g,u)]\mathbb{E}[h_{\overline{\boldsymbol{\theta}}}(b,g,u)].\tag{3.5}$$

Application of the regularity assumptions imposed on σ and the dominated convergence theorem evidently reveals the continuity of $c(\cdot,\cdot)$. We denote by $\mathbb{G}$ the Gaussian process indexed by $\boldsymbol{\theta}$ with mean zero and covariance function $c(\cdot,\cdot)$. The following lemma establishes that $\mathbb{G}_m$ converges weakly to $\mathbb{G}$.

Lemma 3.4. *Let $\Omega := \{\boldsymbol{x} \in \mathbb{R}^3 : \|\boldsymbol{x}\|_\infty \le 1\}$, and $C(\Omega)$ be the space of continuous functions on Ω endowed with the supremum norm. Under the conditions of Theorem 2.1, if we further assume that there exists a constant $S_0 > 0$, such that $s_d \to S_0$ as $d \to \infty$, then $\{\mathbb{G}_m\}_{m\ge 1}$ converges weakly in $C(\Omega)$ to $\mathbb{G}$, which is a Gaussian process with mean zero and covariance defined in equation (3.5).*

The proof of Lemma 3.4 is deferred to Section A.4.

Remark 3.1. Lemma 3.4 is the main step in which we lose quantitative control of success probability for the FGSM attack. We prove this lemma by an application of the uniform central limit theorem. A more explicit approach should be able to provide concrete probability bounds.

For $\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \in \Omega$, we define $\rho(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) := \mathbb{E}[(\mathbb{G}(\boldsymbol{\theta}) - \mathbb{G}(\overline{\boldsymbol{\theta}}))^2]^{1/2}$. Then by [25, Lemma 18.15], without any loss, we can and will assume that $\mathbb{G}$ almost surely has ρ-continuous sample path.

In the following parts, we fix some positive ϵ a priori, and define

$$S_\epsilon(\mathbb{G}) := \sup_{\|\boldsymbol{\theta}\|_\infty \le \epsilon} |\mathbb{G}(\boldsymbol{\theta})|.$$

Note that S_ϵ is a continuous function with respect to the supremum norm on Ω, thus $S_\epsilon(\mathbb{G}_m)$ converges weakly to $S_\epsilon(\mathbb{G})$ due to the continuous mapping theorem. Recall that we showed in Lemma 3.3 that $\mu, \beta, \gamma \xrightarrow{\mathbb{P}} 0$ as $d \to \infty$, thus for any $\epsilon > 0$,

$$\begin{aligned}
&\left|F(\boldsymbol{g}^s) - F(\boldsymbol{g}) + \sqrt{m}\beta\mathbb{E}[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)]\right| \\
&\quad \overset{\text{(i)}}{=} \Bigg|\frac{1}{\sqrt{m}}\sum_{i=1}^m b_i\big(\sigma((1-\mu)g_i - \beta b_i\sigma'(g_i) - \gamma u_i) - \sigma(g_i)\big) \\
&\qquad\quad - \sqrt{m}\mathbb{E}\big[b\big(\sigma((1-\mu)g - \beta b\sigma'(g) - \gamma u) - \sigma(g)\big)\big]\Bigg| \\
&\quad = \left|\frac{1}{\sqrt{m}}\sum_{i=1}^m \big(h_{(\mu,\beta,\gamma)}(b_i, g_i, u_i) - \mathbb{E}[h_{(\mu,\beta,\gamma)}(b_i, g_i, u_i)]\big)\right| \\
&\quad \overset{\text{(ii)}}{\le} S_\epsilon(\mathbb{G}_m) + \delta_\epsilon(m),
\end{aligned}$$

where $\delta_\epsilon(m) \xrightarrow{\mathbb{P}} 0$ as $d \to \infty$. In the above display, (i) is by Stein's lemma (see Lemma 3.5 below), (ii) is by Lemma 3.3, and the expectations are taken over

$$\{b, b_i, g, g_i, u, u_i\}_{i\in[m]} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1).$$

Lemma 3.5 (Stein's lemma). *Suppose Z is a normally distributed random variable with expectation x_1 and variance x_2. Further suppose g is a function for which the two expectations $\mathbb{E}[g(Z)(Z - x_1)]$ and $\mathbb{E}[g'(Z)]$ both exist. Then*

$$\mathbb{E}[g(Z)(Z - x_1)] = x_2 \mathbb{E}[g'(Z)].$$

We next prove that $S_\epsilon(\mathbb{G}_m)$ is small, which consists of two major steps. In the first step, we show that $S_\epsilon(\mathbb{G})$ is small, then we establish that $S_\epsilon(\mathbb{G})$ and $S_\epsilon(\mathbb{G}_m)$ are close for large d.

Note that $c(\mathbf{0}, \mathbf{0}) = 0$. Since $\mathbb{G}$ has ρ-continuous sample path and the covariance function c is continuous, we then obtain that $S_\epsilon(\mathbb{G}) \xrightarrow{\mathbb{P}} 0$ as $\epsilon \to 0^+$. For all $\epsilon' > 0$, we first choose $\epsilon > 0$ small enough, such that $\mathbb{P}(S_\epsilon(\mathbb{G}) \geq \epsilon'/3) \leq \epsilon'/3$. Since

$$S_\epsilon(\mathbb{G}_m) \xrightarrow{d} S_\epsilon(\mathbb{G}),$$

$S_\epsilon(\mathbb{G})$ obviously has continuous cumulative distribution function, and $\delta_\epsilon(m) \xrightarrow{\mathbb{P}} 0$ as $m, d \to \infty$, putting these together we conclude that there exists $m_{\epsilon,\epsilon'} \in \mathbb{N}_+$, such that for all $m \geq m_{\epsilon,\epsilon'}$,

$$\mathbb{P}(|\delta_\epsilon(m)| \geq \epsilon'/3) \leq \epsilon'/3 \quad \text{and} \quad \mathbb{P}(|S_\epsilon(\mathbb{G}_m)| \geq \epsilon'/3) \leq \epsilon'/3 + \mathbb{P}(|S_\epsilon(\mathbb{G})| \geq \epsilon'/3).$$

In summary, for all $m \geq m_{\epsilon,\epsilon'}$,

$$\mathbb{P}\left(\left|F(\boldsymbol{g}^s) - F(\boldsymbol{g}) + \sqrt{m}\beta\mathbb{E}[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)]\right| \geq \epsilon'\right) \leq \epsilon'.$$

As the choice of ϵ', we then have

$$\left|F(\boldsymbol{g}^s) - F(\boldsymbol{g}) + \sqrt{m}\beta\mathbb{E}[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)]\right| = o_P(1). \quad (3.6)$$

Note that in the above equation the expectation is taken over $(b, g, u) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_3)$, and $\mathbb{E}[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)]$ is a random variable which depends on the values of the random vector (μ, β, γ).

Next, we consider $\langle \nabla F(\boldsymbol{g}), \boldsymbol{g} - \boldsymbol{g}^s \rangle$. Notice that this formula admits the following decomposition:

$$\begin{aligned}\langle \nabla F(\boldsymbol{g}), \boldsymbol{g} - \boldsymbol{g}^s \rangle &= \mu \sum_{i=1}^m a_i \sigma'(g_i) g_i + \beta\sqrt{m} \sum_{i=1}^m a_i^2 \sigma'(g_i)^2 + \gamma \sum_{i=1}^m a_i u_i \sigma'(g_i) \\ &= \mu T_1 + \beta\sqrt{m} T_2 + \gamma T_3.\end{aligned}$$

Since $\{\sqrt{m}a_i, u_i, g_i\}_{i\in[m]}$ are i.i.d. standard Gaussian random variables, as an immediate consequence of the law of large numbers and the central limit theorem, we can conclude that T_1 and T_3 are both $O_P(1)$, and $T_2 = \mathbb{E}[\sigma'(g)^2] + O_P(m^{-1/2})$. Using this result and Lemma 3.3, we further deduce that

$$\langle \nabla F(\boldsymbol{g}), \boldsymbol{g} - \boldsymbol{g}^s \rangle = \beta\sqrt{m}\,\mathbb{E}[\sigma'(g)^2] + o_P(1). \tag{3.7}$$

The law of large numbers implies that as $d \to \infty$, we have $\beta\sqrt{m} = \tau S_0 + o_P(1)$. As a result,

$$\langle \nabla F(\boldsymbol{g}), \boldsymbol{g} - \boldsymbol{g}^s \rangle = \tau S_0 \mathbb{E}[\sigma'(g)^2] + o_P(1).$$

Recall that according to Lemma 3.3, $\mu, \beta, \gamma = o_P(1)$, thus intuitively we would expect that the expectations displayed in equation (3.6) and equation (3.7) are close to each other. To make such heuristic rigorous, we notice that by assumption, σ' is almost everywhere continuous and $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. Then we apply the dominated convergence theorem, and obtain that as $d \to \infty$,

$$\mathbb{E}[\sigma'(g)\sigma'((1-\mu)g - \beta b \sigma'(g) - \gamma u)] = \mathbb{E}[\sigma'(g)^2] + o_P(1). \tag{3.8}$$

Substituting equations (3.7) and (3.8) into equation (3.6) gives

$$|F(\boldsymbol{g}^s) - F(\boldsymbol{g}) - \langle \nabla F(\boldsymbol{g}), \boldsymbol{g}^s - \boldsymbol{g} \rangle| = o_P(1),$$

which further leads to

$$F(\boldsymbol{g}^s) = F(\boldsymbol{g}) - \tau S_0 \mathbb{E}[\sigma'(g)^2] + o_P(1).$$

Furthermore, the central limit theorem implies that $F(\boldsymbol{g}) \xrightarrow{d} \mathsf{N}(0, \mathbb{E}[\sigma(g)^2])$, thus as $d \to \infty$,

$$\mathbb{P}\big(\mathrm{sign}(F(\boldsymbol{g})) \ne \mathrm{sign}(F(\boldsymbol{g}^s))\big) \to \mathbb{P}\big(\mathrm{sign}(z) \ne \mathrm{sign}\big(z - \mathrm{sign}(z) S_0 \mathbb{E}[\sigma'(g)^2]\big)\big),$$

where $z \sim \mathsf{N}(0, \mathbb{E}[\sigma(g)^2])$, $g \sim \mathsf{N}(0,1)$, and are independent of each other. Since S_0 is arbitrary, using a standard diagonal argument, we derive that there exists a sequence of step sizes $\{s_d\}_{d\in\mathbb{N}_+}$, such that as $d \to \infty$,

$$\mathbb{P}\big(\mathrm{sign}(F(\boldsymbol{g})) \ne \mathrm{sign}(F(\boldsymbol{g}^s))\big) \to 1 \quad \text{and} \quad \|\boldsymbol{x}^s - \boldsymbol{x}\|_2 / \|\boldsymbol{x}\|_2 \xrightarrow{\mathbb{P}} 0.$$

More precisely, for all $n \in \mathbb{N}_+$, there exists $S_0^n > 0$ and $d_n \in \mathbb{N}_+$, such that if we set $s_d = S_0^n$ for all $d \in \mathbb{N}_+$, then for all $d \ge d_n$,

$$\mathbb{P}\big(\mathrm{sign}(F(\boldsymbol{g})) \ne \mathrm{sign}(F(\boldsymbol{g}^s))\big) \ge 1 - \frac{1}{n}.$$

Without loss of generality, we can assume $d_n < d_{n+1}$, $S_0^n \le \xi_{d_n}$, and $S_0^n / \sqrt{d}_n < n^{-1}$ by fixing S_0^n and taking d_n large enough. We set $s_d = S_0^n$ if and only if $d_n \le d < d_{n+1}$.

Under such choice of $\{s_d\}_{d\geq 1}$, for all $d_{n+1} > d \geq d_n$, we have

$$\frac{s_d}{\sqrt{d}} = \frac{S_0^n}{\sqrt{d}} \leq \frac{S_0^n}{\sqrt{d_n}} \leq \frac{1}{n}, \quad \mathbb{P}\big(\mathrm{sign}(F(\boldsymbol{g})) \neq \mathrm{sign}(F(\boldsymbol{g}^s))\big) \geq 1 - \frac{1}{n},$$
$$s_d \leq \xi_{d_n} \leq \xi_d.$$

Since n is arbitrary, then we combine the equations above with the first claim of the theorem and conclude the proof of the second claim.

4. Non-asymptotic result for two-layer networks

As emphasized above, Theorem 2.1 does not provide a quantitative convergence rate for the probability that the attack succeeds, namely, $\mathbb{P}(\mathrm{sign}(f(\boldsymbol{x})) \neq \mathrm{sign}(f(\boldsymbol{x}^s)))$. We remedy to this by establishing a non-asymptotic bound below, at the price of assuming that the activation function σ is Lipschitz continuous.

Theorem 4.1. *Consider the random two-layer neural network of equation* (2.1)*. Let* $\boldsymbol{x} \in \mathbb{R}^d$ *be a deterministic vector with* $\|\boldsymbol{x}\|_2 = \sqrt{d}$*. Assume that* σ *is* L*-Lipschitz over* $\mathbb{R}$ *for some* $L \geq 1$*,* σ *is not a constant, and* σ' *is almost everywhere continuous. Then there exist numerical constants* $c, C_0 > 0$*, such that for all* $\xi > 0$*, if the following conditions hold:*

$$d \geq \max\left\{\frac{C_\xi^2 C_0(\sigma(0)^2 + L^2)}{\xi}, \frac{4L^4 C_\xi^2}{c^2}\left(\log\frac{C_0}{\xi}\right)^2, 16C_\xi^4\big(1 + \mathbb{E}_{g\sim\mathsf{N}(0,1)}[\sigma(g)^2]\big)^2\right\},$$
$$m \geq C_\xi^4, \quad \widetilde{Q}_{d,m} \geq \mathbb{E}_{g\sim\mathsf{N}(0,1)}[\sigma'(g)^2]/2, \quad \xi \leq C_0 e^{-9c}, \quad s_d = C_\xi, \tag{4.1}$$

where

$$C_\xi = \frac{4\sqrt{\log(C_0/\xi)\cdot(\mathbb{E}_{g\sim\mathsf{N}(0,1)}[\sigma(g)^2] + 1)}}{\sqrt{c}\,\mathbb{E}_{g\sim\mathsf{N}(0,1)}[\sigma'(g)^2]},$$
$$\widetilde{Q}_{d,m} = \min_{\substack{|\theta|_1\leq d^{-1/2},\\ |\theta_2|\leq 2m^{-1/4}, |\theta_3|\leq d^{-1/4}}} \mathbb{E}\big[\sigma'(g)\sigma'((1-\theta_1)g - \theta_2 b\sigma'(g) - \theta_3 u)\big],$$
$$g, b, u \sim_{\text{i.i.d.}} \mathsf{N}(0,1).$$

Then with probability at least

$$1 - 3\xi - C_0\big(\exp(-cd) + m^{-1}(\sigma(0)^4 + L^4) + 2L(d^{-1/4} + m^{-1/4})\big),$$

it holds that $\mathrm{sign}(f(\boldsymbol{x})) \neq \mathrm{sign}(f(\boldsymbol{x}^s))$.

Remark 4.1. In order to parse the last statement, think of ξ as a small constant controlling the probability that the adversarial attack fails. Then the conditions (4.1) require d and m to be larger than some constants $d_0(\xi)$ and $m_0(\xi)$, where $d_0(\xi)$ is of order $\log(1/\xi)/\xi$, and $m_0(\xi)$ is only polylogarithmic in $1/\xi$.

In order to prove Theorem 4.1, it suffices to prove the following lemma, which can be regarded as a more general version of the statement.

Lemma 4.1. *Under the conditions of Theorem 4.1, for $\eta_1, \eta_2 > 0$, we define*

$$\begin{aligned} \mathcal{S}_{d,m} &:= \Big\{ \boldsymbol{\theta} : |\theta_1| \le \frac{s_d \eta_1}{d}, \Big|\theta_2 - \frac{\tau s_d}{\sqrt{m}}\Big| \le \frac{2 s_d \eta_2}{\sqrt{dm}}, \\ &\qquad |\theta_3| \le \frac{2 s_d}{\sqrt{d}} \cdot \sqrt{1 + \mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2]} \Big\}, \\ Q_{d,m} &:= \min_{\boldsymbol{\theta}\in \mathcal{S}_{d,m}} \mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)} \big[\sigma'(g)\sigma'((1-\theta_1)g - \theta_2 b \sigma'(g) - \theta_3 u)\big], \\ \delta_{d,m} &:= \max \Big\{ \frac{s_d \eta_1}{d}, \frac{2 s_d \eta_2}{\sqrt{dm}} + \frac{\tau s_d}{\sqrt{m}}, \frac{2 s_d}{\sqrt{d}} \cdot \sqrt{1 + \mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2]} \Big\}. \end{aligned} \tag{4.2}$$

In the above definitions, we ignore the dependence on (η_1, η_2) for the sake of simplicity. Then there exist numerical constants $c, C_0 > 0$, such that with probability at least

$$\begin{aligned} 1 - C_0 \big\{ &\eta_1^{-2}(\sigma(0)^2 + L^2) + \exp(-c\eta_2) + \exp(-cd) \\ &+ m^{-1}(\sigma(0)^4 + L^4) + \exp(-c\eta_3^2) + \eta^{-1} L \delta_{d,m} \big\}, \end{aligned}$$

we have $\mathrm{sign}(f(\boldsymbol{x})) \ne \mathrm{sign}(f(\boldsymbol{x}^s))$. *In the above display,*

$$\eta_3 = \frac{s_d Q_{d,m} - \eta - 2d^{-1/2} s_d L^2 \eta_2 - 1}{\sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1}}.$$

Remark 4.2. In Lemma 4.1, if we set

$$\begin{aligned} \eta = 1, \quad \eta_1 &= \sqrt{\frac{C_0(\sigma(0)^2 + L^2)}{\xi}}, \quad \eta_2 = c^{-1} \log \frac{C_0}{\xi}, \\ s_d &= \frac{2\sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1} \cdot \big(\sqrt{c^{-1}\log(C_0/\xi)} + 1\big) + 2}{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] - 4d^{-1/2}\eta_2}, \end{aligned}$$

then Theorem 4.1 reduces to a direct corollary of Lemma 4.1.

We will prove Lemma 4.1 in the rest of the parts of this section. Recall that μ, β, γ were defined in equation (3.3). We first give a non-asymptotic characterization of these random quantities.

Lemma 4.2. *There exist numerical constants $c, C > 0$, such that the following results hold:*

(1) *For any $\eta_1 > 0$, with probability at least $1 - C\eta_1^{-2}(\sigma(0)^2 + L^2)$,*

$$|\mu| \le \frac{s_d \eta_1}{d}.$$

(2) *For any $\eta_2 \ge 1$, with probability at least $1 - C\exp(-c\eta_2)$,*

$$\left|\beta - \frac{\tau s_d}{\sqrt{m}}\right| \le \frac{2 s_d \eta_2}{\sqrt{dm}}.$$

(3) *With probability at least $1 - 2\exp(-cd) - Cm^{-1}(\sigma(0)^4 + L^4)$,*

$$|\gamma| \le \frac{2 s_d}{\sqrt{d}} \cdot \sqrt{1 + \mathbb{E}_{g\sim \mathrm{N}(0,1)}[\sigma(g)^2]}.$$

We postpone the proof of Lemma 4.2 to Section B.1. In what follow, we will always assume that the events described in Lemma 4.2 occur. Namely, we will be working on event $\mathcal{S}$ defined as follows:

$$\mathcal{S} = \left\{|\mu| \le \frac{s_d \eta_1}{d}, \left|\beta - \frac{\tau s_d}{\sqrt{m}}\right| \le \frac{2 s_d \eta_2}{\sqrt{dm}}, |\gamma| \le \frac{2 s_d}{\sqrt{d}} \cdot \sqrt{1 + \mathbb{E}_{g\sim \mathrm{N}(0,1)}[\sigma(g)^2]}\right\}.$$

By Lemma 4.2,

$$\begin{aligned}\mathbb{P}(\mathcal{S}) \ge 1 - C\eta_1^{-2}(\sigma(0)^2 + L^2) - C\exp(-c\eta_2) \\ - 2\exp(-cd) - Cm^{-1}(\sigma(0)^4 + L^4).\end{aligned}$$

Recall from equations (3.2) and (3.4) that

$$\begin{aligned}&F(\boldsymbol{g}^s) - F(\boldsymbol{g}) = \sum_{i=1}^m a_i\big(\sigma((1-\mu)g_i - \beta\sqrt{m}a_i\sigma'(g_i) - \gamma u_i) - \sigma(g_i)\big) \\ &= \mathbb{G}_m((\mu,\beta,\gamma)) - \sqrt{m}\beta\,\mathbb{E}_{g,b,u\sim_{\text{i.i.d.}}\mathrm{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big].\end{aligned}$$

We then show that $\mathbb{G}_m((\mu,\beta,\gamma))$ is close to zero. More precisely, for $\delta > 0$, we define the set $\Theta_\delta := \{\boldsymbol{\theta} \in \mathbb{R}^3 : \|\boldsymbol{\theta}\|_\infty \le \delta\}$. We immediately see that if $\max\{|\mu|, |\beta|, |\gamma|\} \le \delta$, then

$$\mathbb{G}_m((\mu,\beta,\gamma)) \le \sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{G}_m(\boldsymbol{\theta}).$$

For $\boldsymbol{\theta} \in \mathbb{R}^3$, we define

$$\mathbb{L}_m(\boldsymbol{\theta}) = \frac{1}{\sqrt{m}}\sum_{i=1}^m \varepsilon_i b_i\big(\sigma((1-\theta_1)g_i - \theta_2 b_i\sigma'(g_i) - \theta_3 u_i) - \sigma(g_i)\big),$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1, +1\}$ and are independent of everything else. By symmetrization, it holds that

$$\mathbb{E}\big[\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{G}_m(\boldsymbol{\theta})\big] \le 2\,\mathbb{E}\big[\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{L}_m(\boldsymbol{\theta})\big]. \tag{4.3}$$

Since $\mathbb{L}_m(\vec{0}) = \mathbb{G}_m(\vec{0}) = 0$, we see that both $\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{G}_m(\boldsymbol{\theta})$ and $\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{L}_m(\boldsymbol{\theta})$ are non-negative. Furthermore, conditioning on $\{(b_i, g_i, u_i) : i \in [m]\}$, it is not hard to see that $\{\mathbb{L}_m(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_\delta\}$ is a sub-Gaussian process indexed by parameters in Θ_δ. In addition, the sub-Gaussian norm of this process $\|\cdot\|_{\Psi_2}$ satisfies the following inequality.

Lemma 4.3. *There exists a numerical constant $C > 0$, such that for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^3$,*

$$\|\mathbb{L}_m(\boldsymbol{\theta}) - \mathbb{L}_m(\boldsymbol{\theta}')\|_{\Psi_2} \le C \cdot \sqrt{\frac{1}{m}\sum_{i=1}^m M(b_i, g_i, u_i)^2} \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,$$

where

$$M(b, g, u) = L \cdot |b| \cdot \sqrt{g^2 + L^2 b^2 + u^2}.$$

We defer the proof of the lemma to Section B.2. Using the Dudley's integral inequality, we conclude that there exists another numerical constant $C' > 0$, such that

$$\mathbb{E}_{\varepsilon_i \sim_{\text{i.i.d.}} \text{Unif}\{\pm 1\}}\Bigg[\sup_{\boldsymbol{\theta}\in\Theta_\delta} \frac{1}{\sqrt{m}}\sum_{i=1}^m \varepsilon_i b_i \big(\sigma((1-\theta_1)g_i - \theta_2 b_i - \theta_3 u_i) - \sigma(g_i)\big)\Bigg]$$
$$\le C' \sqrt{\frac{1}{m}\sum_{i=1}^m M(b_i, g_i, u_i)^2} \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta_\delta, \|\cdot\|_2, x)}\, \mathrm{d}x, \tag{4.4}$$

where $\mathcal{N}(\Theta, \|\cdot\|_2, x)$ is the smallest number of closed balls with centers in Θ and radius x whose union covers Θ. In our case, since Θ_δ is contained in the ball centered at the origin with radius $\sqrt{3}\delta$, we have $\mathcal{N}(\Theta_\delta, \|\cdot\|_2, x) \le (1 + 4\delta/x)^3$. Plugging this upper bound into equation (4.4) further leads to the following result:

$$\mathbb{E}\big[\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{L}_m(\boldsymbol{\theta})\big] \le C'' L \int_0^{\sqrt{3}\delta} \sqrt{\log(1 + 4\delta/x)}\, \mathrm{d}x \le 4C'' L\delta, \tag{4.5}$$

where $C'' > 0$ is a numerical constant. Combining (4.3) and (4.5), we can further upper bound the expectation of the non-negative random variable $\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{G}_m(\boldsymbol{\theta})$ with $8C''L\delta$. By Markov's inequality, with probability at least $1 - 8\eta^{-1}C''L\delta$,

$$\sup_{\boldsymbol{\theta}\in\Theta_\delta} \mathbb{G}_m(\boldsymbol{\theta}) \le \eta$$

for any $\eta > 0$. Recall that $\delta_{d,m}$ is defined in equation (4.2). By Lemma 4.2, we see that with probability at least

$$\begin{aligned} &1 - C\eta_1^{-2}(\sigma(0)^2 + L^2) - C\exp(-c\eta_2) \\ &\quad - 2\exp(-cd) - Cm^{-1}(\sigma(0)^4 + L^4) - 8\eta^{-1}C''L\delta_{d,m}, \end{aligned}$$

we have

$$\begin{aligned} &F(\boldsymbol{g}^s) - F(\boldsymbol{g}) \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta_{\delta_{d,m}}} \mathbb{G}_m(\boldsymbol{\theta}) - \sqrt{m}\beta\, \mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big] \\ &\leq \eta - \sqrt{m}\beta\, \mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big]. \end{aligned}$$

Analogously, similar lower bound holds with at least the same amount of probability:

$$\begin{aligned} &F(\boldsymbol{g}^s) - F(\boldsymbol{g}) \\ &\qquad \geq -\eta - \sqrt{m}\beta\, \mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big]. \end{aligned}$$

Furthermore, on the set $\mathcal{S}$, it holds that

$$|\sqrt{m}\beta - \tau s_d| \leq \frac{2 s_d \eta_2}{\sqrt{d}}.$$

As a result, with probability at least

$$\begin{aligned} &1 - 2C\eta_1^{-2}(\sigma(0)^2 + L^2) - 2C\exp(-c\eta_2) \\ &\quad - 4\exp(-cd) - 2Cm^{-1}(\sigma(0)^4 + L^4) - 16\eta^{-1}C''L\delta_{d,m}, \end{aligned}$$

we have

$$\begin{aligned} &\big|F(\boldsymbol{g}^s) - F(\boldsymbol{g}) + \tau s_d \mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big]\big| \\ &\qquad \leq \eta + \frac{2 s_d L^2 \eta_2}{\sqrt{d}}. \end{aligned}$$

Since μ, β, γ are all $o_P(1)$, we expect that

$$\mathbb{E}_{g,b,u \sim_{\text{i.i.d.}} \mathsf{N}(0,1)}\big[\sigma'(g)\sigma'((1-\mu)g - \beta b\sigma'(g) - \gamma u)\big]$$

should be approximately equal to $\mathbb{E}_{g \sim \mathsf{N}(0,1)}[\sigma'(g)^2]$, which is strictly positive and does not depend on (m, d), provided that σ is not a constant function. In this case, we only need to choose the step size s_d large enough to flip the sign of $F(\boldsymbol{g})$. This argument can be made rigorous via the following lemma. In this lemma, we upper bound the magnitude of $F(\boldsymbol{g})$.

Lemma 4.4. *For any $\eta_3 \geq 0$, with probability at least*

$$1 - C\exp\Big(-\frac{cm}{\sigma(0)^4 + L^4}\Big) - C\exp(-c\eta_3^2)$$

for some numerical constants $c, C > 0$, we have

$$|F(\boldsymbol{g})| \leq \eta_3 \cdot \sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1}.$$

We prove the lemma in Section B.3. Let

$$\mathcal{S}' := \mathcal{S} \cap \big\{|F(\boldsymbol{g})| \leq \eta_3 \cdot \sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1}\big\}.$$

On $\mathcal{S}'$, if in addition, we have

$$s_d Q_{d,m} - \eta - \frac{2 s_d L^2 \eta_2}{\sqrt{d}} \geq \eta_3 \cdot \sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1},$$

then $\operatorname{sign}(F(\boldsymbol{g})) \neq \operatorname{sign}(F(\boldsymbol{g}^s))$. In the rest parts of the proof, we will always take

$$\eta_3 = \frac{s_d Q_{d,m} - \eta - 2d^{-1/2} s_d L^2 \eta_2 - 1}{\sqrt{\mathbb{E}_{g\sim \mathsf{N}(0,1)}[\sigma(g)^2] + 1}}.$$

With such choice of η_3, we can finally put together all above analysis and conclude that the adversarial example succeeds with probability at least

$$1 - C_0\big\{\eta_1^{-2}(\sigma(0)^2 + L^2) + \exp(-c\eta_2) + \exp(-cd) \\ + m^{-1}(\sigma(0)^4 + L^4) + \exp(-c\eta_3^2) + \eta^{-1} L\delta_{d,m}\big\}$$

for some absolute positive constants c, C_0.

5. Random multi-layer networks

We generalize the model considered in Section 2 in the current section. More precisely, we consider a multi-layer neural network with $l + 1$ layers for $l \in \mathbb{N}_+$:

$$f(\boldsymbol{x}) = \boldsymbol{W}_{l+1}\sigma\big(\boldsymbol{W}_l\sigma(\cdots\sigma(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))\cdots)\big).$$

In the above equation, the random weight matrix $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ has i.i.d. Gaussian entries: $(\boldsymbol{W}_i)_{jj'} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/d_{i-1})$ for all $j \in [d_i]$, $j' \in [d_{i-1}]$, and further $\{\boldsymbol{W}_i\}_{i\in[l+1]}$ are independent of each other. We assume $d_0 = d$, $d_{l+1} = 1$, and $d_i = d_i(d) \to \infty$ for all $0 \leq i \leq l$. The d-dimensional input vector $\boldsymbol{x}$ is a deterministic vector with Euclidean norm $\sqrt{d}$. The activation function $\sigma: \mathbb{R} \to \mathbb{R}$ is understood to act on vectors entrywise.

For the simplicity of notations, we define recursively the following vectors:

$$\boldsymbol{h}_0 := \boldsymbol{x}, \quad \boldsymbol{g}_1 := \boldsymbol{W}_1\boldsymbol{x}, \quad \boldsymbol{h}_j := \sigma(\boldsymbol{g}_j), \quad \text{and} \quad \boldsymbol{g}_{j+1} := \boldsymbol{W}_{j+1}\boldsymbol{h}_j$$

for $j \in [l]$. The gradient of f can be expressed as

$$\nabla f(\boldsymbol{x}) = \boldsymbol{W}_1^\intercal \boldsymbol{D}_\sigma^1 \boldsymbol{W}_2^\intercal \boldsymbol{D}_\sigma^2 \cdots \boldsymbol{W}_l^\intercal \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\intercal,$$

where $\boldsymbol{D}_\sigma^j = \operatorname{diag}(\{\sigma'(\boldsymbol{g}_j)\}) \in \mathbb{R}^{d_j \times d_j}$. As before, we denote by $\tau \in \{\pm 1\}$ the sign of $f(\boldsymbol{x})$, and let $\{s_d\}_{d\in\mathbb{N}_+} \subseteq \mathbb{R}^+$ be a sequence of step sizes to be determined.

Theorem 5.1. *Assume that σ satisfies the conditions in Theorem 2.1. Then the following results hold:*

(1) *There exists a constant $C > 0$ depending uniquely on (σ, l), such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\frac{\|\boldsymbol{x}-\boldsymbol{x}^s\|_2}{\|\boldsymbol{x}\|_2} \le \frac{C s_d}{\sqrt{d}} \big(\sqrt{\log(1/\delta)}+1\big)^{l-1} \big(1+\log(1/\delta) d^{-1/2}\big) \times \prod_{i=1}^{l}\prod_{j=1}^{i} \big(1+\delta^{-1/2} d_j^{-1/2}\big)^{k^{i-j}}, \tag{5.1}$$

where we recall that k is a fixed positive integer such that

$$|\sigma'(x)| \le C_\sigma(1+|x|^{k-1}).$$

(2) *Let $\{\xi_d\}_{d\in\mathbb{N}_+} \subseteq \mathbb{R}^+$ be an increasing sequence such that $\xi_d \to \infty$ as $d \to \infty$. Then there exists $\{s_d\}_{d\in\mathbb{N}_+} \subseteq \mathbb{R}^+$, such that $s_d \le \xi_d$ and the following limits hold:*

$$\text{p-lim}_{d\to\infty} \frac{\|\boldsymbol{x}-\boldsymbol{x}^s\|_2}{\|\boldsymbol{x}\|_2} = 0, \quad \lim_{d\to\infty} \mathbb{P}\big(\operatorname{sign}(f(\boldsymbol{x})) \ne \operatorname{sign}(f(\boldsymbol{x}^s))\big) = 1.$$

Remark 5.1. The proof of Theorem 5.1 applies without changes to neural networks which have different activation functions at different layers, provided they satisfy the assumptions as stated. We refrain from stating such a generalization to avoid cumbersome notations.

Remark 5.2. The bound on the perturbation size in equation (5.1) deteriorates when the depth l becomes exponentially large in the input dimension d. A similar behavior is observed in [2] which also provides an example of a random network with exponential depth for which the output is nearly constant, and in particular is immune to FGSM attacks.

For general random networks, the example of [2] implies that subexponential depth is a required assumption. On the other hand, the example of [2] is special in that the network is non-balanced: it takes the same sign for any input. As we pointed out in the introduction, if the output takes each sign on a fraction of the inputs with measure bounded away from zero (it is 'balanced'), then it must have adversarial examples by isoperimetry.

On the other hand, it is unclear whether subexponential depth is necessary for FGSM attacks to be successful on random networks, after balancing. As the depth increases, the random function $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ becomes 'rougher,' as it can be seen by computing its covariance function. While such a function will contain adversarial examples, it is likely to be more difficult to find them by a single gradient step as in FGSM. (Of course a special case is the one of linear activations: in that case depth is irrelevant.)

6. Proof of Theorem 5.1

Proof of the first claim. For $m \in [l]$, we define

$$\boldsymbol{\eta}_m := \boldsymbol{D}_\sigma^m \boldsymbol{W}_{m+1}^\top \boldsymbol{D}_\sigma^{m+1} \cdots \boldsymbol{W}_l^\top \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\top \in \mathbb{R}^{d_m} \quad \text{and} \quad \boldsymbol{y}_m := \boldsymbol{W}_m^\top \boldsymbol{\eta}_m \in \mathbb{R}^{d_{m-1}}.$$

The following lemma shows that the normalized Euclidean norms of $\boldsymbol{\eta}_m, \boldsymbol{y}_m, \boldsymbol{h}_m, \boldsymbol{g}_m$ converge in probability to some deterministic constants as $d \to \infty$. Furthermore, such constants are independent of the choice of $\{s_d\}_{d \in \mathbb{N}_+}$.

Lemma 6.1. *Under the conditions of Theorem 5.1, the following sequences of random variables converge in probability to strictly positive constants as $d \to \infty$:*

(1) $\{\|\boldsymbol{h}_m\|_2^2/d_m\}_{d \geq 1}$ *for all* $1 \leq m \leq l$.

(2) $\{\|\boldsymbol{g}_m\|_2^2/d_m\}_{d \geq 1}$ *for all* $1 \leq m \leq l$.

(3) $\{\|\boldsymbol{\eta}_m\|_2^2\}_{d \geq 1}$ *for all* $1 \leq m \leq l$.

(4) $\{\|\boldsymbol{y}_m\|_2^2\}_{d \geq 1}$ *for all* $1 \leq m \leq l$.

Furthermore,

(5) $\boldsymbol{h}_{m-1}^\top \boldsymbol{W}_m^\top \boldsymbol{\eta}_m = O_P(1)$ *for all* $1 \leq m \leq l$.

Remark 6.1. The above sequences of random variables are independent of the choice of $\{s_d\}_{d \in \mathbb{N}_+}$.

The proof of Lemma 6.1 is deferred to Section A.5. As in the two-layers case, in the next lemma we provide a finite sample upper bound on the Euclidean norm of the gradient $\nabla f(\boldsymbol{x})$.

Lemma 6.2. *Under the conditions of Theorem 5.1, there exists a constant $Q > 0$, which is a function of (σ, l) only, such that for any $\delta > 0$, with probability at least $1 - \delta$,* *we have*

$$\|\nabla f(\boldsymbol{x})\|_2 \leq Q\big(\sqrt{\log(1/\delta)} + 1\big)^{l-1}\big(1 + \log(1/\delta)d^{-1/2}\big) \times \prod_{i=1}^{l}\prod_{j=1}^{i}\big(1 + \delta^{-1/2}d_j^{-1/2}\big)^{k^{i-j}}.$$

The proof of Lemma 6.2 is deferred to Section A.6. Recall that $\|\boldsymbol{x}\|_2 = \sqrt{d}$, thus the first claim of the theorem is just a straightforward consequence of Lemma 6.2.

Proof of the second claim. Our proof of the second claim proceeds by induction. Before stating our induction hypothesis, we analyze the first layer to gain some intuition.

We define

$$\boldsymbol{g}_1^s := \boldsymbol{W}_1\boldsymbol{x}^s = \boldsymbol{g}_1 - \tau s_d \boldsymbol{W}_1\boldsymbol{W}_1^\intercal\boldsymbol{\eta}_1,$$
$$\mathscr{F}_1 := \sigma\{\boldsymbol{g}_1, \boldsymbol{\eta}_1, \{\boldsymbol{W}_i\}_{2\leq i\leq l+1}, \boldsymbol{x}\}.$$

Notice that $\boldsymbol{g}_1$, $\boldsymbol{g}_1^s$ can be regarded as the outputs of the first layer with the inputs being $\boldsymbol{x}$ and $\boldsymbol{x}^s$, respectively.

Since $\boldsymbol{W}_1$ has i.i.d. Gaussian entries, and $\mathscr{F}_1$ depends on $\boldsymbol{W}_1$ only through $\boldsymbol{g}_1 = \boldsymbol{W}_1\boldsymbol{x}$. Invoking Lemma 3.1, we can write

$$\boldsymbol{W}_1 = \boldsymbol{g}_1\boldsymbol{x}^\intercal/d + \widetilde{\boldsymbol{W}}_1\Pi_{\boldsymbol{x}}^\perp,$$

where $\widetilde{\boldsymbol{W}}_1$ has the same marginal distribution as $\boldsymbol{W}_1$ and is independent of $\mathscr{F}_1$. Then we have

$$\boldsymbol{g}_1^s = \boldsymbol{g}_1(1 - \tau s_d \boldsymbol{g}_1^\intercal\boldsymbol{\eta}_1/d) - \tau s_d \widetilde{\boldsymbol{W}}_1\Pi_{\boldsymbol{x}}^\perp\widetilde{\boldsymbol{W}}_1^\intercal\boldsymbol{\eta}_1.$$

Furthermore, using the property of Gaussian distribution, we have

$$\widetilde{\boldsymbol{W}}_1\Pi_{\boldsymbol{x}}^\perp\widetilde{\boldsymbol{W}}_1^\intercal = \bar{\boldsymbol{W}}_1\bar{\boldsymbol{W}}_1^\intercal,$$

where $\bar{\boldsymbol{W}}_1 \in \mathbb{R}^{d_1\times(d-1)}$ is a matrix that has i.i.d. Gaussian entries with mean zero and variance $1/d$ that is further independent of $\mathscr{F}_1$. By Lemma 3.1, $\bar{\boldsymbol{W}}_1$ admits the decomposition

$$\bar{\boldsymbol{W}}_1 = \Pi_{\boldsymbol{\eta}_1}^\perp\boldsymbol{W}_1' + \Pi_{\boldsymbol{\eta}_1}\bar{\boldsymbol{W}}_1,$$

where $\boldsymbol{W}_1'$ has the same marginal distribution with $\bar{\boldsymbol{W}}_1$ and is further independent of $\sigma\{\mathscr{F}_1, \bar{\boldsymbol{W}}_1^\intercal\boldsymbol{\eta}_1\}$. Therefore,

$$\begin{aligned}\boldsymbol{g}_1^s &= \boldsymbol{g}_1(1 - \tau s_d\boldsymbol{g}_1^\intercal\boldsymbol{\eta}_1/d)\\ &\quad - \tau s_d\frac{\|\bar{\boldsymbol{W}}_1^\intercal\boldsymbol{\eta}_1\|_2^2 - \langle(\boldsymbol{W}_1')^\intercal\boldsymbol{\eta}_1, \bar{\boldsymbol{W}}_1^\intercal\boldsymbol{\eta}_1\rangle}{\|\boldsymbol{\eta}_1\|_2^2}\boldsymbol{\eta}_1 - \tau s_d\boldsymbol{W}_1'\bar{\boldsymbol{W}}_1^\intercal\boldsymbol{\eta}_1\\ &= (1-\mu_1)\boldsymbol{g}_1 - \beta_1\boldsymbol{\eta}_1 - \gamma_1\boldsymbol{u}_1,\end{aligned}$$

where

$$\boldsymbol{u}_1 := \sqrt{d}\tau \boldsymbol{W}_1' \bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1 / \|\bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1\|_2.$$

Note that $\tau \in \mathcal{F}_1$, and $\boldsymbol{W}_1'$ is independent of $\sigma\{\mathcal{F}_1, \bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1\}$. We hence obtain that $\boldsymbol{u}_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, and is independent of $\sigma\{\bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1, \mathcal{F}_1\}$. In the equation above, we have

$$\mu_1 := \frac{1}{d}\tau s_d \boldsymbol{g}_1^\top \boldsymbol{\eta}_1, \qquad \beta_1 := \tau s_d \frac{\|\bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1\|_2^2 - \langle (\boldsymbol{W}_1')^\top \boldsymbol{\eta}_1, \bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1 \rangle}{\|\boldsymbol{\eta}_1\|_2^2},$$
$$\gamma_1 := \frac{s_d}{\sqrt{d}} \|\bar{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1\|_2.$$

We define the sigma algebra $\mathcal{G}_1 := \sigma\{\mu_1, \boldsymbol{g}_1, \beta_1, \boldsymbol{\eta}_1, \gamma_1, \boldsymbol{u}_1, \boldsymbol{x}, \boldsymbol{g}_2, \{\boldsymbol{W}_i\}_{3\le i\le l+1}\}$. Note that $\mathcal{G}_1$ depends on $\boldsymbol{W}_2$ only through $\boldsymbol{g}_2 = \boldsymbol{W}_2 \boldsymbol{h}_1$ and $\boldsymbol{y}_2 = \boldsymbol{W}_2^\top \boldsymbol{\eta}_2$. In the following parts of the proof, we will assume $s_d \to S_0$ for some positive constant S_0. Under such choice of s_d, we can prove the following two lemmas.

Lemma 6.3. *Under the assumptions of Theorem 5.1, if we further assume that $s_d \to S_0$ for some positive constant S_0, then as $d \to \infty$, we have $\mu_1 = o_P(1)$, $\beta_1 = O_P(1)$, $\gamma_1 = o_P(1)$.*

Lemma 6.4. *Under the assumptions of Theorem 5.1, if we further assume that $s_d \to S_0$ for some positive constant S_0, then the following limits hold as $d \to \infty$:*

$$\frac{1}{d_1}\|\Pi_{\boldsymbol{h}_1}^\perp \sigma(\boldsymbol{g}_1^s)\|_2^2 \xrightarrow{\mathbb{P}} 0, \qquad \frac{\langle \boldsymbol{h}_1, \sigma(\boldsymbol{g}_1^s)\rangle}{\|\boldsymbol{h}_1\|_2^2} \xrightarrow{\mathbb{P}} 1.$$

The proofs of Lemmas 6.3 and 6.4 are deferred to Appendices A.7 and A.8, respectively.

For $2 \le i \le l$, we define $\boldsymbol{g}_i^s := \boldsymbol{W}_i \sigma(\boldsymbol{g}_{i-1}^s) \in \mathbb{R}^{d_i}$ as the output of an intermediate layer of the neural network with the input being the adversarial example. In summary, we have shown $\mathcal{H}_m$ holds for $m = 1$ with $\mathcal{H}_m$ stated below. Next, we proceed by induction and show that $\mathcal{H}_m$ holds for all $m \in [l]$.

$\boldsymbol{\mathcal{H}_m}$**.** We make five claims.

(i) There exists $\boldsymbol{u}_m \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m})$, that is independent of

$$\mathcal{F}_m := \sigma\{\boldsymbol{g}_m, \boldsymbol{\eta}_m, \{\boldsymbol{W}_i\}_{m+1\le i\le l+1}, \boldsymbol{h}_{m-1}\},$$

such that

$$\boldsymbol{g}_m^s = (1 - \mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{\eta}_m - \gamma_m \boldsymbol{u}_m.$$

(ii) Let

$$\mathcal{G}_m := \sigma\{\mu_m, \boldsymbol{g}_m, \beta_m, \boldsymbol{\eta}_m, \gamma_m, \boldsymbol{u}_m, \boldsymbol{h}_{m-1}, \boldsymbol{g}_{m+1}, \{\boldsymbol{W}_i\}_{m+2\le i\le l+1}\}.$$

Then $\mathcal{G}_m$ depends on $\boldsymbol{W}_{m+1}$ only through $\boldsymbol{g}_{m+1} = \boldsymbol{W}_{m+1}\boldsymbol{h}_m$ and $\boldsymbol{y}_{m+1} = \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1}$. In particular, $\mathcal{G}_m \perp \Pi_{\boldsymbol{\eta}_{m+1}}^\perp \boldsymbol{W}_{m+1} \Pi_{\boldsymbol{h}_m}^\perp$.

(iii) For $(\mu_m, \beta_m, \gamma_m)$ in (i), the following results hold: $\mu_m = o_P(1)$, $\beta_m = O_P(1)$, $\gamma_m = o_P(1)$.

(iv) As $d \to \infty$, $\|\Pi^{\perp}_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)\|_2^2/d_m \xrightarrow{\mathbb{P}} 0$ and $\langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle/\|\boldsymbol{h}_m\|_2^2 \xrightarrow{\mathbb{P}} 1$.

(v) There exists a random variable R_m and a positive constant α_m, whose distribution and value depend only on (σ, m, l). In particular, they are independent of the input dimension and the number of neurons. Furthermore, they satisfy $\beta_m \geq \alpha_m\beta_{m-1} + R_m + o_P(1)$.

Note that claim (v) does not apply for the base case $m = 1$. Next, we will show that if $\mathcal{H}_m$ holds for all $m \leq l-1$, then this further implies that $\mathcal{H}_{m+1}$ holds.

Proofs of $\mathcal{H}_{m+1}$ claims (i) *and* (ii). By $\mathcal{H}_m$ claim (i), we have

$$\begin{aligned} \boldsymbol{g}^s_{m+1} &= \boldsymbol{W}_{m+1}\sigma(\boldsymbol{g}^s_m) \\ &= \boldsymbol{W}_{m+1}\sigma((1-\mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{D}^m_\sigma \boldsymbol{W}^\top_{m+1}\boldsymbol{\eta}_{m+1} - \gamma_m \boldsymbol{u}_m). \end{aligned} \tag{6.1}$$

From $\mathcal{H}_m$ claim (ii), we see that $\mathcal{G}_m$ depends on $\boldsymbol{W}_{m+1}$ only through

$$\boldsymbol{g}_{m+1} = \boldsymbol{W}_{m+1}\boldsymbol{h}_m \quad \text{and} \quad \boldsymbol{y}_{m+1} = \boldsymbol{W}^\top_{m+1}\boldsymbol{\eta}_{m+1},$$

$\boldsymbol{h}_m$ is independent of $\boldsymbol{W}_{m+1}$, and $\boldsymbol{\eta}_{m+1}$ depends on $\boldsymbol{W}_{m+1}$ only through

$$\boldsymbol{g}_{m+1} = \boldsymbol{W}_{m+1}\boldsymbol{h}_m.$$

Therefore, invoking Lemma 3.1, we find that exists $\widetilde{\boldsymbol{W}}_{m+1}$ that has the same marginal distribution as $\boldsymbol{W}_{m+1}$ and is independent of $\mathcal{G}_m$, such that

$$\boldsymbol{W}_{m+1} = \frac{\boldsymbol{g}_{m+1}\boldsymbol{h}^\top_m}{\|\boldsymbol{h}_m\|_2^2} + \Pi^\perp_{\boldsymbol{\eta}_{m+1}}\widetilde{\boldsymbol{W}}_{m+1}\Pi^\perp_{\boldsymbol{h}_m} + \frac{\boldsymbol{\eta}_{m+1}\boldsymbol{y}^\top_{m+1}\Pi^\perp_{\boldsymbol{h}_m}}{\|\boldsymbol{\eta}_{m+1}\|_2^2}, \tag{6.2}$$

Next, we substitute equation (6.2) into equation (6.1), which leads to the following equality

$$\begin{aligned} \boldsymbol{g}^s_{m+1} = &\frac{\langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle}{\|\boldsymbol{h}_m\|_2^2}\boldsymbol{g}_{m+1} \\ &+ \frac{\boldsymbol{\eta}^\top_{m+1}\boldsymbol{W}_{m+1}\Pi^\perp_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m) - \boldsymbol{\eta}^\top_{m+1}\widetilde{\boldsymbol{W}}_{m+1}\Pi^\perp_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)}{\|\boldsymbol{\eta}_{m+1}\|_2^2}\boldsymbol{\eta}_{m+1} \\ &+ \widetilde{\boldsymbol{W}}_{m+1}\Pi^\perp_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m) \\ = &(1-\mu_{m+1})\boldsymbol{g}_{m+1} - \beta_{m+1}\boldsymbol{\eta}_{m+1} - \gamma_{m+1}\boldsymbol{u}_{m+1}, \end{aligned}$$

where

$$\boldsymbol{u}_{m+1} := -\sqrt{d_m}\widetilde{\boldsymbol{W}}_{m+1}\Pi^\perp_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)\|\Pi^\perp_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)\|_2^{-1}.$$

Since $\widetilde{\boldsymbol{W}}_{m+1}$ is independent of $\mathscr{G}_m$, and $\Pi^{\perp}_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m) \in \mathscr{G}_m$, we then conclude that $\boldsymbol{u}_{m+1} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m+1}})$, and $\boldsymbol{u}_{m+1}$ is independent of $\mathscr{G}_m$. Since $\boldsymbol{h}_m = \sigma(\boldsymbol{g}_m)$ and $\boldsymbol{\eta}_{m+1}$ is a function of $\boldsymbol{g}_{m+1}$ and $\{\boldsymbol{W}_i\}_{m+2\le i\le l+1}$, we obtain that

$$\mathscr{F}_{m+1} \subseteq \mathscr{G}_m \quad \text{and} \quad \boldsymbol{u}_{m+1} \perp \mathscr{F}_{m+1}.$$

Thus, we have completed the proof of $\mathscr{H}_{m+1}$ claim (i). Furthermore,

$$(\mu_{m+1}, \beta_{m+1}, \gamma_{m+1})$$

can be expressed as follows:

$$\begin{aligned}
\mu_{m+1} &= 1 - \frac{\langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle}{\|\boldsymbol{h}_m\|_2^2},\\
\beta_{m+1} &= \frac{-\boldsymbol{\eta}_{m+1}^{\mathsf{T}}\boldsymbol{W}_{m+1}\Pi^{\perp}_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m) + \boldsymbol{\eta}_{m+1}^{\mathsf{T}}\widetilde{\boldsymbol{W}}_{m+1}\Pi^{\perp}_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)}{\|\boldsymbol{\eta}_{m+1}\|_2^2},\\
\gamma_{m+1} &= \frac{1}{\sqrt{d_m}}\|\Pi^{\perp}_{\boldsymbol{h}_m}\sigma(\boldsymbol{g}^s_m)\|_2.
\end{aligned} \tag{6.3}$$

Notice that $\mathscr{G}_{m+1}$ depends on $\boldsymbol{W}_{m+2}$ only through

$$\boldsymbol{g}_{m+2} = \boldsymbol{W}_{m+2}\boldsymbol{h}_{m+1} \quad \text{and} \quad \boldsymbol{y}_{m+2} = \boldsymbol{W}_{m+2}^{\mathsf{T}}\boldsymbol{\eta}_{m+2},$$

thus proving $\mathscr{H}_{m+1}$ claim (ii).

Proofs of $\mathscr{H}_{m+1}$ claims (iii) *and* (v). The following lemma is a direct consequence of the induction hypothesis.

Lemma 6.5. *Under the assumptions of Theorem 5.1, if we further assume that $s_d \to S_0$ for some positive constant S_0, and $\mathscr{H}_m$ holds, then $\mu_{m+1} = o_P(1)$ and $\gamma_{m+1} = o_P(1)$.*

The proof of Lemma 6.5 is deferred to Section A.9. We define the random object

$$\mathcal{V}_{m+1} := (\boldsymbol{g}_m, \boldsymbol{D}^m_\sigma, \boldsymbol{h}_m, \boldsymbol{h}_{m-1}, \boldsymbol{\eta}_{m+1}, \boldsymbol{u}_m, \boldsymbol{W}_m).$$

Note that $\mathcal{V}_{m+1}$ depends on $\boldsymbol{W}_{m+1}$ only through $\boldsymbol{g}_{m+1} = \boldsymbol{W}_{m+1}\boldsymbol{h}_m$. Since $\boldsymbol{h}_m$ is independent of $\boldsymbol{W}_{m+1}$, by Lemma 3.1, we can write

$$\boldsymbol{W}_{m+1} = \bar{\boldsymbol{W}}_{m+1}\Pi^{\perp}_{\boldsymbol{h}_m} + \boldsymbol{W}_{m+1}\Pi_{\boldsymbol{h}_m},$$

where $\bar{\boldsymbol{W}}_{m+1} \in \mathbb{R}^{d_m \times d_{m+1}}$ has the same marginal distribution with $\boldsymbol{W}_{m+1}$, and is independent of $\mathcal{V}_{m+1}$. Next we prove that $\beta_{m+1} = O_P(1)$.

We consider the first term in the enumerator of the definition of β_{m+1} in equation (6.3), and substitute in the decompositions we just obtained, which gives

$$\begin{aligned}
&\langle \Pi^{\perp}_{\boldsymbol{h}_m} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}, \sigma(\boldsymbol{g}^s_m)\rangle \\
&\quad = \langle \Pi^{\perp}_{\boldsymbol{h}_m} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}, \sigma((1-\mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{D}^m_\sigma \Pi^{\perp}_{\boldsymbol{h}_m} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} \\
&\qquad - \beta_m \boldsymbol{D}^m_\sigma \Pi_{\boldsymbol{h}_m} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} - \gamma_m \boldsymbol{u}_m)\rangle \\
&\quad = \langle \Pi^{\perp}_{\boldsymbol{h}_m} \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}, \sigma((1-\mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{D}^m_\sigma \Pi^{\perp}_{\boldsymbol{h}_m} \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} \\
&\qquad - \beta_m \boldsymbol{D}^m_\sigma \Pi_{\boldsymbol{h}_m} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} - \gamma_m \boldsymbol{u}_m)\rangle \\
&\quad = \langle \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}, \sigma((1-\mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{D}^m_\sigma \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} \\
&\qquad - \beta_m \delta_{m+1} \boldsymbol{D}^m_\sigma \boldsymbol{h}_m - \gamma_m \boldsymbol{u}_m)\rangle \\
&\qquad - \langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle \langle \boldsymbol{h}_m, \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}\rangle / \|\boldsymbol{h}_m\|_2^2,
\end{aligned} \tag{6.4}$$

where

$$\delta_{m+1} := \frac{1}{\|\boldsymbol{h}_m\|_2^2}(\boldsymbol{h}_m^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1} - \boldsymbol{h}_m^{\mathsf{T}} \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}).$$

Lemma 6.1 together with the fact that $\bar{\boldsymbol{W}}_{m+1}$ is independent of $(\boldsymbol{h}_m, \boldsymbol{\eta}_{m+1})$ implies that $\delta_{m+1} = O_P(d_m^{-1})$.

From $\mathcal{H}_m$ claim (iv), we see that as $d \to \infty$, $\langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle / \|\boldsymbol{h}_m\|_2^2 \xrightarrow{\mathbb{P}} 1$. Furthermore, since $\bar{\boldsymbol{W}}_{m+1} \perp \mathcal{V}_{m+1}$, then conditioning on $(\boldsymbol{h}_m, \boldsymbol{\eta}_{m+1})$, we have

$$\langle \boldsymbol{h}_m, \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}\rangle \stackrel{d}{=} \mathsf{N}(0, \|\boldsymbol{h}_m\|_2^2 \|\boldsymbol{\eta}_{m+1}\|_2^2 / d_m).$$

Putting together these results and Lemma 6.1, we conclude that

$$\langle \boldsymbol{h}_m, \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}\rangle \xrightarrow{d} \mathsf{N}(0, H_m^2 E_{m+1}^2),$$

where $H_m :=$ p-lim$\|\boldsymbol{h}_m\|_2/\sqrt{d_m}$ and $E_{m+1} :=$ p-lim$\|\boldsymbol{\eta}_{m+1}\|_2$. In summary, we have

$$\frac{\langle \boldsymbol{h}_m, \sigma(\boldsymbol{g}^s_m)\rangle}{\|\boldsymbol{h}_m\|_2^2} \langle \boldsymbol{h}_m, \bar{\boldsymbol{W}}^{\mathsf{T}}_{m+1}\boldsymbol{\eta}_{m+1}\rangle \xrightarrow{d} \mathsf{N}(0, H_m^2 E_{m+1}^2). \tag{6.5}$$

Thus, the second term in the last line of equation (6.4) is $O_P(1)$.

Next, we consider the first term in the last line of equation (6.4). Conditioning on $\boldsymbol{h}_{m-1}$, we have $\boldsymbol{g}_m = \boldsymbol{z}_m \nu_m$, where

$$\nu_m := \sqrt{\|\boldsymbol{h}_{m-1}\|_2^2 / d_{m-1}}, \quad \boldsymbol{z}_m = \nu_m^{-1} \boldsymbol{W}_m \boldsymbol{h}_{m-1} \in \mathbb{R}^{d_m}.$$

Since $\boldsymbol{W}_m$ is independent of $\boldsymbol{h}_{m-1}$, we can conclude that $\boldsymbol{z}_m \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m})$ and is independent of $\boldsymbol{h}_{m-1}$. By $\mathcal{H}_m$ claim (i), $\boldsymbol{u}_m$ is independent of $\boldsymbol{g}_m$ and $\boldsymbol{h}_{m-1}$. Therefore, $\boldsymbol{u}_m$ is further independent of $(\boldsymbol{z}_m, \boldsymbol{h}_{m-1})$. Then since $\bar{\boldsymbol{W}}_{m+1}$ is independent of $(\boldsymbol{g}_m, \boldsymbol{h}_{m-1}, \boldsymbol{u}_m, \boldsymbol{\eta}_{m+1})$, we can write

$$\overline{\boldsymbol{W}}_{m+1}^{\mathsf{T}} \boldsymbol{\eta}_{m+1} = \overline{\gamma}_{m+1} \overline{\boldsymbol{u}}_{m+1} / \sqrt{d_m},$$

where $\overline{\boldsymbol{u}}_{m+1} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m})$, is independent of $(\boldsymbol{z}_m, \boldsymbol{u}_m, \boldsymbol{h}_{m-1})$ and $\overline{\gamma}_{m+1} = \|\boldsymbol{\eta}_{m+1}\|_2$. In summary, we have

$$\boldsymbol{z}_m, \boldsymbol{u}_m, \overline{\boldsymbol{u}}_{m+1} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m}), \quad (\boldsymbol{z}_m, \boldsymbol{u}_m, \overline{\boldsymbol{u}}_{m+1}) \perp \boldsymbol{h}_{m-1}. \tag{6.6}$$

For $\boldsymbol{\theta} \in \mathbb{R}^6$, we define

$$\begin{aligned}
h_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}, z, u) &:= \theta_1 \overline{u} \sigma\big((1-\theta_2)\theta_3 z - \theta_4 \sigma'(\theta_3 z)\theta_1 \overline{u} - \theta_5 \sigma'(\theta_3 z)\sigma(\theta_3 z) - \theta_6 u\big), \\
\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}, z, u) &:= \theta_1 \overline{u} \sigma\big((1-\theta_2)\theta_3 z - \theta_4 \sigma'(H_{m-1} z)\theta_1 \overline{u} \\
&\qquad - \theta_5 \sigma'(H_{m-1} z)\sigma(\theta_3 z) - \theta_6 u\big),
\end{aligned}$$

where $H_{m-1} = \text{p-lim}\sqrt{\|\boldsymbol{h}_{m-1}\|_2^2 / d_{m-1}}$. We further define the empirical processes evaluated at $\boldsymbol{\theta}$ as

$$\begin{aligned}
\mathbb{G}_d^{(m+1)}(\boldsymbol{\theta}) &:= \frac{1}{\sqrt{d_m}} \sum_{i=1}^{d_m} \big(h_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i}) \\
&\qquad - \mathbb{E}\big[h_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i})\big]\big), \\
\overline{\mathbb{G}}_d^{(m+1)}(\boldsymbol{\theta}) &:= \frac{1}{\sqrt{d_m}} \sum_{i=1}^{d_m} \big(\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i}) \\
&\qquad - \mathbb{E}\big[\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i})\big]\big),
\end{aligned}$$

where the expectations are taken over $\{(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i})\}_{i \le d_m} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)$. Here, $\overline{u}_{m+1,i}$ is the i-th coordinate of $\overline{\boldsymbol{u}}_{m+1}$, $z_{m,i}$ is the i-th coordinate of $\boldsymbol{z}_m$, and $u_{m,i}$ is the i-th coordinate of $\boldsymbol{u}_m$. For $\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \in \mathbb{R}^6$, we define the covariance function $\overline{c}^{(m+1)}(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}})$ as

$$\begin{aligned}
\overline{c}^{(m+1)}(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) &:= \mathbb{E}\big[\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}, z, u)\overline{h}_{\overline{\boldsymbol{\theta}}}^{(m+1)}(\overline{u}, z, u)\big] \\
&\qquad - \mathbb{E}\big[\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}, z, u)\big]\mathbb{E}\big[\overline{h}_{\overline{\boldsymbol{\theta}}}^{(m+1)}(\overline{u}, z, u)\big],
\end{aligned} \tag{6.7}$$

where the expectations are taken over $(\overline{u}, z, u) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)$. Since σ' is almost everywhere continuous, and by assumption almost everywhere we have

$$|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1}),$$

then standard application of the dominated convergence theorem shows that the covariance function $\overline{c}^{(m+1)}(\cdot, \cdot)$ is continuous. Recall that

$$H_{m-1} = \text{p-lim}\sqrt{\|\boldsymbol{h}_{m-1}\|_2^2 / d_{m-1}} = \text{p-lim}\nu_m.$$

Furthermore, we define $E_{m+1} := \text{p-lim}\|\boldsymbol{\eta}_{m+1}\|_2 = \text{p-lim}\overline{\gamma}_{m+1}$. The following lemma establishes a weak convergence result for $\overline{\mathbb{G}}_d^{(m+1)}$.

Lemma 6.6. *Let $\Omega_{m+1} := \{\boldsymbol{x} \in \mathbb{R}^6 : \|\boldsymbol{x}\|_\infty \le H_{m-1} + E_{m+1}\}$, and $C(\Omega_{m+1})$ be the space of continuous functions on Ω_{m+1} endowed with the supremum norm. Under the conditions of Theorem 5.1, if we further assume that there exists $S_0 > 0$ such that $s_d \to S_0$ as $d \to \infty$, and induction hypothesis $\mathcal{H}_m$ holds, then $\{\overline{\mathbb{G}}_d^{(m+1)}\}_{d\ge 1}$ converges weakly in $C(\Omega_{m+1})$ to $\overline{\mathbb{G}}^{(m+1)}$ as $d \to \infty$, which is a Gaussian process with mean zero and covariance defined in equation (6.7).*

The proof of Lemma 6.6 is deferred to Section A.10.

Lemma 6.7. *Under the conditions of Theorem 5.1, if we further assume that there exists $S_0 > 0$ such that $s_d \to S_0$ as $d \to \infty$, and induction hypothesis $\mathcal{H}_m$ holds, then as $d \to \infty$, we have*

$$\begin{aligned}&\mathbb{G}_d^{(m+1)}(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m\delta_{m+1}, \gamma_m)\\ &\quad = \overline{\mathbb{G}}_d^{(m+1)}(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m\delta_{m+1}, \gamma_m) + o_P(1).\end{aligned}$$

The proof of Lemma 6.7 is deferred to Section A.11. Next, we will apply Lemmas 6.6 and 6.7 to show that $\beta_{m+1} = O_P(1)$ (thus $\mathcal{H}_{m+1}$ claim (iii) holds) and $\mathcal{H}_{m+1}$ claim (v). Note that

$$\begin{aligned}&\langle \overline{\boldsymbol{W}}_{m+1}^{\mathsf{T}}\boldsymbol{\eta}_{m+1}, \sigma((1-\mu_m)\boldsymbol{g}_m - \beta_m \boldsymbol{D}_\sigma^m \overline{\boldsymbol{W}}_{m+1}^{\mathsf{T}}\boldsymbol{\eta}_{m+1} - \beta_m\delta_{m+1}\boldsymbol{D}_\sigma^m \boldsymbol{h}_m - \gamma_m \boldsymbol{u}_m)\rangle\\ &= \frac{1}{\sqrt{d_m}}\sum_{i=1}^{d_m} \overline{\gamma}_{m+1}\overline{u}_{m+1,i} \times \sigma\big((1-\mu_m)\nu_m z_{m,i} - d_m^{-1/2}\beta_m\overline{\gamma}_{m+1}\sigma'(\nu_m z_{m,i})\overline{u}_{m+1,i}\\ &\qquad - \beta_m\delta_{m+1}\sigma'(\nu_m z_{m,i})\sigma(\nu_m z_{m,i}) - \gamma_m u_{m,i}\big)\\ &\overset{(a)}{=} \mathbb{G}_d^{(m+1)}(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m\delta_{m+1}, \gamma_m)\\ &\qquad - \beta_m\overline{\gamma}_{m+1}^2\mathbb{E}\big[\sigma'(\nu_m z)\sigma'\big((1-\mu_m)\nu_m z - d_m^{-1/2}\beta_m\overline{\gamma}_{m+1}\sigma'(\nu_m z)\overline{u}\\ &\qquad - \beta_m\delta_{m+1}\sigma'(\nu_m z)\sigma(\nu_m z) - \gamma_m u\big)\big],\end{aligned} \tag{6.8}$$

where the expectation is taken over $(z, u, \overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)$. In step (a) we apply Stein's lemma to derive the equality. By $\mathcal{H}_m$ claim (iii), we have

$$\mu_m = o_P(1), \quad \beta_m/\sqrt{d_m} = o_P(1), \quad \gamma_m = o_P(1).$$

Recall that we have shown $\delta_{m+1} = O_P(d_m^{-1})$, thus $\beta_m\delta_{m+1} = o_P(1)$. By Lemma 6.1, we have

$$\nu_m = H_{m-1} + o_P(1), \quad \overline{\gamma}_{m+1} = E_{m+1} + o_P(1).$$

Therefore, combining Lemmas 6.6 and 6.7, we conclude that for any $\epsilon > 0$,

$$\left|\mathbb{G}_d^{(m+1)}(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m\delta_{m+1}, \gamma_m)\right| \le \sup_{\boldsymbol{\theta}\in A_\epsilon^{(m+1)}} |\overline{\mathbb{G}}_d^{(m+1)}| + o_P(1),$$

where

$$\begin{aligned} A_\epsilon^{(m+1)} := \big\{\boldsymbol{\theta} \in \mathbb{R}^6 : |\theta_1 - E_{m+1}| \le \epsilon, |\theta_2| \le \epsilon, \\ |\theta_3 - H_{m-1}| \le \epsilon, |\theta_4| \le \epsilon, |\theta_5| \le \epsilon, |\theta_6| \le \epsilon\big\}. \end{aligned}$$

Furthermore, invoking the dominated convergence theorem, we see that as $d \to \infty$,

$$\begin{aligned} &\overline{\gamma}_{m+1}^2 \mathbb{E}\big[\sigma'(\nu_m z)\sigma'\big((1-\mu_m)\nu_m z - d_m^{-1/2}\beta_m\overline{\gamma}_{m+1}\sigma'(\nu_m z)\overline{u} \\ &\qquad - \beta_m\delta_{m+1}\sigma'(\nu_m z)\sigma(\nu_m z) - \gamma_m u\big)\big] \\ &\quad = E_{m+1}^2\mathbb{E}\big[\sigma'(H_{m-1}z_m)^2\big] + o_P(1). \end{aligned} \tag{6.9}$$

We define $M_\epsilon^{(m+1)}(\mathbb{G}) := \sup_{\boldsymbol{\theta}\in A_\epsilon^{(m+1)}} |\mathbb{G}(\boldsymbol{\theta})|$, then $M_\epsilon^{(m+1)}$ is a continuous function with respect to the supremum norm $\ell^\infty(\Omega_{m+1})$. Using Lemma 6.6 together with the continuous mapping theorem, we see that $M_\epsilon^{(m+1)}(\overline{\mathbb{G}}_d^{(m+1)})$ converges in distribution to $M_\epsilon^{(m+1)}(\overline{\mathbb{G}}^{(m+1)})$. Notice that if we let $\epsilon = 1$, then E_{m+1}, H_{m-1}, $\overline{c}^{(m+1)}$, $A_1^{(m+1)}$ depend only on $(l, m+1, \sigma)$, thus the distribution of $M_1^{(m+1)}(\overline{\mathbb{G}}^{(m+1)})$ also depends uniquely on $(l, m+1, \sigma)$. By $\mathcal{H}_m$ claim (iii), we have $\beta_m = O_P(1)$. Putting this together with equations (6.8) and (6.9), we obtain that there exists a random variable $R_{m+1}^{(1)}$, the distribution of which depends only on $(l, m+1, \sigma)$, such that as $d \to \infty$,

$$\begin{aligned} &\big|\big\langle \overline{\boldsymbol{W}}_{m+1}^\mathsf{T}\boldsymbol{\eta}_{m+1}, \sigma\big((1-\mu_m)\boldsymbol{g}_m - \beta_m\boldsymbol{D}_\sigma^m\overline{\boldsymbol{W}}_{m+1}^\mathsf{T}\boldsymbol{\eta}_{m+1} - \beta_m\delta_{m+1}\boldsymbol{D}_\sigma^m\boldsymbol{h}_m - \gamma_m\boldsymbol{u}_m\big)\big\rangle \\ &\qquad + \beta_m E_{m+1}^2\mathbb{E}\big[\sigma'(H_{m-1}z_m)^2\big]\big| \\ &\quad \le R_{m+1}^{(1)} + o_P(1). \end{aligned} \tag{6.10}$$

Finally, we consider the second term in the enumerator of the definition of β_{m+1} given in equation (6.3). Conditioning on $(\boldsymbol{\eta}_{m+1}, \Pi_{\boldsymbol{h}_m}^\perp\sigma(\boldsymbol{g}_m^s))$,

$$\boldsymbol{\eta}_{m+1}^\mathsf{T}\widetilde{\boldsymbol{W}}_{m+1}\Pi_{\boldsymbol{h}_m}^\perp\sigma(\boldsymbol{g}_m^s) \overset{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_{m+1}\|_2^2\|\Pi_{\boldsymbol{h}_m}^\perp\sigma(\boldsymbol{g}_m^s)\|_2^2/d_m\big), \tag{6.11}$$

which is $o_P(1)$ by Lemma 6.1 and $\mathcal{H}_m$ claim (iv). Note that

$$\|\boldsymbol{\eta}_{m+1}\|_2^{-2} = E_{m+1}^{-2} + o_P(1).$$

Taking this collectively with equations (6.3)–(6.5), (6.10), (6.11), we obtain that

$$\beta_{m+1} = O_P(1).$$

Furthermore, there exists a random variable R_{m+1} and a positive number α_{m+1}, the distribution (and the value) of which is a function of $(\sigma, m+1, l)$ only, such that

$$\beta_{m+1} \geq \beta_m \alpha_{m+1} + R_{m+1} + o_P(1).$$

This completes the proof of $\mathcal{H}_{m+1}$ claims (iii), (v).

Proof of $\mathcal{H}_{m+1}$ claim (iv). Finally, we prove $\mathcal{H}_{m+1}$ claim (iv), which is achieved by the following lemma.

Lemma 6.8. *Under the assumptions of Theorem 5.1, if we further assume that $s_d \to S_0$ for some positive constant S_0, $\mathcal{H}_m$ holds, and claims* (i), (ii), (iii), (v) *from $\mathcal{H}_{m+1}$ hold, then as $d \to \infty$ we have the following convergences:*

$$\frac{1}{d_{m+1}} \|\Pi^{\perp}_{\boldsymbol{h}_{m+1}} \sigma(\boldsymbol{g}^s_{m+1})\|_2^2 \xrightarrow{\mathbb{P}} 0, \qquad \frac{\langle \boldsymbol{h}_{m+1}, \sigma(\boldsymbol{g}^s_{m+1})\rangle}{\|\boldsymbol{h}_{m+1}\|_2^2} \xrightarrow{\mathbb{P}} 1.$$

The proof of Lemma 6.8 is deferred to Section A.12. We note that $\mathcal{H}_{m+1}$ claim (iv) is a direct consequence of Lemma 6.8. By induction, we have completed the proof of $\mathcal{H}_i$ for all $i \in [l]$.

Back to the proof of the theorem. Next, we will apply results from $\mathcal{H}_l$ to prove Theorem 5.1. By $\mathcal{H}_l$ claim (i),

$$\boldsymbol{g}^s_l = (1-\mu_l)\boldsymbol{g}_l - \beta_l \boldsymbol{D}^l_\sigma \boldsymbol{W}^{\mathsf{T}}_{l+1} - \gamma_l \boldsymbol{u}_l,$$

Using our modeling assumption and $\mathcal{H}_l$ claim (i), we obtain that $(\boldsymbol{g}_l, \boldsymbol{D}^l_\sigma, \boldsymbol{u}_l)$ is independent of $\boldsymbol{W}_{l+1}$. By $\mathcal{H}_l$ claim (iii), we have

$$\mu_l = o_P(1), \quad \beta_l/\sqrt{d_l} = o_P(1), \quad \gamma_l = o_P(1).$$

We define $\zeta_l := \beta_l/\sqrt{d_l}$, $F_l: \mathbb{R}^{d_l} \to \mathbb{R}$, such that $F_l(\boldsymbol{y}) := \sum_{i=1}^{d_l} W_{l+1,i}\sigma(y_i)$. Then we have

$$\begin{aligned} &F_l(\boldsymbol{g}^s_l) - F_l(\boldsymbol{g}_l) \\ &\quad = \sum_{i=1}^{d_l} W_{l+1,i}\big(\sigma((1-\mu_l)g_{l,i} - \zeta_l \sigma'(g_{l,i})\sqrt{d_l} W_{l+1,i} - \gamma_l u_{l,i}) - \sigma(g_{l,i})\big). \end{aligned}$$

Let $z_i := \sqrt{d_l} W_{l+1,i}$, then $\{z_i\}_{i\in[d_l]} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$. We define $\boldsymbol{g}_l = \nu_l \boldsymbol{z}_l$, where

$$\nu_l := \sqrt{\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1}},$$

$\boldsymbol{z}_l = \nu_l^{-1} \boldsymbol{W}_l \boldsymbol{h}_{l-1}$. Since $\boldsymbol{W}_l$ is independent of $\boldsymbol{h}_{l-1}$, we have $\boldsymbol{z}_l \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l})$ and is independent of $\boldsymbol{h}_{l-1}$. Since $\boldsymbol{W}_{l+1}$ is independent of $(\boldsymbol{g}_l, \boldsymbol{h}_{l-1})$, we conclude that it is also independent of $(\boldsymbol{z}_l, \boldsymbol{h}_{l-1})$. By $\mathcal{H}_l$ claim (i), we know that $\boldsymbol{u}_l$ is independent of $(\boldsymbol{g}_l, \boldsymbol{h}_{l-1}, \boldsymbol{W}_{l+1})$, thus we obtain that $\boldsymbol{u}_l, \sqrt{d_l}\boldsymbol{W}_{l+1}, \boldsymbol{z}_l \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l})$. Furthermore, $(\boldsymbol{u}_l, \sqrt{d_l}\boldsymbol{W}_{l+1}, \boldsymbol{z}_l)$ is independent of $\boldsymbol{h}_{l-1}$.

By Lemma 6.1, ν_l converges in probability to a positive constant H_{l-1}. For $\boldsymbol{\theta} \in \mathbb{R}^4$, we define

$$h_{\boldsymbol{\theta}}^{(l+1)}(z, z_l, u_l) := z\big(\sigma((1-\theta_1)\theta_2 z_l - \theta_3\sigma'(\theta_2 z_l)z - \theta_4 u) - \sigma(\theta_2 z_l)\big),$$
$$\bar{h}_{\boldsymbol{\theta}}^{(l+1)}(z, z_l, u_l) := z\big(\sigma((1-\theta_1)\theta_2 z_l - \theta_3\sigma'(H_{l-1} z_l)z - \theta_4 u) - \sigma(\theta_2 z_l)\big).$$

For $\boldsymbol{\theta} \in \mathbb{R}^4$, we define the empirical processes $\mathbb{G}_d^{(l+1)}, \bar{\mathbb{G}}_d^{(l+1)}$ indexed by $\boldsymbol{\theta}$ as

$$\mathbb{G}_d^{(l+1)}(\boldsymbol{\theta}) := \frac{1}{\sqrt{d_l}}\sum_{i=1}^{d_l}\big(h_{\boldsymbol{\theta}}(z_i, z_{l,i}, u_{l,i}) - \mathbb{E}\big[h_{\boldsymbol{\theta}}(z_i, z_{l,i}, u_{l,i})\big]\big),$$
$$\bar{\mathbb{G}}_d^{(l+1)}(\boldsymbol{\theta}) := \frac{1}{\sqrt{d_l}}\sum_{i=1}^{d_l}\big(\bar{h}_{\boldsymbol{\theta}}(z_i, z_{l,i}, u_{l,i}) - \mathbb{E}\big[\bar{h}_{\boldsymbol{\theta}}(z_i, z_{l,i}, u_{l,i})\big]\big),$$

where the expectations are taken over $\{(z_i, z_{l,i}, u_{l,i})\}_{i\in[d_l]} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)$. For $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^4$, we define the covariance function $\bar{c}^{(l+1)}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ as

$$\bar{c}^{(l+1)}(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) := \mathbb{E}\big[\bar{h}_{\boldsymbol{\theta}}(z, z_l, u_l)\bar{h}_{\bar{\boldsymbol{\theta}}}(z, z_l, u_l)\big] - \mathbb{E}\big[\bar{h}_{\boldsymbol{\theta}}(z, z_l, u_l)\big]\mathbb{E}\big[\bar{h}_{\bar{\boldsymbol{\theta}}}(z, z_l, u_l)\big],$$

where the expectations are taken over $(z, z_l, u_l) \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1)$. Using the assumptions imposed on σ, σ', we can apply the dominated convergence theorem and conclude that $\bar{c}^{(l+1)}(\cdot, \cdot)$ is a continuous function. We denote by $\bar{\mathbb{G}}^{(l+1)}$ the Gaussian process with mean zero and covariance function $\bar{c}^{(l+1)}$. We define

$$\Omega_{l+1} := \{\boldsymbol{x} \in \mathbb{R}^4 : \|\boldsymbol{x}\|_\infty \le 2H_{l-1}\}.$$

Similar to the proof of Lemma 6.6, we can prove that equipped with the supremum norm $\ell^\infty(\Omega_{l+1})$, $\{\bar{\mathbb{G}}_d^{(l+1)}\}_{d\ge 1}$ converges weakly in $C(\Omega_{l+1})$ to $\bar{\mathbb{G}}^{(l+1)}$. We skip the detailed proof here for the sake of compactness. For $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \Omega_{l+1}$, we define

$$\rho(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) := \mathbb{E}\big[(\bar{\mathbb{G}}^{(l+1)}(\boldsymbol{\theta}) - \bar{\mathbb{G}}^{(l+1)}(\bar{\boldsymbol{\theta}}))^2\big]^{1/2}.$$

Then again by [25, Lemma 18.15], we can and will assume that $\bar{\mathbb{G}}^{(l+1)}$ almost surely has ρ-continuous sample path.

For $H_{l-1} > \epsilon > 0$, we let

$$B_\epsilon := \{\boldsymbol{x} \in \mathbb{R}^4 : |x_1| < \epsilon, |x_2 - H_{l-1}| < \epsilon, |x_3| < \epsilon, |x_4| < \epsilon\},$$
$$S_\epsilon^{(l+1)}(\mathbb{G}) := \sup_{\boldsymbol{\theta}\in B_\epsilon} |\mathbb{G}(\boldsymbol{\theta})|.$$

Note that $S_\epsilon^{(l+1)}$ is continuous with respect to $\ell^\infty(\Omega_{l+1})$, thus

$$S_\epsilon^{(l+1)}(\overline{\mathbb{G}}_d^{(l+1)}) \xrightarrow{d} S_\epsilon^{(l+1)}(\overline{\mathbb{G}}^{(l+1)}).$$

Recall that we have proved

$$\mu_l = o_P(1), \quad \nu_l = H_{l-1} + o_P(1), \quad \zeta_l = O_P(d_l^{-1/2}),$$
$$\gamma_l = o_P(1), \quad z_i = \sqrt{d_l} W_{l+1,i}.$$

Therefore,

$$\begin{aligned}
&\left|F_l(\boldsymbol{g}_l^s) - F_l(\boldsymbol{g}_l) + \sqrt{d_l}\zeta_l \mathbb{E}\big[\sigma'(\nu_l z_l)\sigma'((1-\mu_l)\nu_l z_l - \zeta_l\sigma'(\nu_l z_l)z - \gamma_l u_l)\big]\right| \\
&\overset{\text{(i)}}{=} \left|\frac{1}{\sqrt{d_l}}\sum_{i=1}^{d_l} z_i\big(\sigma((1-\mu_l)\nu_l z_{l,i} - \zeta_l\sigma'(\nu_l z_{l,i})z_i - \gamma_l u_{l,i}) - \sigma(\nu_l z_{l,i})\big)\right. \\
&\qquad \left. - \sqrt{d_l}\mathbb{E}\big[z\big(\sigma((1-\mu_l)\nu_l z_l - \zeta_l\sigma'(\nu_l z_l)z - \gamma_l u_l) - \sigma(\nu_l z_l)\big)\big]\right| \\
&= \left|\frac{1}{\sqrt{d_l}}\sum_{i=1}^{d_l}\big(h_{(\mu_l,\nu_l,\zeta_l,\gamma_l)}(z_i, z_{l,i}, u_{l,i}) - \mathbb{E}\big[h_{(\mu_l,\nu_l,\zeta_l,\gamma_l)}(z, z_l, u_l)\big]\big)\right| \\
&\overset{\text{(ii)}}{\le} S_\epsilon^{(l+1)}(\overline{\mathbb{G}}_d^{(l+1)}) + \delta_\epsilon(d),
\end{aligned}$$

where $\delta_\epsilon(d) \xrightarrow{\mathbb{P}} 0$ as $d \to \infty$. In the above equations, the expectations are taken over $\{z_i, z_{l,i}, u_{l,i}, z, z_l, u_l\}_{i\in[d_l]} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, (i) is by Stein's lemma, and (ii) is by an argument that is similar to the proof of Lemma 6.7. More precisely, we show that as $d \to \infty$

$$\mathbb{G}_d^{(l+1)}((\mu_l, \nu_l, \zeta_l, \gamma_l)) = \overline{\mathbb{G}}_d^{(l+1)}((\mu_l, \nu_l, \zeta_l, \gamma_l)) + o_P(1).$$

We ignore the proof of this part for the sake of simplicity, as it is basically the same as the proof of Lemma 6.7.

Since $\overline{\mathbb{G}}^{(l+1)}$ has ρ-continuous sample path, one can verify that

$$S_\epsilon^{(l+1)}(\overline{\mathbb{G}}^{(l+1)}) \xrightarrow{\mathbb{P}} 0 \quad \text{as } \epsilon \to 0^+.$$

For any $\epsilon' > 0$, we first choose ϵ small enough, such that

$$\mathbb{P}\big(S_\epsilon^{(l+1)}(\overline{\mathbb{G}}^{(l+1)}) \ge \epsilon'/3\big) \le \epsilon'/3.$$

Since $S_\epsilon^{(l+1)}(\overline{\mathbb{G}}_d^{(l+1)}) \xrightarrow{d} S_\epsilon^{(l+1)}(\overline{\mathbb{G}}^{(l+1)})$, and $\delta_\epsilon(d) \xrightarrow{\mathbb{P}} 0$ as $d \to \infty$, there exists $d_{\epsilon,\epsilon'} \in \mathbb{N}_+$, such that for all $d \ge d_{\epsilon,\epsilon'}$,

$$\mathbb{P}\big(|\delta_\epsilon(d)| \ge \epsilon'/3\big) \le \epsilon'/3$$

and

$$\mathbb{P}\big(|S_\epsilon^{(l+1)}(\overline{\mathbb{G}}_d^{(l+1)})| \geq \epsilon'/3\big) \leq \mathbb{P}\big(|S_\epsilon^{(l+1)}(\overline{\mathbb{G}}^{(l+1)})| \geq \epsilon'/3\big) + \epsilon'/3.$$

Combining these results, we conclude that for all $d \geq d_{\epsilon,\epsilon'}$,

$$\begin{aligned}&\mathbb{P}\big(\big|F_l(\boldsymbol{g}_l^s) - F_l(\boldsymbol{g}_l)\\ &\quad + \sqrt{d_l}\zeta_l \mathbb{E}\big[\sigma'(\nu_l z_l)\sigma'((1-\mu_l)\nu_l z_l - \zeta_l \sigma'(\nu_l z_l) z - \gamma_l u_l)\big]\big| \geq \epsilon'\big) \leq \epsilon',\end{aligned}$$

Application of the dominated convergence theorem shows that

$$\mathbb{E}\big[\sigma'(\nu_l z_l)\sigma'((1-\mu_l)\nu_l z_l - \alpha_l \sigma'(\nu_l z_l) z - \gamma_l u_l)\big] = \mathbb{E}\big[\sigma'(H_{l-1} z)^2\big] + o_P(1)$$

as $d \to \infty$. Recall that we have proved $\beta_l = O_P(1)$, thus

$$f(\boldsymbol{x}^s) - f(\boldsymbol{x}) = F_l(\boldsymbol{g}_l^s) - F_l(\boldsymbol{g}_l) = -\beta_l \mathbb{E}\big[\sigma'(H_{l-1} z)^2\big] + o_P(1).$$

Since σ is not a constant function, σ' is almost everywhere continuous, and $H_{l-1} > 0$, we then have $\mathbb{E}[\sigma'(H_{l-1} z)^2] > 0$. Recall that we have shown that there exist random variables $\{\mathbb{R}_m\}_{2 \leq m \leq l}$, the distributions of which depend uniquely on $(\sigma, [l])$. In addition, we have shown that there exist positive constants $\{\alpha_m\}_{2 \leq m \leq l}$, with the values of which depend uniquely on $(\sigma, [l])$, such that

$$\beta_m \geq \alpha_m \beta_{m-1} + R_m + o_P(1).$$

Furthermore, using the law of large numbers, we have $\beta_1 = \tau S_0 + o_P(1)$. By Lemma 6.1, $F_l(\boldsymbol{g}_l)$ converges in distribution to a Gaussian distribution with mean zero and variance depending only on (σ, l) (especially, independent of $\{s_d\}_{d \in \mathbb{N}_+}$). Therefore, we deduce that

$$\lim_{S_0 \to \infty} \liminf_{d \to \infty} \mathbb{P}\big(\text{sign}(f(\boldsymbol{x})) \neq \text{sign}(f(\boldsymbol{x}^s))\big) = 1.$$

Finally, we prove Theorem 5.1 via a standard diagonal argument. Note that for all $n \in \mathbb{N}_+$, there exists $S_0^n > 0$ and $d_n \in \mathbb{N}_+$, such that if we set $s_d = S_0^n$ for all $d \in \mathbb{N}_+$, then for all $d \geq d_n$,

$$\mathbb{P}\big(\text{sign}(f(\boldsymbol{x}^s)) \neq \text{sign}(f(\boldsymbol{x}))\big) \geq 1 - n^{-1}.$$

Without loss of generality, we assume that $d_{n+1} \geq d_n$, $S_0^n/\sqrt{d_n} < n^{-1}$ and $S_0^n < \xi_{d_n}$. Indeed, to achieve these conditions, we simply need to take d_n large enough. Then we set $s_d = S_0^n$ if and only if $d_n \leq d < d_{n+1}$. Under such choice of $\{s_d\}_{d \in \mathbb{N}_+}$, for all $d_{n+1} > d \geq d_n$, we have

$$\begin{aligned}&\frac{s_d}{\sqrt{d}} = \frac{S_0^n}{\sqrt{d}} \leq \frac{S_0^n}{\sqrt{d_n}} \leq \frac{1}{n}, \quad \mathbb{P}\big(\text{sign}(f(\boldsymbol{x}^s)) \neq \text{sign}(f(\boldsymbol{x}))\big) \geq 1 - n^{-1},\\ &s_d \leq \xi_{d_n} \leq \xi_d.\end{aligned}$$

Note that n is arbitrary, thus by combining the above results with the first claim of the theorem, we complete the proof of the second claim.

A. Proofs of the supporting lemmas

A.1. Proof of Lemma 3.1

We prove Lemma 3.1 in this section. Note that $\boldsymbol{X} = \Pi_{\boldsymbol{A}_1}\boldsymbol{X} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}$. Let $\boldsymbol{X}'$ be an independent copy of $\boldsymbol{X}$ that is independent of $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$. We consider the matrix $\boldsymbol{X}_1 := \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X} + \Pi_{\boldsymbol{A}_1}\boldsymbol{X}'$. Conditioning on any value of $(\boldsymbol{A}_1, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$, the conditional distribution of $\boldsymbol{X}_1$ is equal to the distribution of $\boldsymbol{X}$. Therefore, we conclude that $\boldsymbol{X}_1 \overset{d}{=} \boldsymbol{X}$ and $\boldsymbol{X}_1$ is independent of $(\boldsymbol{A}_1, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$. Notice that $\boldsymbol{X}$ admits the decomposition

$$\begin{aligned}\boldsymbol{X} &= \Pi_{\boldsymbol{A}_1}\boldsymbol{X} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{A}_2} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{A}_2}^{\perp}\\ &= \Pi_{\boldsymbol{A}_1}\boldsymbol{X} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{A}_2} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}_1\Pi_{\boldsymbol{A}_2}^{\perp}.\end{aligned}$$

We let $\boldsymbol{X}''$ be an independent copy of $\boldsymbol{X}$ that is independent of $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{Z}_1, \boldsymbol{Z}_2)$. Define

$$\boldsymbol{X}_2 = \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}_1\Pi_{\boldsymbol{A}_2}^{\perp} + \Pi_{\boldsymbol{A}_1}\boldsymbol{X}'' + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}''\Pi_{\boldsymbol{A}_2}.$$

Using the distributional property of the Gaussian ensemble and the fact that $\boldsymbol{X}_1$ is independent of $(\boldsymbol{A}_1, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$ (thus is independent of $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$), we can conclude that conditioning on any specific value of $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{A}_1\boldsymbol{X}_1, \boldsymbol{X}_1\boldsymbol{A}_2, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$, the conditional distribution of $\boldsymbol{X}_2$ is equal to the distribution of $\boldsymbol{X}$. Therefore, we deduce that $\boldsymbol{X}_2 \perp (\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{A}_1\boldsymbol{X}_1, \boldsymbol{X}_1\boldsymbol{A}_2, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$. Notice that

$$\begin{aligned}\boldsymbol{X}\boldsymbol{A}_2 &= \Pi_{\boldsymbol{A}_1}\boldsymbol{X}\boldsymbol{A}_2 + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\boldsymbol{A}_2\\ &= \Pi_{\boldsymbol{A}_1}\boldsymbol{X}\boldsymbol{A}_2 + \boldsymbol{X}_1\boldsymbol{A}_2 - \Pi_{\boldsymbol{A}_1}\boldsymbol{X}_1\boldsymbol{A}_2,\end{aligned}$$

thus $\boldsymbol{X}_2 \perp (\boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{X}\boldsymbol{A}_2, \boldsymbol{Z}_1)$. Combining the above analysis, we obtain that

$$\boldsymbol{X} = \Pi_{\boldsymbol{A}_1}\boldsymbol{X} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{A}_2} + \Pi_{\boldsymbol{A}_1}^{\perp}\boldsymbol{X}_2\Pi_{\boldsymbol{A}_2}^{\perp},$$

with $\boldsymbol{X}_2$ independent of $\boldsymbol{Y}$. Thus, we have completed the proof of the lemma.

A.2. Proof of Lemma 3.2

Conditioning on $\boldsymbol{W}\boldsymbol{x}$, the following two matrices are equal in distribution:

$$\boldsymbol{W} \overset{d}{=} \frac{1}{d}\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} + \widetilde{\boldsymbol{W}}\Pi_{\boldsymbol{x}}^{\perp},$$

where $\widetilde{\boldsymbol{W}}$ is an independent copy of $\boldsymbol{W}$ which is independent of everything else, $\Pi_{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$ is the projection operator projecting onto the linear subspace spanned by $\boldsymbol{x}$, and $\Pi_{\boldsymbol{x}}^{\perp} := \boldsymbol{I}_d - \Pi_{\boldsymbol{x}}$. Using this decomposition, we can express the distribution of the gradient as

$$\nabla f(\boldsymbol{x}) \stackrel{d}{=} \Pi_{\boldsymbol{x}}^{\perp} \widetilde{\boldsymbol{W}}^{\mathsf{T}} \boldsymbol{D}_{\sigma} \boldsymbol{a} + \frac{1}{d} \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}} \boldsymbol{D}_{\sigma} \boldsymbol{a}.$$

By assumption, almost everywhere we have $|\sigma'(x)| \le C_{\sigma}(1 + |x|^{k-1})$. Notice that $\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$ for $i \in [m]$. Therefore, we can apply Chebyshev's inequality and conclude that there exists a constant $C_1 > 0$ which depends only on σ, such that with probability at least $1 - \delta/4$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 \sigma'(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 - \mathbb{E}\big[G^2 \sigma'(G)^2\big] \right| \le \sqrt{\frac{C_1}{m\delta}},$$

where $G \sim \mathsf{N}(0,1)$. Conditioning on $\boldsymbol{g} = \boldsymbol{W}\boldsymbol{x}$, we have

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{W}^{\mathsf{T}}\boldsymbol{D}_{\sigma}\boldsymbol{a} \stackrel{d}{=} \mathsf{N}\left(0, \sum_{i=1}^{m} (\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 \sigma'(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2/m\right),$$

then using Gaussian concentration, we conclude that there exists a numerical constant $C_2 > 0$, such that with probability at least $1 - \delta/4$,

$$|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{W}^{\mathsf{T}}\boldsymbol{D}_{\sigma}\boldsymbol{a}| \le \sqrt{C_2(\log(1/\delta) + 1) \times \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 \sigma(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2}.$$

Note that $\|\Pi_{\boldsymbol{x}}^{\perp} \widetilde{\boldsymbol{W}}^{\mathsf{T}} \boldsymbol{D}_{\sigma}\boldsymbol{a}\|_2^2 \le \|\widetilde{\boldsymbol{W}}^{\mathsf{T}} \boldsymbol{D}_{\sigma}\boldsymbol{a}\|_2^2$. Let $z_1, \dots, z_d \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, then

$$\|\widetilde{\boldsymbol{W}}^{\mathsf{T}} \boldsymbol{D}_{\sigma}\boldsymbol{a}\|_2^2 \stackrel{d}{=} \left(\sum_{i=1}^{m} \sigma'(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 a_i^2\right) \times \left(\frac{1}{d}\sum_{j=1}^{d} z_j^2\right).$$

Using Bernstein's inequality, with probability at least $1 - \delta/4$,

$$\left| \frac{1}{d}\sum_{j=1}^{d} z_j^2 - 1 \right| \le C_3(\log(1/\delta) + 1)/\sqrt{d}$$

for some absolute constant C_3. Again by Chebyshev's inequality, with probability at least $1 - \delta/4$,

$$\left| \sum_{i=1}^{m} \sigma'(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x})^2 a_i^2 - \mathbb{E}\big[\sigma'(G_1)^2 G_2^2\big] \right| \le \sqrt{C_4/m\delta},$$

where $G_1, G_2 \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1)$ and $C_4 > 0$ is a constant depending only on σ. Then we combine the above results, and conclude that there exists $C > 0$ that depends only on σ, such that with probability at least $1 - \delta$,

$$\|\nabla f(\boldsymbol{x})\|_2 \leq C\big(1 + d^{-1/2}\log(1/\delta) + (m\delta)^{-1/2} + (md\delta)^{-1/2}\log(1/\delta)\big),$$

thus concluding the proof of the lemma.

A.3. Proof of Lemma 3.3

By assumption, almost everywhere we have $|\sigma'(x)| \leq C_\sigma(1 + |x|^{k-1})$. Let g_1 be the first coordinate of $\boldsymbol{g}$, then $\mathbb{E}[\sigma'(g_1)^2 g_1^2] < \infty$. By the law of large numbers,

$$\|\boldsymbol{g}^\top \boldsymbol{D}_\sigma\|_2^2/m = \mathbb{E}[g_1^2\sigma'(g_1)^2] + o_P(1).$$

Conditioning on $\boldsymbol{g}^\top \boldsymbol{D}_\sigma$, we have

$$\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a} \overset{d}{=} \mathsf{N}(\boldsymbol{0}, \|\boldsymbol{g}^\top \boldsymbol{D}_\sigma\|_2^2/m),$$

thereby $\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a} = O_P(1)$. By assumption we have $s_d \to S_0$, then we can conclude that $\tau s_d \boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}/d = o_P(1)$. This proves $\mu = o_P(1)$.

Similarly, we apply the law of large numbers, and obtain that

$$\|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2 = \mathbb{E}[\sigma'(g_1)^2]^{1/2} + o_P(1).$$

By assumption, σ is not a constant function and σ' is almost everywhere continuous, thus we have $\mathbb{E}[\sigma'(g_1)^2] > 0$. Given $\boldsymbol{D}_\sigma \boldsymbol{a}$, the conditional distributions of $\bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}$ and $\bar{\boldsymbol{W}}_c^\top \boldsymbol{D}_\sigma \boldsymbol{a}$ are both $\mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d \|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2/d)$, thus by the law of large numbers and Cauchy–Schwarz inequality, we can conclude that

$$\|\bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2 = O_P(1), \quad \langle \bar{\boldsymbol{W}}_c^\top \boldsymbol{D}_\sigma \boldsymbol{a}, \bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\rangle = O_P(1).$$

Combining the equations above gives $\beta = o_P(1)$ and $\gamma = o_P(1)$ as $m, d \to \infty$.

A.4. Proof of Lemma 3.4

By our assumptions imposed on σ, we see that there exists a deterministic constant $C_0 > 0$, which is a function of the activation function σ only, such that for all $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \Omega$,

$$\begin{aligned}|h_{\boldsymbol{\theta}}(b, g, u) - h_{\bar{\boldsymbol{\theta}}}(b, g, u)| \leq C_0\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2 |b| \sqrt{g^2 + b^2 + b^2 g^{2k} + u^2} \\ \times \big(1 + |g|^k + |b|^k + |b|^k|g|^{k^2} + |u|^k\big).\end{aligned}$$

We define

$$m(b,g,u) := C_0|b|\sqrt{g^2+b^2+b^2g^{2k}+u^2}\times\left(1+|g|^k+|b|^k+|b|^k|g|^{k^2}+|u|^k\right).$$

One can verify that $\|m\|_2^2 := \mathbb{E}[m(b,g,u)^2] < \infty$, where the expectation is taken over $b, g, u \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$.

Let $\mathcal{F} := \{h_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$. For the sake of completeness, we reproduce the definition of bracketing number introduced in [25]. Given two functions $e_1, e_2 : \mathbb{R}^3 \to \mathbb{R}$, we define the bracket $[e_1, e_2]$ as the set of all functions h such that $e_1(z) \le h_z \le e_2(z)$ for all $z \in \Omega$. An ϵ-bracket in L_2 is a bracket $[e_1, e_2]$ such that

$$\mathbb{E}[(e_1(b,g,u) - e_2(b,g,u))^2]^{1/2} < \epsilon,$$

where again the expectation is taken over $b, g, u \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_2)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. We define the *bracketing integral* as

$$J_{[]}(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2)}\, \mathrm{d}\epsilon.$$

The following lemma is from [25, Theorem 19.5].

Lemma A.1. *If $J_{[]}(1, \mathcal{F}, L_2) < \infty$, then $\mathbb{G}_m$ converges weakly in $C(\Omega)$ to $\mathbb{G}$.*

By Lemma A.1, to prove the theorem, we only need to show that the bracketing integral is finite. By [25, Example 19.7], there exists a numerical constant $K > 0$, such that

$$N_{[]}(\epsilon\|m\|_2, \mathcal{F}, L_2) \le K\epsilon^{-3}.$$

It is not hard to see that there exists another constant $K_0 > 0$, which depends only on $(K, \|m\|_2)$, such that

$$J_{[]}(1, \mathcal{F}, L_2) \le K_0 \int_0^1 (\log(1/\epsilon)+1)^{1/2}\, \mathrm{d}\epsilon \le K_0 \int_0^1 (\epsilon^{-1}+1)^{1/2}\, \mathrm{d}\epsilon < \infty,$$

which concludes the proof of the lemma.

A.5. Proof of Lemma 6.1

Proof of claims (1) *and* (2). We first prove claims (1) and (2) via induction over m. For the base case $m = 1$, the claims hold by the law of large numbers and the assumption that σ is not a constant function. Suppose the claims hold for $1 \le m \le m_0$, then we prove it also holds for $m = m_0 + 1$ by induction. Conditioning on $\boldsymbol{h}_{m_0}$, notice that

$$\boldsymbol{g}_{m_0+1} \overset{d}{=} \sqrt{\|\boldsymbol{h}_{m_0}\|_2^2/d_{m_0}}\boldsymbol{z},$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m_0+1}})$ and is independent of $\boldsymbol{h}_{m_0}$. Therefore,

$$\frac{1}{d_{m_0+1}} \|\boldsymbol{g}_{m_0+1}\|_2^2 \stackrel{d}{=} \frac{1}{d_{m_0}} \|\boldsymbol{h}_{m_0}\|_2^2 \times \frac{1}{d_{m_0+1}} \|\boldsymbol{z}_{m_0+1}\|_2^2,$$

which converges to some positive deterministic constant by the law of large numbers and induction hypothesis. Similarly, conditioning on $\boldsymbol{h}_{m_0}$, we have

$$\frac{1}{d_{m_0+1}} \|\boldsymbol{h}_{m_0+1}\|_2^2 \stackrel{d}{=} \frac{1}{d_{m_0+1}} \sum_{i=1}^{d_{m_0+1}} \sigma\big((\|\boldsymbol{h}_{m_0}\|_2^2/d_{m_0})^{1/2} z_i\big)^2,$$

where z_i is the i-th entry of $\boldsymbol{z}$. By our induction hypothesis, $\|\boldsymbol{h}_{m_0}\|_2^2/d_{m_0}$ converges in probability to some constant $H_{m_0} > 0$. Since almost everywhere

$$|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$$

and σ is continuous, we can conclude that

$$\frac{1}{d_{m_0+1}} \sum_{i=1}^{d_{m_0+1}} \sigma\big((\|\boldsymbol{h}_{m_0}\|_2^2/d_{m_0})^{1/2} z_i\big)^2 = \frac{1}{d_{m_0+1}} \sum_{i=1}^{d_{m_0+1}} \sigma(H_{m_0}^{1/2} z_i)^2 + o_P(1),$$

which further converges in probability to some positive constant by the law of large numbers and the non-degeneracy assumption on σ. Thus, we have completed the proof of the first two claims by induction.

Proof of claims (3)*,* (4) *and* (5)*.* Then we prove claims (3), (4) and (5), again via induction. We start with the base case $m = l$. Conditioning on $\boldsymbol{h}_{l-1}$,

$$\|\boldsymbol{\eta}_l\|_2^2 \stackrel{d}{=} \frac{1}{d_l} \sum_{i=1}^{d_l} |z_i^{(l+1)}|^2 \sigma'\big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2} z_i^{(l)}\big)^2,$$

where $\boldsymbol{z}^{(l)}, \boldsymbol{z}^{(l+1)} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l})$ and are independent of $\boldsymbol{h}_{l-1}$. Recall that almost everywhere we have $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. By Chebyshev's inequality and claim (1) of Lemma 6.1,

$$\begin{aligned}
&\frac{1}{d_l} \sum_{i=1}^{d_l} |z_i^{(l+1)}|^2 \sigma'\big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2} z_i^{(l)}\big)^2 \\
&\quad = \mathbb{E}_{z \sim \mathsf{N}(0,1)}\big[\sigma'\big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2} z\big)^2\big] + o_P(1),
\end{aligned}$$

where the expectation on the right-hand side is taken over $z \sim \mathsf{N}(0, 1)$. By claim (2), there exists a constant $H_{l-1} > 0$, such that

$$\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1} \xrightarrow{\mathbb{P}} H_{l-1}.$$

Since σ' is almost everywhere continuous, and $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$, we can apply the dominated convergence theorem and conclude that as $d \to \infty$,

$$\mathbb{E}_{z\sim\mathsf{N}(0,1)}\big[\sigma'\big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2}z\big)^2\big] \xrightarrow{\mathbb{P}} \mathbb{E}_{z\sim\mathsf{N}(0,1)}\big[\sigma'(H_{l-1}^{1/2}z)^2\big].$$

In summary, we have

$$\|\boldsymbol{\eta}_l\|_2^2 \xrightarrow{\mathbb{P}} \mathbb{E}_{z\sim\mathsf{N}(0,1)}\big[\sigma'(H_{l-1}^{1/2}z)^2\big],$$

thus completing the proof of claim (3) for the base case.

Then we consider $\boldsymbol{h}_{l-1}^\top \boldsymbol{W}_l^\top \boldsymbol{\eta}_l$ and the proof of claim (5). Since $\boldsymbol{W}_{l+1}$ is independent of $\boldsymbol{D}_\sigma^l \boldsymbol{W}_l \boldsymbol{h}_{l-1}$, conditioning on $\boldsymbol{D}_\sigma^l \boldsymbol{W}_l \boldsymbol{h}_{l-1}$, we have

$$\boldsymbol{h}_{l-1}^\top \boldsymbol{W}_l^\top \boldsymbol{\eta}_l = \langle \boldsymbol{D}_\sigma^l \boldsymbol{W}_l \boldsymbol{h}_{l-1}, \boldsymbol{W}_{l+1}^\top \rangle \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{D}_\sigma^l \boldsymbol{W}_l \boldsymbol{h}_{l-1}\|_2^2/d_l\big).$$

Conditioning on $\boldsymbol{h}_{l-1}$, we have

$$\frac{1}{d_l}\|\boldsymbol{D}_\sigma^l \boldsymbol{W}_l \boldsymbol{h}_{l-1}\|_2^2 \stackrel{d}{=} \frac{1}{d_l}\sum_{i=1}^{d_l} \sigma'\big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2} z_i^{(l)}\big)^2 \big((\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1})^{1/2} z_i^{(l)}\big)^2,$$

which converges in probability to $\mathbb{E}_{z\sim\mathsf{N}(0,1)}[H_{l-1}\sigma'(H_{l-1}^{1/2}z)^2 z^2]$ as $d \to \infty$, again by Chebyshev's inequality and the dominated convergence theorem. Therefore, as $d \to \infty$,

$$\boldsymbol{h}_{l-1}^\top \boldsymbol{W}_l^\top \boldsymbol{\eta}_l \xrightarrow{d} \mathsf{N}\big(0, \mathbb{E}_{z\sim\mathsf{N}(0,1)}\big[H_{l-1}\sigma'(H_{l-1}^{1/2}z)^2 z^2\big]\big),$$

which implies that $\boldsymbol{h}_{l-1}^\top \boldsymbol{W}_l^\top \boldsymbol{\eta}_l = O_P(1)$. We have completed the proof of claim (5) for the base case.

As the last step towards proving the base case, we consider the Euclidean norm of $\boldsymbol{y}_l$, i.e., claim (4). Notice that $\boldsymbol{\eta}_l$ depends on $\boldsymbol{W}_l$ only through $\boldsymbol{W}_l \boldsymbol{h}_{l-1}$. Therefore, conditioning on $\boldsymbol{W}_l \boldsymbol{h}_{l-1}$, we have

$$\begin{aligned}
\boldsymbol{y}_l &= \boldsymbol{W}_l^\top \boldsymbol{\eta}_l \\
&\stackrel{d}{=} \Pi_{\boldsymbol{h}_{l-1}}^\perp \widetilde{\boldsymbol{W}}_l^\top \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\top + \frac{\boldsymbol{h}_{l-1}}{\|\boldsymbol{h}_{l-1}\|_2^2}\langle \boldsymbol{W}_l \boldsymbol{h}_{l-1}, \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\top \rangle \\
&= \widetilde{\boldsymbol{W}}_l^\top \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\top + \frac{\boldsymbol{h}_{l-1}}{\|\boldsymbol{h}_{l-1}\|_2^2}\big(\langle \boldsymbol{W}_l \boldsymbol{h}_{l-1}, \boldsymbol{\eta}_l \rangle - \langle \widetilde{\boldsymbol{W}}_l \boldsymbol{h}_{l-1}, \boldsymbol{\eta}_l \rangle\big),
\end{aligned}$$

where $\widetilde{\boldsymbol{W}}_l$ is an independent copy of $\boldsymbol{W}_l$ and is independent of everything else. By claim (2) of the lemma, the vector $\boldsymbol{h}_{l-1}/\|\boldsymbol{h}_{l-1}\|_2^2$ has Euclidean norm $O_P(d_{l-1}^{-1/2})$. Conditioning on $(\boldsymbol{h}_{l-1}, \boldsymbol{\eta}_l)$, we have

$$\langle \widetilde{\boldsymbol{W}}_l \boldsymbol{h}_{l-1}, \boldsymbol{\eta}_l \rangle \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_l\|_2^2 \|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1}\big),$$

which is $O_P(1)$ as $\|\boldsymbol{\eta}_l\|_2^2$, $\|\boldsymbol{h}_{l-1}\|_2^2/d_{l-1}$ are both $O_P(1)$ by claim (1) and claim (3). As proven above, we have

$$\langle \boldsymbol{W}_l \boldsymbol{h}_{l-1}, \boldsymbol{\eta}_l \rangle = O_P(1).$$

Finally, conditioning on $\boldsymbol{\eta}_l = \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\mathsf{T}$, we have

$$\|\widetilde{\boldsymbol{W}}_l^\mathsf{T} \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\mathsf{T}\|_2^2 \stackrel{d}{=} \sum_{i=1}^{d_{l-1}} z_i^2 \|\boldsymbol{\eta}_l\|_2^2/d_{l-1},$$

where $\{z_i\}_{i\in[d_{l-1}]} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$. By the law of large numbers, $\|\widetilde{\boldsymbol{W}}_l^\mathsf{T} \boldsymbol{D}_\sigma^l \boldsymbol{W}_{l+1}^\mathsf{T}\|_2^2$ converges in probability to the same limit of $\|\boldsymbol{\eta}_l\|_2^2$ as $d \to \infty$. Combining the results above, we can conclude that $\|\boldsymbol{y}_l\|_2^2$ converges in probability to a positive constant as $d \to \infty$, thus concluding the proof of claim (4) of the base case.

Suppose claims (3) to (5) hold for all $m_0 + 1 \le m \le l$, then we prove that they also hold for $m = m_0$. First notice that $(\boldsymbol{\eta}_{m_0+1}, \boldsymbol{D}_\sigma^{m_0})$ depends on $\boldsymbol{W}_{m_0+1}$ only through $\boldsymbol{W}_{m_0+1}\boldsymbol{h}_{m_0}$, thus conditioning on $(\boldsymbol{D}_\sigma^{m_0}, \boldsymbol{\eta}_{m_0+1})$, we have

$$\begin{aligned} \boldsymbol{\eta}_{m_0} &= \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1} \\ &\stackrel{d}{=} \boldsymbol{D}_\sigma^{m_0} \widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1} + \frac{\boldsymbol{h}_{m_0}^\mathsf{T} \boldsymbol{W}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1} - \boldsymbol{h}_{m_0}^\mathsf{T} \widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1}}{\|\boldsymbol{h}_{m_0}\|_2^2} \boldsymbol{D}_\sigma^{m_0} \boldsymbol{h}_{m_0}, \end{aligned} \tag{A.1}$$

where $\widetilde{\boldsymbol{W}}_{m_0+1}$ has the same marginal distribution as $\boldsymbol{W}_{m_0+1}$, and is independent of everything else. By claim (1), $\|\boldsymbol{h}_{m_0}\|_2^{-2} = O_P(d_{m_0}^{-1})$. By induction hypothesis,

$$\boldsymbol{h}_{m_0}^\mathsf{T} \boldsymbol{W}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1} = O_P(1).$$

Conditioning on $(\boldsymbol{h}_{m_0}, \boldsymbol{\eta}_{m_0+1})$, we have

$$\boldsymbol{h}_{m_0}^\mathsf{T} \widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T} \boldsymbol{\eta}_{m_0+1} \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{h}_{m_0}\|_2^2 \|\boldsymbol{\eta}_{m_0+1}\|_2^2/d_{m_0}\big),$$

which is $O_P(1)$ by induction hypothesis and claim (1). Conditioning on $\boldsymbol{h}_{m_0-1}$,

$$\begin{aligned} \frac{1}{d_{m_0}} \|\boldsymbol{D}_\sigma^{m_0} \boldsymbol{h}_{m_0}\|_2^2 \stackrel{d}{=} \frac{1}{d_{m_0}} \sum_{i=1}^{d_{m_0}} &\sigma'\big((\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1})^{1/2} z_i^{(m_0)}\big)^2 \\ &\times \sigma\big((\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1})^{1/2} z_i^{(m_0)}\big)^2, \end{aligned}$$

where $\{z_i^{(m_0)}\}_{i\in[d_{m_0}]} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$ and are independent of $\boldsymbol{h}_{m_0-1}$. By claim (2) of the lemma, there exists a constant $H_{m_0-1} > 0$, such that

$$\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1} \stackrel{\mathbb{P}}{\to} H_{m_0-1}.$$

By the assumption that σ' is almost everywhere continuous, and

$$|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1}),$$

we can apply Chebyshev's inequality and dominated convergence theorem and conclude that $\|\boldsymbol{D}_\sigma^{m_0} \boldsymbol{h}_{m_0}\|_2^2 / d_{m_0}$ converges in probability to some constant. In summary, the second term in equation (A.1) has Euclidean norm $O_P(d_{m_0}^{-1/2})$.

Then we consider the first term in equation (A.1). Notice that conditioning on $\boldsymbol{h}_{m_0-1}$, we have

$$\|\boldsymbol{D}_\sigma^{m_0} \widetilde{\boldsymbol{W}}_{m_0+1}^\top \boldsymbol{\eta}_{m_0+1}\|_2^2 \overset{d}{=} \frac{1}{d_{m_0}} \sum_{i=1}^{d_{m_0}} \sigma'\big((\|\boldsymbol{h}_{m_0-1}\|_2^2 / d_{m_0-1})^{1/2} z_i^{(m_0)}\big)^2 z_i^2 \|\boldsymbol{\eta}_{m_0+1}\|_2^2,$$

where $\boldsymbol{z}, \boldsymbol{z}^{(m_0)} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m_0}})$ and are independent of $\boldsymbol{h}_{m_0-1}$. Again by Chebyshev's inequality and dominated convergence theorem, we have

$$\frac{1}{d_{m_0}} \sum_{i=1}^{d_{m_0}} \sigma'\big((\|\boldsymbol{h}_{m_0-1}\|_2^2 / d_{m_0-1})^{1/2} z_i^{(m_0)}\big)^2 z_i^2 \overset{\mathbb{P}}{\to} \mathbb{E}_{z \sim \mathsf{N}(0,1)}\big[\sigma'(H_{m_0-1}^{1/2} z)^2\big].$$

By induction hypothesis, $\|\boldsymbol{\eta}_{m_0+1}\|_2^2$ converges in probability to a positive constant. Therefore, we conclude that $\|\boldsymbol{D}_\sigma^{m_0} \widetilde{\boldsymbol{W}}_{m_0+1}^\top \boldsymbol{\eta}_{m_0+1}\|_2^2$ converges in probability to a positive constant. Then we plug the above results into equation (A.1), and conclude that $\|\boldsymbol{\eta}_{m_0}\|_2^2$ converges in probability to a positive constant which depends only on (σ, l). Thus, we have completed the proof of claim (3) for $m = m_0$.

We next show that

$$\boldsymbol{h}_{m_0-1}^\top \boldsymbol{W}_{m_0}^\top \boldsymbol{\eta}_{m_0} = O_P(1).$$

Notice that $(\boldsymbol{\eta}_{m_0+1}, \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1})$ depends on $\boldsymbol{W}_{m_0+1}$ only through $\boldsymbol{W}_{m_0+1} \boldsymbol{h}_{m_0}$. Then conditioning on $(\boldsymbol{\eta}_{m_0+1}, \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1})$, we have

$$\boldsymbol{W}_{m_0+1} \overset{d}{=} \widetilde{\boldsymbol{W}}_{m_0+1} \Pi_{\boldsymbol{h}_{m_0}}^\perp + \boldsymbol{W}_{m_0+1} \Pi_{\boldsymbol{h}_{m_0}},$$

where $\widetilde{\boldsymbol{W}}_{m_0+1}$ is an independent copy of $\boldsymbol{W}_{m_0+1}$ and is independent of everything else. Therefore,

$$\begin{aligned}
\langle \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1}, \boldsymbol{\eta}_{m_0} \rangle &= \langle \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1}, \boldsymbol{W}_{m_0+1}^\top \boldsymbol{\eta}_{m_0+1} \rangle \\
&\overset{d}{=} \frac{\langle \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1}, \boldsymbol{h}_{m_0} \rangle}{\|\boldsymbol{h}_{m_0}\|_2^2} (\boldsymbol{h}_{m_0}^\top \boldsymbol{W}_{m_0+1}^\top \boldsymbol{\eta}_{m_0+1} - \boldsymbol{h}_{m_0}^\top \widetilde{\boldsymbol{W}}_{m_0+1} \boldsymbol{\eta}_{m_0+1}) \\
&\quad + \langle \boldsymbol{D}_\sigma^{m_0} \boldsymbol{W}_{m_0} \boldsymbol{h}_{m_0-1}, \widetilde{\boldsymbol{W}}_{m_0+1}^\top \boldsymbol{\eta}_{m_0+1} \rangle. \qquad \text{(A.2)}
\end{aligned}$$

By claim (2) of the lemma, $\|\boldsymbol{h}_{m_0}\|_2^{-2} = O_P(d_{m_0}^{-1})$. Conditioning on $\boldsymbol{h}_{m_0-1}$, we have

$$\frac{1}{d_{m_0}}\langle \boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \boldsymbol{h}_{m_0}\rangle \stackrel{d}{=} \frac{1}{d_{m_0}}\sum_{i=1}^{d_{m_0}} \sigma'\big((\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1})^{1/2}z_i\big)^2$$
$$\times\, \sigma\big((\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1})^{1/2}z_i\big)^2 z_i^2 \|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1},$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m_0}})$ and is independent of $\boldsymbol{h}_{m_0-1}$. Again we apply Chebyshev's inequality and dominated convergence theorem, and conclude that

$$\frac{1}{d_{m_0}}\langle \boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \boldsymbol{h}_{m_0}\rangle \stackrel{\mathbb{P}}{\to} H_{m_0-1}\mathbb{E}_{z\sim\mathsf{N}(0,1)}\big[\sigma'(H_{m_0-1}^{1/2}z)^2\sigma(H_{m_0-1}^{1/2}z)^2z^2\big]$$

as $d \to \infty$. By induction hypothesis, $\boldsymbol{h}_{m_0}^\mathsf{T}\boldsymbol{W}_{m_0+1}^\mathsf{T}\boldsymbol{\eta}_{m_0+1} = O_P(1)$. Conditioning on $(\boldsymbol{h}_{m_0}, \boldsymbol{\eta}_{m_0+1})$, we have

$$\boldsymbol{h}_{m_0}^\mathsf{T}\widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T}\boldsymbol{\eta}_{m_0+1} \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{h}_{m_0}\|_2^2\|\boldsymbol{\eta}_{m_0+1}\|_2^2/d_{m_0}\big),$$

which is $O_P(1)$ since by claim (1) of the lemma we have $\|\boldsymbol{h}_{m_0}\|_2^2/d_{m_0} = O_P(1)$, and by induction hypothesis we have $\|\boldsymbol{\eta}_{m_0+1}\|_2^2 = O_P(1)$. Combining the above analysis, we can conclude that the first summand in equation (A.2) is $O_P(1)$.

Then we consider the second summand in equation (A.2).

Conditioning on $(\boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \boldsymbol{\eta}_{m_0+1})$, we have

$$\langle \boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T}\boldsymbol{\eta}_{m_0+1}\rangle \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_{m_0+1}\|_2^2\|\boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0}\big).$$

Again we apply the conditioning technique. Conditioning on $\boldsymbol{h}_{m_0-1}$, we have

$$\frac{1}{d_{m_0}}\|\boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}\|_2^2$$
$$\stackrel{d}{=} \frac{1}{d_{m_0}}\sum_{i=1}^{d_{m_0}}\sigma'\big((\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1})^{1/2}z_i\big)^2\big(\|\boldsymbol{h}_{m_0-1}\|_2^2/d_{m_0-1}\big)z_i^2 = O_P(1),$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m_0}})$ and is independent of $\boldsymbol{h}_{m_0-1}$. By induction hypothesis,

$$\|\boldsymbol{\eta}_{m_0+1}\|_2^2 = O_P(1),$$

thus $\langle \boldsymbol{D}_\sigma^{m_0}\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \widetilde{\boldsymbol{W}}_{m_0+1}^\mathsf{T}\boldsymbol{\eta}_{m_0+1}\rangle = O_P(1)$. Next, we plug the above analysis into equation (A.2) and conclude that

$$\langle \boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}, \boldsymbol{\eta}_{m_0}\rangle = O_P(1),$$

thus proving claim (5) for $m = m_0$.

As the last step of our induction proof, we show that $\|\boldsymbol{y}_{m_0}\|_2^2$ converges in probability to some positive constant.

Note that $\boldsymbol{\eta}_{m_0}$ depends on $\boldsymbol{W}_{m_0}$ only through $\boldsymbol{W}_{m_0}\boldsymbol{h}_{m_0-1}$. As a result, conditioning on $\boldsymbol{\eta}_{m_0}$, we have

$$\boldsymbol{y}_{m_0} = \boldsymbol{W}_{m_0}^\top \boldsymbol{\eta}_{m_0} \stackrel{d}{=} \widetilde{\boldsymbol{W}}_{m_0}^\top \boldsymbol{\eta}_{m_0} + \frac{\boldsymbol{h}_{m_0-1}^\top \boldsymbol{W}_{m_0}^\top \boldsymbol{\eta}_{m_0} - \boldsymbol{h}_{m_0-1}^\top \widetilde{\boldsymbol{W}}_{m_0}^\top \boldsymbol{\eta}_{m_0}}{\|\boldsymbol{h}_{m_0-1}\|_2^2} \boldsymbol{h}_{m_0-1},$$

where $\widetilde{\boldsymbol{W}}_{m_0}$ is an independent copy of $\boldsymbol{W}_{m_0}$ and is independent of everything else. By induction hypothesis, $\|\boldsymbol{\eta}_{m_0}\|_2^2$ converges in probability to some positive constant. By the law of large numbers, we have $\|\widetilde{\boldsymbol{W}}_{m_0}^\top \boldsymbol{\eta}_{m_0}\|_2^2$ converges in probability to the same limit of $\|\boldsymbol{\eta}_{m_0}\|_2^2$ as $d \to \infty$. By claim (2) of the lemma,

$$\|\boldsymbol{h}_{m_0-1}\|_2^{-2} = O_P(d_{m_0-1}^{-1}) \quad \text{and} \quad \|\boldsymbol{h}_{m_0-1}\|_2 = O_P(d_{m_0-1}^{1/2}).$$

Note that we have proved $\boldsymbol{h}_{m_0-1}^\top \boldsymbol{W}_{m_0}^\top \boldsymbol{\eta}_{m_0} = O_P(1)$. Conditioning on $(\boldsymbol{h}_{m_0-1}, \boldsymbol{\eta}_{m_0})$, we have

$$\boldsymbol{h}_{m_0-1}^\top \widetilde{\boldsymbol{W}}_{m_0}^\top \boldsymbol{\eta}_{m_0} \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{h}_{m_0-1}\|_2^2 \|\boldsymbol{\eta}_{m_0}\|_2^2 / d_{m_0-1}\big),$$

which is $O_P(1)$ by claim (1) and induction hypothesis. In summary, we can conclude that $\|\boldsymbol{y}_{m_0}\|_2^2$ and $\|\boldsymbol{\eta}_m\|_2^2$ converges in probability to the same limit, thus completing the proof of the lemma by induction.

A.6. Proof of Lemma 6.2

We first provide finite sample upper bounds on the Euclidean norms of

$$\{\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{D}_\sigma^i \boldsymbol{g}_i\}_{i\in[l]}.$$

Lemma A.2. *Under the conditions of Theorem 5.1, there exist positive constants $\{Q_i\}_{1\le i\le l}$, which depend only on (σ, l), such that for all $1 \le i \le l$, with probability at least $1-\delta$, we have*

$$\frac{1}{\sqrt{d_i}}\|\boldsymbol{h}_i\|_2 \le Q_i \prod_{j=1}^{i} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{i-j}}, \tag{A.3}$$

$$\frac{1}{\sqrt{d_i}}\|\boldsymbol{g}_i\|_2 \le Q_i \big(1 + \log(1/\delta) d_i^{-1/2}\big) \prod_{j=1}^{i-1} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{i-1-j}}, \tag{A.4}$$

$$\frac{1}{\sqrt{d_i}} \frac{\|\boldsymbol{D}_\sigma^i \boldsymbol{g}_i\|_2}{\|\boldsymbol{h}_{i-1}\|_2} \le \frac{Q_i}{\sqrt{d_{i-1}}} \prod_{j=1}^{i} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{i-j}}. \tag{A.5}$$

Proof. We prove this lemma by induction over i. For the base case $i = 1$, $\boldsymbol{g}_1 \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_{d_1})$, thus, equation (A.4) follows from Bernstein's inequality. Since almost everywhere we have $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$, then there exists $C'_\sigma > 0$ which is a function of σ only, such that almost everywhere

$$|\sigma(x)| \le C'_\sigma(1 + |x|^k) \quad \text{and} \quad |\sigma'(x)x| \le C'_\sigma(1 + |x|^k).$$

Then with probability 1, we have

$$\frac{1}{d_1}\|\boldsymbol{h}_1\|_2^2 \le \frac{2}{d_1}\sum_{i=1}^{d_1}(C'_\sigma)^2(1 + |g_{1,i}|^{2k}),$$

$$\frac{1}{d_1\|\boldsymbol{x}\|_2^2}\|\boldsymbol{D}_\sigma^1\boldsymbol{g}_1\|_2^2 \le \frac{2}{d_1 d}\sum_{i=1}^{d_1}(C'_\sigma)^2(1 + |g_{1,i}|^{2k}),$$

thus, equations (A.3) and (A.5) for the base case follow from Chebyshev's inequality. Now suppose the lemma holds for all $i \le m$, then we prove it also holds for $i = m + 1$ by induction. Conditioning on $\boldsymbol{h}_m$, we have

$$\boldsymbol{g}_{m+1} \overset{d}{=} \mathsf{N}\big(0, (\|\boldsymbol{h}_m\|_2^2/d_m)\boldsymbol{I}_{d_{m+1}}\big).$$

Thus, by Bernstein's inequality, there exists an absolute constant $C > 0$, such that with probability at least $1 - \delta/3$,

$$\|\boldsymbol{g}_{m+1}\|_2^2/d_{m+1} \le C\|\boldsymbol{h}_m\|_2^2\big(1 + \log(1/\delta)/\sqrt{d_{m+1}}\big)/d_m.$$

By induction hypothesis, with probability at least $1 - \delta/3$,

$$\frac{1}{\sqrt{d_m}}\|\boldsymbol{h}_m\|_2 \le 3^{mk^m} Q_m \prod_{j=1}^{m}(1 + \delta^{-1/2}d_j^{-1/2})^{k^{m-j}},$$

thus, equation (A.4) for $i = m + 1$ follows. To prove equations (A.3) and (A.5), note that

$$\begin{aligned}\frac{1}{d_{m+1}}\|\boldsymbol{h}_{m+1}\|_2^2 &\le \frac{1}{d_{m+1}}\sum_{i=1}^{d_{m+1}}(C'_\sigma)^2\big(1 + |g_{m+1,i}|^k\big)^2\\ &\overset{d}{=} \frac{1}{d_{m+1}}\sum_{i=1}^{d_{m+1}} 2(C'_\sigma)^2\big(1 + |z_i|^{2k}\|\boldsymbol{h}_m\|_2^{2k}/d_m^k\big),\end{aligned}$$

$$\frac{1}{d_{m+1}}\frac{\|\boldsymbol{D}_\sigma^{m+1}\boldsymbol{g}_{m+1}\|_2^2}{\|\boldsymbol{h}_m\|_2^2} \le \frac{1}{d_{m+1}\|\boldsymbol{h}_m\|_2^2}\sum_{i=1}^{d_{m+1}}\sigma'(g_{m+1,i})^2 g_{m+1,i}^2$$

$$\stackrel{d}{=} \frac{1}{d_{m+1}d_m} \sum_{i=1}^{d_{m+1}} \sigma'\big(z_i \|\boldsymbol{h}_m\|_2/\sqrt{d_m}\big)^2 z_i^2$$
$$\leq \frac{1}{d_{m+1}d_m} \sum_{i=1}^{d_{m+1}} 4(C_\sigma)^2 z_i^2 \big(1 + z_i^{2k-2}\|\boldsymbol{h}_m\|_2^{2k-2}/d_m^{k-1}\big),$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m+1}})$, and is independent $\boldsymbol{h}_m$. Thus, equations (A.3) and (A.5) for $i = m+1$ follows from Chebyshev's inequality, and we complete the proof of the lemma by induction. ∎

Next, we control the Euclidean norms of $\{\boldsymbol{\eta}_i\}_{i\in[l]}$.

Lemma A.3. *Under the conditions of Theorem 5.1, there exist constants $\{\widetilde{Q}_i\}_{1\leq i\leq l}$, which depend only on (σ, l), such that for all $1 \leq i \leq l$, with probability at least $1-\delta$, we have*

$$\|\boldsymbol{\eta}_i\|_2 \leq \widetilde{Q}_i\big(\sqrt{\log(1/\delta)}+1\big)^{l-i} \prod_{m=i}^{l}\prod_{j=1}^{m}(1+\delta^{-1/2}d_j^{-1/2})^{k^{m-j}}, \tag{A.6}$$

$$\frac{|\boldsymbol{h}_{i-1}^\top \boldsymbol{W}_i^\top \boldsymbol{\eta}_i|}{\|\boldsymbol{h}_{i-1}\|_2} \leq \frac{\widetilde{Q}_i\big(\sqrt{\log(1/\delta)}+1\big)^{l-i}}{\sqrt{d_{i-1}}} \prod_{m=i}^{l}\prod_{j=1}^{m}(1+\delta^{-1/2}d_j^{-1/2})^{k^{m-j}}. \tag{A.7}$$

Proof. We prove this lemma by induction over i. We start with the base case $i = l$. Conditioning on $\boldsymbol{h}_{l-1}$, we have

$$\|\boldsymbol{\eta}_l\|_2^2 \stackrel{d}{=} \frac{1}{d_l}\sum_{i=1}^{d_l} z_{l+1,i}^2 \sigma'\big(\|\boldsymbol{h}_{l-1}\|_2/\sqrt{d_{l-1}} z_{l,i}\big)^2$$
$$\leq \frac{8C_\sigma^2}{d_l}\sum_{i=1}^{d_l} z_{l+1,i}^2 + \frac{8C_\sigma^2}{d_l}\sum_{i=1}^{d_l} z_{l+1,i}^2 z_{l,i}^{2k-2}\big(\|\boldsymbol{h}_{l-1}\|_2/\sqrt{d_{l-1}}\big)^{2k-2},$$

where $\boldsymbol{z}_l, \boldsymbol{z}_{l+1} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l})$ and is independent of $\boldsymbol{h}_{l-1}$. Then equation (A.6) for $i = l$ follows from Lemma A.2 and Chebyshev's inequality. Note that $\boldsymbol{W}_{l+1}$ is independent of $(\boldsymbol{D}_\sigma^l, \boldsymbol{g}_l)$, then

$$\frac{|\boldsymbol{h}_{l-1}^\top \boldsymbol{W}_l^\top \boldsymbol{\eta}_l|^2}{\|\boldsymbol{h}_{l-1}\|_2^2} \stackrel{d}{=} \frac{1}{d_l d_{l-1}}\sum_{i=1}^{d_l} z_{l+1,i}^2 z_{l,i}^2 \sigma'\big(z_{l,i}\|\boldsymbol{h}_{l-1}\|_2/\sqrt{d_{l-1}}\big)^2$$
$$\leq \frac{8C_\sigma^2}{d_l d_{l-1}}\sum_{i=1}^{d_l} z_{l+1,i}^2 z_{l,i}^2\big(1 + z_{l,i}^{2k-2}\|\boldsymbol{h}_{l-1}\|_2^{2k-2} d_{l-1}^{-k+1}\big).$$

As a result, equation (A.7) for $i = l$ follows from Lemma A.2 and Chebyshev's inequality. Thus, we have completed the proof for the base case.

Suppose the lemma is true for all $m+1 \le i \le l$, then we prove it also holds for $i = m$ via induction hypothesis. By equation (A.1),

$$\boldsymbol{\eta}_m \stackrel{d}{=} \boldsymbol{D}_\sigma^m \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} + \frac{\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1} - \boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}}{\|\boldsymbol{h}_m\|_2^2} \boldsymbol{D}_\sigma^m \boldsymbol{h}_m, \tag{A.8}$$

where $\widetilde{\boldsymbol{W}}_{m+1}$ has the same distribution as $\boldsymbol{W}_{m+1}$, and is independent of everything else. The right-hand side of equation (A.8) has Euclidean norm no larger than

$$\|\boldsymbol{D}_\sigma^m \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}\|_2 + \frac{|\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1}| + |\boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}|}{\|\boldsymbol{h}_m\|_2^2} \|\boldsymbol{D}_\sigma^m \boldsymbol{h}_m\|_2.$$

Note that

$$\|\boldsymbol{D}_\sigma^m \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}\|_2^2 \stackrel{d}{=} \frac{1}{d_m} \sum_{i=1}^{d_m} \sigma'(z_{m,i} \|\boldsymbol{h}_{m-1}\|_2 / \sqrt{d_{m-1}})^2 z_{m+1,i}^2 \|\boldsymbol{\eta}_{m+1}\|_2^2,$$

where $\boldsymbol{z}_m, \boldsymbol{z}_{m+1} \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m})$, and are independent of $\boldsymbol{h}_{m-1}$. Using the fact that almost everywhere $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$, together with Lemma A.2 and Chebyshev's inequality, we conclude that there exists a constant $\widetilde{Q}_m^{(1)} > 0$, depending only on (σ, l), such that with probability at least $1 - \delta/6$,

$$\|\boldsymbol{D}_\sigma^m \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}\|_2 \le \widetilde{Q}_m^{(1)} \prod_{j=1}^{m} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}} \|\boldsymbol{\eta}_{m+1}\|_2.$$

Conditioning on $(\boldsymbol{h}_m, \boldsymbol{\eta}_{m+1})$, we have

$$\boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} / \|\boldsymbol{h}_m\|_2 \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_{m+1}\|_2^2 / d_m\big).$$

Therefore, with probability at least $1 - \delta/6$, we have

$$\frac{|\boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1}|}{\|\boldsymbol{h}_m\|_2} \le d_m^{-1/2} \widetilde{Q}_m^{(2)} \sqrt{\log(1/\delta)} \|\boldsymbol{\eta}_{m+1}\|_2,$$

where $\widetilde{Q}_m^{(2)} > 0$ is a numerical constant.

Conditioning on $\boldsymbol{h}_{m-1}$, we have

$$|\boldsymbol{D}_\sigma^m \boldsymbol{h}_m\|_2^2 / \|\boldsymbol{h}_m\|_2^2 \le \|\boldsymbol{D}_\sigma^m\|_F^2 \stackrel{d}{=} \sum_{i=1}^{d_m} \sigma'(\|\boldsymbol{h}_{m-1}\|_2 / \sqrt{d_{m-1}} z_{m,i})^2,$$

where $\boldsymbol{z}_m \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_m})$ and is independent of $\boldsymbol{h}_{m-1}$. Since

$$\frac{1}{d_m} \sum_{i=1}^{d_m} \sigma'\big(\|\boldsymbol{h}_{m-1}\|_2 / \sqrt{d_{m-1}} z_{m,i}\big)^2 \le \frac{8 C_\sigma^2}{d_m} \sum_{i=1}^{d_m} \big(1 + |z_{m,i}|^{2k-2} \|\boldsymbol{h}_{m-1}\|_2^{2k-2} d_{m-1}^{-k+1}\big),$$

by Lemma A.2 and Chebyshev's inequality, we conclude that there exists $\widetilde{Q}_m^{(3)} > 0$, depending only on (σ, l), such that with probability at least $1 - \delta/6$,

$$\frac{\|\boldsymbol{D}_\sigma^m \boldsymbol{h}_m\|_2}{\|\boldsymbol{h}_m\|_2} \leq \sqrt{d_m}\,\widetilde{Q}_m^{(3)} \prod_{j=1}^{m} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}}. \tag{A.9}$$

Combining equations (A.8)–(A.9), we conclude that there exists $\widetilde{Q}_m^{(4)} > 0$, depending only on (σ, l), such that with probability at least $1 - \delta/2$,

$$\begin{aligned} \|\boldsymbol{\eta}_m\|_2 \leq \widetilde{Q}_m^{(4)} \big(\sqrt{\log(1/\delta)} + 1\big) \times \left\{ \prod_{j=1}^{m} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}} \right\} \\ \times \left\{ \|\boldsymbol{\eta}_{m+1}\|_2 + \frac{\sqrt{d_m}\|\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1}\|}{\|\boldsymbol{h}_m\|_2} \right\}, \end{aligned} \tag{A.10}$$

thus equation (A.6) for $i = m$ follows from equation (A.10) and induction hypothesis. Then we proceed to prove equation (A.7) for $i = m$. By equation (A.2), there exists $\widetilde{\boldsymbol{W}}_{m+1}$ that has the same marginal distribution as $\boldsymbol{W}_{m+1}$ and is independent of everything else, such that

$$\begin{aligned} \frac{|\boldsymbol{h}_{m-1}^\top \boldsymbol{W}_m^\top \boldsymbol{\eta}_m|}{\|\boldsymbol{h}_{m-1}\|_2} &\overset{d}{=} \Bigg| \frac{\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \boldsymbol{h}_m \rangle}{\|\boldsymbol{h}_m\|_2^2 \|\boldsymbol{h}_{m-1}\|_2} (\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1} - \boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1} \boldsymbol{\eta}_{m+1}) \\ &\quad + \frac{\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} \rangle}{\|\boldsymbol{h}_{m-1}\|_2} \Bigg| \\ &\leq \frac{|\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \boldsymbol{h}_m \rangle|}{\|\boldsymbol{h}_m\|_2^2 \|\boldsymbol{h}_{m-1}\|_2} |\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1} - \boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1} \boldsymbol{\eta}_{m+1}| \\ &\quad + \frac{|\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} \rangle|}{\|\boldsymbol{h}_{m-1}\|_2}. \end{aligned} \tag{A.11}$$

By Lemma A.2, with probability at least $1 - \delta/12$,

$$\frac{\|\boldsymbol{D}_\sigma^m \boldsymbol{g}_m\|_2}{\|\boldsymbol{h}_{m-1}\|_2} \leq \frac{6^{mk^m} Q_m \sqrt{d_m}}{\sqrt{d_{m-1}}} \prod_{j=1}^{m} (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}}. \tag{A.12}$$

Conditioning on $(\boldsymbol{h}_m, \boldsymbol{\eta}_{m+1})$, we have

$$\boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1} \boldsymbol{\eta}_{m+1} / \|\boldsymbol{h}_m\|_2 \overset{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_{m+1}\|_2^2 / d_m\big).$$

Therefore, with probability at least $1 - \delta/6$, we have

$$\frac{|\boldsymbol{h}_m^\top \widetilde{\boldsymbol{W}}_{m+1} \boldsymbol{\eta}_{m+1}|}{\|\boldsymbol{h}_m\|_2} \leq \frac{\widetilde{Q}_m^{(5)} \sqrt{\log(1/\delta)} \|\boldsymbol{\eta}_{m+1}\|_2}{\sqrt{d_m}},$$

where $\widetilde{Q}_m^{(5)} > 0$ is a numerical constant.

Given $(\boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \boldsymbol{\eta}_{m+1})$, the conditional distribution of

$$\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} \rangle / \|\boldsymbol{h}_{m-1}\|_2$$

is $\mathsf{N}(0, d_m^{-1} \|\boldsymbol{D}_\sigma^m \boldsymbol{g}_m\|_2^2 \|\boldsymbol{\eta}_{m+1}\|_2^2 / \|\boldsymbol{h}_{m-1}\|_2^2)$. Therefore, using equation (A.12), with probability at least $1 - \delta/6$,

$$\frac{|\langle \boldsymbol{D}_\sigma^m \boldsymbol{W}_m \boldsymbol{h}_{m-1}, \widetilde{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+1} \rangle|}{\|\boldsymbol{h}_{m-1}\|_2} \le \frac{\widetilde{Q}_m^{(6)} \|\boldsymbol{\eta}_{m+1}\|_2}{\sqrt{d_{m-1}}} \times \left\{ \prod_{j=1}^m (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}} \right\} \times \sqrt{\log(1/\delta)}, \tag{A.13}$$

where $\widetilde{Q}_m^{(6)} > 0$ is a constant depending only on (σ, l). Using equations (A.11)–(A.13) and induction hypothesis, we conclude that with probability at least $1 - \delta/2$, the last line in equation (A.11) is no larger than

$$\frac{\widetilde{Q}_m^{(7)} (\sqrt{\log(1/\delta)} + 1)}{\sqrt{d_{m-1}}} \times \left\{ \prod_{j=1}^m (1 + \delta^{-1/2} d_j^{-1/2})^{k^{m-j}} \right\} \times \left\{ \|\boldsymbol{\eta}_{m+1}\|_2 + \frac{\sqrt{d_m} \|\boldsymbol{h}_m^\top \boldsymbol{W}_{m+1}^\top \boldsymbol{\eta}_{m+1}\|}{\|\boldsymbol{h}_m\|_2} \right\}, \tag{A.14}$$

where $\widetilde{Q}_m^{(7)} > 0$ is a constant depending only on (σ, l). Thus, equation (A.7) for $i = m$ follows from equation (A.14) and induction hypothesis. Therefore, we have completed the proof of the lemma by induction. ■

Finally, with Lemmas A.2 and A.3, we are ready to prove Lemma 6.2. Note that $\boldsymbol{\eta}_1$ depends on $\boldsymbol{W}_1$ only through $\boldsymbol{W}_1 \boldsymbol{x}$, thus by property of Gaussian distribution, we have

$$\nabla f(\boldsymbol{x}) = \boldsymbol{W}_1^\top \boldsymbol{\eta}_1 \stackrel{d}{=} \frac{\boldsymbol{x}^\top \boldsymbol{W}_1^\top \boldsymbol{\eta}_1 - \boldsymbol{x}^\top \widetilde{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1}{\|\boldsymbol{x}\|_2^2} \boldsymbol{x} + \widetilde{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1,$$

where $\widetilde{\boldsymbol{W}}_1$ is an independent copy of $\boldsymbol{W}_1$, and is independent of everything else. We let $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, independent of everything else, then we have

$$\|\widetilde{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1\|_2 \stackrel{d}{=} \frac{\|\boldsymbol{\eta}_1\|_2 \|\boldsymbol{z}\|_2}{\sqrt{d}},$$

$$\frac{|\boldsymbol{x}^\top \boldsymbol{W}_1^\top \boldsymbol{\eta}_1|}{\|\boldsymbol{x}\|_2} = \frac{|\boldsymbol{h}_0^\top \boldsymbol{W}_1^\top \boldsymbol{\eta}_1|}{\|\boldsymbol{h}_0\|_2},$$

$$\frac{|\boldsymbol{x}^\top \widetilde{\boldsymbol{W}}_1^\top \boldsymbol{\eta}_1|}{\|\boldsymbol{x}\|_2} \stackrel{d}{=} \mathsf{N}\big(0, \|\boldsymbol{\eta}_1\|_2^2 / d\big).$$

Therefore, by Lemma A.3 and Bernstein's inequality, with probability at least $1-\delta$, we have

$$\|\nabla f(\boldsymbol{x})\|_2 \le Q\big(\sqrt{\log(1/\delta)}+1\big)^{l-1}\big(1+\log(1/\delta)d^{-1/2}\big) \times \prod_{i=1}^{l}\prod_{j=1}^{i}(1+\delta^{-1/2}d_j^{-1/2})^{k^{i-j}},$$

where $Q>0$ is a constant depending only on (σ, l), thus completing the proof of Lemma 6.2.

A.7. Proof of Lemma 6.3

By definition, $\mu_1 = \tau s_d \boldsymbol{x}^\top \boldsymbol{y}_1/d$, thus by the Cauchy–Schwarz inequality we have $|\mu_1| \le s_d\|\boldsymbol{y}_1\|_2/\sqrt{d}$. By Lemma 6.1, as $d\to\infty$, we have $\|\boldsymbol{y}_1\|_2 = O_P(1)$, thus $\mu_1 = o_P(1)$. Next, we consider γ_1. Since $\bar{\boldsymbol{W}}_1$ is independent of $\boldsymbol{\eta}_1$, by the law of large numbers, as $d\to\infty$, we have

$$\text{p-lim}_{d\to\infty}\|\bar{\boldsymbol{W}}_1^\top\boldsymbol{\eta}_1\|_2 = \text{p-lim}_{d\to\infty}\|\boldsymbol{\eta}_1\|_2,$$

which is finite and positive. As a result, $\gamma_1 = o_P(1)$. Finally, we consider β_1. Notice that

$$|\langle(\boldsymbol{W}_1')^\top\boldsymbol{\eta}_1, \bar{\boldsymbol{W}}_1^\top\boldsymbol{\eta}_1\rangle| \le \|(\boldsymbol{W}_1')^\top\boldsymbol{\eta}_1\|_2\|\bar{\boldsymbol{W}}_1^\top\boldsymbol{\eta}_1\|_2,$$

which further converges to $(\text{p-lim}\|\boldsymbol{\eta}_1\|_2)^2$ as $d\to\infty$. Similarly, we can show that $\|\bar{\boldsymbol{W}}_1\boldsymbol{\eta}_1\|_2^2$ converges to $(\text{p-lim}\|\boldsymbol{\eta}_1\|_2)^2$ as $d\to\infty$. Therefore, we have $\beta_1 = O_P(1)$, thus completing the proof of the lemma.

A.8. Proof of Lemma 6.4

By definition, $\boldsymbol{\eta}_1 = \boldsymbol{D}_\sigma^1\boldsymbol{W}_2^\top\boldsymbol{\eta}_2$. Recall that $\boldsymbol{u}_1$ is independent of $\sigma\{\bar{\boldsymbol{W}}_1^\top\boldsymbol{\eta}_1, \mathcal{F}_1\}$, then we can conclude that $(\boldsymbol{g}_1, \boldsymbol{D}_\sigma^1, \boldsymbol{\eta}_2, \boldsymbol{u}_1)$ depends on $\boldsymbol{W}_2$ only through $\boldsymbol{W}_2\boldsymbol{h}_1$. Therefore, by Lemma 3.1, there exists $\bar{\boldsymbol{W}}_2\in\mathbb{R}^{d_2\times d_1}$, which has the same marginal distribution with $\boldsymbol{W}_2$ and is independent of $(\boldsymbol{g}_1, \boldsymbol{D}_\sigma^1, \boldsymbol{\eta}_2, \boldsymbol{u}_1)$, such that

$$\begin{aligned}\boldsymbol{g}_1^s &= \boldsymbol{g}_1(1-\mu_1) - \beta_1\boldsymbol{D}_\sigma^1\Pi_{\boldsymbol{h}_1}\boldsymbol{W}_2^\top\boldsymbol{\eta}_2 - \beta_1\boldsymbol{D}_\sigma^1\Pi_{\boldsymbol{h}_1}^\perp\bar{\boldsymbol{W}}_2^\top\boldsymbol{\eta}_2 - \gamma_1\boldsymbol{u}_1\\ &= \boldsymbol{g}_1(1-\mu_1) + \beta_1\boldsymbol{D}_\sigma^1\boldsymbol{h}_1\frac{\boldsymbol{h}_1^\top\bar{\boldsymbol{W}}_2^\top\boldsymbol{\eta}_2 - \boldsymbol{h}_1^\top\boldsymbol{W}_2^\top\boldsymbol{\eta}_2}{\|\boldsymbol{h}_1\|_2^2} - \gamma_1\boldsymbol{u}_1 - \beta_1\boldsymbol{D}_\sigma^1\bar{\boldsymbol{W}}_2^\top\boldsymbol{\eta}_2\\ &= \boldsymbol{g}_1(1-\mu_1) + \zeta_1\boldsymbol{D}_\sigma^1\sigma(\boldsymbol{g}_1) - \gamma_1\boldsymbol{u}_1 - \beta_1\bar{\gamma}_1\boldsymbol{D}_\sigma^1\bar{\boldsymbol{u}}_1/\sqrt{d_1},\end{aligned}$$

where $\bar{\boldsymbol{u}}_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, is independent of $(\boldsymbol{g}_1, \boldsymbol{D}_\sigma^1, \boldsymbol{\eta}_2, \boldsymbol{u}_1)$ and

$$\zeta_1 := \beta_1\frac{\boldsymbol{h}_1^\top\bar{\boldsymbol{W}}_2^\top\boldsymbol{\eta}_2 - \boldsymbol{h}_1^\top\boldsymbol{W}_2^\top\boldsymbol{\eta}_2}{\|\boldsymbol{h}_1\|_2^2}, \quad \bar{\gamma}_1 := \|\boldsymbol{\eta}_2\|_2.$$

Recall that in Lemma 6.3, we have shown that

$$\mu_1 = o_P(1), \quad \beta_1 = O_P(1), \quad \gamma_1 = o_P(1)$$

as $d \to \infty$. By Lemma 6.1 claims (5), (1), and (3),

$$\boldsymbol{h}_1^\top \boldsymbol{W}_2^\top \boldsymbol{\eta}_2 = O_P(1), \quad \|\boldsymbol{h}_1\|_2^{-2} = O_P(d_1^{-1}), \quad \overline{\gamma}_1 = O_P(1).$$

By controlling the second moment then applying Chebyshev's inequality, we can conclude that $\boldsymbol{h}_1^\top \overline{\boldsymbol{W}}_2^\top \boldsymbol{\eta}_2 = O_P(1)$. Therefore, $\zeta_1 = o_P(1)$ and $\beta_1 \overline{\gamma}_1/\sqrt{d_1} = o_P(1)$.

For $\boldsymbol{\theta} \in \mathbb{R}^4$, we define

$$m_{\boldsymbol{\theta}}(g, u, \overline{u}) := \sigma\big(g(1-\theta_1) + \theta_2 \sigma'(g)\sigma(g) - \theta_3 u - \theta_4 \sigma'(g)\overline{u}\big)\sigma(g).$$

By assumption $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. Using this assumption, we can conclude that for any $\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \in \mathbb{R}^4$ satisfying $\|\boldsymbol{\theta}\|_\infty, \|\overline{\boldsymbol{\theta}}\|_\infty \le 1$,

$$|m_{\boldsymbol{\theta}}(g, u, \overline{u}) - m_{\overline{\boldsymbol{\theta}}}(g, u, \overline{u})| \le C'_\sigma \|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|_2 \big(1 + |g|^{n(k)} + |u|^{n(k)} + |\overline{u}|^{n(k)}\big),$$

where $n(k) \in \mathbb{N}_+$ is a function of k and $C'_\sigma > 0$ is a function of σ. Notice that

$$\mathbb{E}_{(g,u,\overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)}\big[1 + |g|^{n(k)} + |u|^{n(k)} + |\overline{u}|^{n(k)}\big] < \infty.$$

Then by [25, Example 19.7, Theorem 19.4], we know that $\{m_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\|_\infty \le 1\}$ is a Glivenko–Cantelli class, thus

$$\sup_{\|\boldsymbol{\theta}\|_\infty \le 1} \left| \frac{1}{d_1} \sum_{i=1}^{d_1} m_{\boldsymbol{\theta}}(g_i, u_i, \overline{u}_i) - \mathbb{E}_{(g,u,\overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)}\big[m_{\boldsymbol{\theta}}(g, u, \overline{u})\big] \right| = o_P(1).$$

Using the equation above and dominated convergence theorem, as $d \to \infty$, we have

$$\begin{aligned}
\frac{1}{d_1}\langle \boldsymbol{h}_1, \sigma(\boldsymbol{g}_1^s)\rangle &= \frac{1}{d_1} \sum_{i=1}^{d_1} m_{(\mu_1, \zeta_1, \gamma_1, \beta_1 \overline{\gamma}_1/\sqrt{d_1})}(g_i, u_i, \overline{u}_i) \\
&= \mathbb{E}_{(g,u,\overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)}\big[m_{(\mu_1, \zeta_1, \gamma_1, \beta_1 \overline{\gamma}_1/\sqrt{d_1})}(g, u, \overline{u})\big] + o_P(1) \\
&= \mathbb{E}_{(g,u,\overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)}\big[m_{\boldsymbol{0}}(g, u, \overline{u})\big] + o_P(1).
\end{aligned}$$

By the law of large numbers,

$$\|\boldsymbol{h}_1\|_2^2 = \mathbb{E}_{(g,u,\overline{u}) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3)}\big[m_{\boldsymbol{0}}(g, u, \overline{u})\big] + o_P(1),$$

thus proving the second claim of the lemma. Again via proof of a uniform convergence-type result, we can conclude that

$$\|\sigma(\boldsymbol{g}_1^s)\|_2^2/d_1 = \|\boldsymbol{h}_1\|_2^2/d_1 + o_P(1),$$

thus completing the proof of the first result.

A.9. Proof of Lemma 6.5

Given equation (6.3), $\mu_{m+1} = o_P(1)$ and $\gamma_{m+1} = o_P(1)$ follow from claim (iv) of the induction hypothesis $\mathcal{H}_m$.

A.10. Proof of Lemma 6.6

By assumption, almost everywhere we have $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. Therefore, there exists $C'_\sigma > 0$ which is a constant depending only on σ, such that

$$|\sigma(x)| \le C'_\sigma(1 + |x|^k).$$

Using these facts, we conclude that there exists $n(k) \in \mathbb{N}_+$ which is a function of k, and $C''_\sigma > 0$ which is a function of (σ, Ω_{m+1}) only, such that for all $\boldsymbol{\theta}, \overline{\boldsymbol{\theta}} \in \Omega_{m+1}$, we have

$$|\overline{h}_{\boldsymbol{\theta}}^{(m+1)}(\overline{u}, z, u) - \overline{h}_{\overline{\boldsymbol{\theta}}}^{(m+1)}(\overline{u}, z, u)| \le C''_\sigma \|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|_2 \big(1 + |\overline{u}|^{n(k)} + |z|^{n(k)} + |u|^{n(k)}\big).$$

Note that

$$\mathbb{E}_{(\overline{u},z,u)\sim \mathsf{N}(\mathbf{0},\boldsymbol{I}_3)}\Big[\big(1 + |\overline{u}|^{n(k)} + |z|^{n(k)} + |u|^{n(k)}\big)^2\Big] < \infty.$$

The rest of the proof is almost identical to that of Lemma 3.4: We apply the well-known results regarding Donsker class in [25] and prove that $\{\overline{h}_{\boldsymbol{\theta}}^{(m+1)} : \boldsymbol{\theta} \in \Omega_{m+1}\}$ is a Donsker class via proving the corresponding bracketing integral is finite. Here, we skip the details to avoid duplication.

A.11. Proof of Lemma 6.7

We define $\mathcal{S}, \overline{\mathcal{S}} : \mathbb{R}^{3\times d_m} \to \mathbb{R}$ as follows:

$$\mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) := \frac{1}{\sqrt{d_m}} \sum_{i=1}^{d_m} h^{(m+1)}_{(\gamma_{m+1},\mu_m,\nu_m,\beta_m/\sqrt{d_m},\beta_m\delta_{m+1},\gamma_m)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i}),$$

$$\overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) := \frac{1}{\sqrt{d_m}} \sum_{i=1}^{d_m} \overline{h}^{(m+1)}_{(\overline{\gamma}_{m+1},\mu_m,\nu_m,\beta_m/\sqrt{d_m},\beta_m\delta_{m+1},\gamma_m)}(\overline{u}_{m+1,i}, z_{m,i}, u_{m,i}).$$

By assumption, almost everywhere we have $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. Therefore, there exists $C'_\sigma > 0$ which is a constant depending only on σ, such that

$$|\sigma(x)| \le C'_\sigma(1 + |x|^k).$$

Using these facts, via some computation we can conclude that there exist $n_0 \in \mathbb{N}_+$ and $\overline{C}_\sigma > 0$ depending only on (σ, Ω_{m+1}), such that

$$\begin{aligned}
\left|\mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) - \overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right| &\le \frac{\overline{C}_\sigma}{d_m} \sum_{i=1}^{d_m} |\sigma'(\nu_m z_{m,i}) - \sigma'(H_{m-1} z_{m,i})| \\
&\times \left(|\beta_m| \overline{\gamma}_{m+1}^2 \overline{u}_{m+1,i}^2 + |d_m^{1/2} \beta_m \delta_{m+1} \overline{\gamma}_{m+1} \overline{u}_{m+1,i}| \left(1 + \nu_m^k |z_{m,i}|^k\right)\right) \\
&\times \left(1 + |\overline{u}_{m+1,i}|^{n_0} + |z_{m,i}|^{n_0} + |u_{m,i}|^{n_0}\right) \\
&\times \left(1 + |\mu_m|^{n_0} + |\nu_m|^{n_0} + |\beta_m|^{n_0} + |\overline{\gamma}_{m+1}|^{n_0} + |\gamma_m|^{n_0} + |\delta_{m+1}|^{n_0}\right).
\end{aligned}$$

By equation (6.6), ν_m is independent of $(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)$. Recall that we have proved

$$\begin{aligned}
&\beta_m = O_P(1), \quad \delta_{m+1} = O_P(d_m^{-1}), \quad \nu_m = H_{m-1} + o_P(1), \\
&\overline{\gamma}_{m+1} = E_{m+1} + o_P(1), \quad \gamma_m = o_P(1), \quad \mu_m = o_P(1).
\end{aligned}$$

Note that σ' is almost everywhere continuous and $|\sigma'(x)| \le C_\sigma(1 + |x|^{k-1})$. Then by applying Chebyshev's inequality and dominated convergence theorem to the above equation, we can conclude that as $d \to \infty$,

$$\left|\mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) - \overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right| = o_P(1). \tag{A.15}$$

Similarly, we have

$$\left|\mathbb{E}\left[\mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right] - \mathbb{E}\left[\overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right]\right| = o_P(1), \tag{A.16}$$

where in the above equation the expectations are taken over $(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)$, assuming

$$(\mu_m, \nu_m, \beta_m, \overline{\gamma}_{m+1}, \gamma_m, \delta_{m+1})$$

are fixed. Note that

$$\begin{aligned}
&\mathbb{G}_d^{(m+1)}\left(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m \delta_{m+1}, \gamma_m\right) \\
&\qquad = \mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) - \mathbb{E}\left[\mathcal{S}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right], \\
&\overline{\mathbb{G}}_d^{(m+1)}\left(\overline{\gamma}_{m+1}, \mu_m, \nu_m, \beta_m/\sqrt{d_m}, \beta_m \delta_{m+1}, \gamma_m\right) \\
&\qquad = \overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m) - \mathbb{E}\left[\overline{\mathcal{S}}(\overline{\boldsymbol{u}}_{m+1}, \boldsymbol{z}_m, \boldsymbol{u}_m)\right],
\end{aligned}$$

thus we have completed the proof of the lemma using equations (A.15) and (A.16).

A.12. Proof of Lemma 6.8

By definition, $\boldsymbol{\eta}_{m+1} = \boldsymbol{D}_\sigma^{m+1} \boldsymbol{W}_{m+2}^\mathsf{T} \boldsymbol{\eta}_{m+2}$. Note that $(\boldsymbol{g}_{m+1}, \boldsymbol{D}_\sigma^{m+1}, \boldsymbol{\eta}_{m+2}, \boldsymbol{u}_{m+1})$ depends on $\boldsymbol{W}_{m+2}$ only through $\boldsymbol{g}_{m+2} = \boldsymbol{W}_{m+2} \boldsymbol{h}_{m+1}$ and $\boldsymbol{h}_{m+1}$ is independent of $\boldsymbol{W}_{m+2}$. By Lemma 3.1, there exists $\bar{\boldsymbol{W}}_{m+2} \in \mathbb{R}^{d_{m+2}\times d_{m+1}}$ that has the same marginal distribution as $\boldsymbol{W}_{m+2}$ and $\bar{\boldsymbol{W}}_{m+2} \perp (\boldsymbol{g}_{m+1}, \boldsymbol{D}_\sigma^{m+1}, \boldsymbol{\eta}_{m+2}, \boldsymbol{u}_{m+1})$, such that

$$\begin{aligned}\boldsymbol{g}^s_{m+1} &= (1-\mu_{m+1})\boldsymbol{g}_{m+1} \\ &\quad + \beta_{m+1}\boldsymbol{D}_\sigma^{m+1}\boldsymbol{h}_{m+1}\frac{\boldsymbol{h}_{m+1}^\top \bar{\boldsymbol{W}}_{m+2}^\top \boldsymbol{\eta}_{m+2} - \boldsymbol{h}_{m+1}^\top \boldsymbol{W}_{m+2}^\top \boldsymbol{\eta}_{m+2}}{\|\boldsymbol{h}_{m+1}\|_2^2} \\ &\quad - \beta_{m+1}\boldsymbol{D}_\sigma^{m+1}\bar{\boldsymbol{W}}_{m+2}^\top\boldsymbol{\eta}_{m+2} - \gamma_{m+1}\boldsymbol{u}_{m+1} \\ &= (1-\mu_{m+1})\boldsymbol{g}_{m+1} + \kappa_{m+1}\boldsymbol{D}_\sigma^{m+1}\sigma(\boldsymbol{g}_{m+1}) - \gamma_{m+1}\boldsymbol{u}_{m+1} \\ &\quad - \beta_{m+1}\bar{\gamma}_{m+2}\boldsymbol{D}_\sigma^{m+1}\bar{\boldsymbol{u}}_{m+2}/\sqrt{d_{m+1}},\end{aligned}$$

where $\bar{\boldsymbol{u}}_{m+2} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m+1}})$ is independent of $(\boldsymbol{g}_{m+1}, \boldsymbol{D}_\sigma^{m+1}, \boldsymbol{\eta}_{m+2}, \boldsymbol{u}_{m+1})$, and

$$\kappa_{m+1} := \beta_{m+1}\frac{\boldsymbol{h}_{m+1}^\top \bar{\boldsymbol{W}}_{m+1}^\top \boldsymbol{\eta}_{m+2} - \boldsymbol{h}_{m+1}^\top \boldsymbol{W}_{m+2}^\top \boldsymbol{\eta}_{m+2}}{\|\boldsymbol{h}_{m+1}\|_2^2}, \quad \bar{\gamma}_{m+2} := \|\boldsymbol{\eta}_{m+2}\|_2.$$

By $\mathcal{H}_{m+1}$ claim (iii),

$$\mu_{m+1} = o_P(1), \quad \gamma_{m+1} = o_P(1), \quad \beta_{m+1}/\sqrt{d_{m+1}} = o_P(1).$$

By Lemma 6.1,

$$\|\boldsymbol{h}_{m+1}\|_2^{-2} = O_P(d_{m+1}^{-1}), \quad \bar{\gamma}_{m+1} = O_P(1), \quad \boldsymbol{h}_{m+1}^\top \boldsymbol{W}_{m+2}^\top \boldsymbol{\eta}_{m+2} = O_P(1).$$

Then we compute the corresponding second moment and apply Chebyshev's inequality, and conclude that

$$\boldsymbol{h}_{m+1}^\top \bar{\boldsymbol{W}}_{m+2}^\top \boldsymbol{\eta}_{m+2} = O_P(1).$$

Combining these analysis, we have $\kappa_{m+1} = o_P(1)$.

We can write $\boldsymbol{g}_{m+1} = \nu_{m+1}\boldsymbol{z}_{m+1}$, where $\nu_{m+1} := \sqrt{\|\boldsymbol{h}_m\|_2^2/d_m}$ and $\boldsymbol{z}_{m+1} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m+1}})$ is independent of $\boldsymbol{h}_m$. Adopting similar arguments we applied to obtain equation (6.6), we can conclude that

$$\boldsymbol{g}_{m+1}, \bar{\boldsymbol{u}}_{m+2}, \boldsymbol{u}_{m+1} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_{d_{m+1}}).$$

Recall that we have proved

$$\mu_{m+1} = o_P(1), \quad \kappa_{m+1} = o_P(1), \quad \gamma_{m+1} = o_P(1), \quad \beta_{m+1}\bar{\gamma}_{m+2}/\sqrt{d_{m+1}} = o_P(1).$$

By assumption, for all $x, y \in \mathbb{R}$,

$$|\sigma(x) - \sigma(y)| \le C_\sigma |x-y|(1 + |x|^{k-1} + |y|^{k-1}). \tag{A.17}$$

Then we apply equation (A.17) to bound the difference between $\boldsymbol{h}_{m+1} = \sigma(\boldsymbol{g}_m)$ and $\sigma(\boldsymbol{g}_m^s)$, and the lemma follows from simple application of the law of large numbers.

B. Proofs for the non-asymptotic results

B.1. Proof of Lemma 4.2

We will heavily rely on the Bernstein inequality to prove the lemma, and we state it here for readers' convenience.

Lemma B.1. *Let $X_1, \dots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-c \min\left\{\frac{t^2}{\sum_{i=1}^{N} \|X_i\|_{\Psi_1}^2}, \frac{t}{\max_{i\in[N]} \|X_i\|_{\Psi_1}}\right\}\right),$$

where $c > 0$ is a numerical constant, and $\|\cdot\|_{\Psi_1}$ is the Orlicz norm.

Proof of the first result. It suffices to prove that with probability at least

$$1 - C\eta_1^{-2}(\sigma(0)^2 + L^2)$$

for some positive numerical constant $C > 0$, the following inequality holds:

$$|\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}| \leq \eta_1.$$

Notice that $\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a} = m^{-1/2} \sum_{i=1}^{m} g_i \sigma(g_i) b_i$. Therefore,

$$\mathrm{Var}\big[\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\big] \leq \mathbb{E}_{b,g\sim_{\text{i.i.d.}}\mathsf{N}(0,1)}\big[b^2 g^2 \sigma(g)^2\big] \leq C(\sigma(0)^2 + L^2)$$

for some numerical constant $C > 0$. By Chebyshev's inequality, with probability at least $1 - C\eta_1^{-2}(\sigma(0)^2 + L^2)$, $|\boldsymbol{g}^\top \boldsymbol{D}_\sigma \boldsymbol{a}| \leq \eta_1$, thus completing the proof for this part.

Proof of the second result. By the definitions of $\bar{\boldsymbol{W}}$ and $\bar{\boldsymbol{W}}_c$, we see that

$$\frac{\|\bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2 - \langle \bar{\boldsymbol{W}}_c^\top \boldsymbol{D}_\sigma \boldsymbol{a}, \bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\rangle}{\|\boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2} \stackrel{d}{=} \frac{1}{d}\sum_{i=1}^{d}(z_i^2 - z_i z_i'),$$

where $z_i, z_i' \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$. By Lemma B.1, we see that there exists a numerical constant $C > 0$, such that for all $\eta_2 \geq 1$, with probability at least $1 - 4\exp(-C\eta_2)$ the following inequalities hold:

$$\left|\frac{1}{d}\sum_{i=1}^{d} z_i^2 - 1\right| \leq \frac{\eta_2}{\sqrt{d}}, \quad \left|\frac{1}{d}\sum_{i=1}^{d} z_i z_i'\right| \leq \frac{\eta_2}{\sqrt{d}}.$$

When the event described above occurs, we see that

$$\left|\beta - \frac{\tau s_d}{\sqrt{m}}\right| \leq \frac{2 s_d \eta_2}{\sqrt{md}},$$

which completes the proof of the second result.

Proof of the third result. Notice that

$$\gamma^2 = \frac{s_d^2}{d} \cdot \|\bar{\boldsymbol{W}}^\top \boldsymbol{D}_\sigma \boldsymbol{a}\|_2^2 \overset{d}{=} \frac{s_d^2}{d} \cdot \left(\frac{1}{d}\sum_{i=1}^{d} z_i^2\right) \cdot \left(\frac{1}{m}\sum_{i=1}^{m} b_i^2 \sigma(g_i)^2\right),$$

where $z_i, b_i, g_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$. Since $|\sigma(x)| \le |\sigma(0)| + L|x|$, we then conclude that there exists a numerical constant $C > 0$, such that

$$\operatorname{Var}\left[\frac{1}{m}\sum_{i=1}^{m} b_i^2 \sigma(g_i)^2\right] \le \frac{C(\sigma(0)^4 + L^4)}{m}.$$

By Chebyshev's inequality, with probability at least $1 - Cm^{-1}(\sigma(0)^4 + L^4)$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} b_i^2 \sigma(g_i)^2 - \mathbb{E}_{g\sim\mathsf{N}(0,1)}\big[\sigma(g)^2\big]\right| \le 1.$$

By Bernstein's inequality (Lemma B.1), with probability at least $1 - 2\exp(-cd)$ for some absolute constant $c > 0$, we have

$$d^{-1}\sum_{i=1}^{d} z_i^2 \le 2.$$

In summary, with probability at least $1 - 2\exp(-cd) - Cm^{-1}(\sigma(0)^4 + L^4)$,

$$\gamma \le \frac{2 s_d}{\sqrt{d}} \cdot \sqrt{1 + \mathbb{E}_{g\sim\mathsf{N}(0,1)}\big[\sigma(g)^2\big]},$$

thus concluding the proof of the third result.

B.2. Proof of Lemma 4.3

Recall that the sub-Gaussian norm of a random variable X is defined as

$$\|X\|_{\Psi_2} := \inf\big\{t > 0 : \mathbb{E}\big[\exp(X^2/t^2)\big] \le 2\big\}.$$

Standard computation implies that for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\big[\exp\big(\lambda(\mathbb{L}_m(\boldsymbol{\theta}) - \mathbb{L}_m(\boldsymbol{\theta}'))\big)\big] \\
&\quad \le \exp\bigg(\frac{\lambda^2}{2m}\sum_{i=1}^{m} b_i^2\big(\sigma\big((1-\theta_1)g_i - \theta_2 b_i \sigma'(g_i) - \theta_3 u_i\big) \\
&\qquad - \sigma\big((1-\theta_1')g_i - \theta_2' b_i \sigma'(g_i) - \theta_3' u_i\big)\big)^2\bigg)
\end{aligned}$$

$$\leq \exp\Big(\frac{\lambda^2 L^2}{2m}\sum_{i=1}^m b_i^2\big(|\theta_1-\theta_1'|\cdot|g_i| + L\cdot|\theta_2-\theta_2'|\cdot|b_i| + |\theta_3-\theta_3'|\cdot|u_i|\big)^2\Big)$$

$$\leq \exp\Big(\frac{\lambda^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2}{2}\cdot\frac{1}{m}\sum_{i=1}^m M(b_i,g_i,u_i)^2\Big).$$

Hence, by the sub-Gaussian property

$$\|\mathbb{L}_m(\boldsymbol{\theta}) - \mathbb{L}_m(\boldsymbol{\theta}')\|_{\Psi_2} \leq C\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2\cdot\sqrt{\frac{1}{m}\sum_{i=1}^m M(b_i,g_i,u_i)^2}$$

for some positive numerical constant C.

B.3. Proof of Lemma 4.4

We observe that

$$F(\boldsymbol{g}) \stackrel{d}{=} m^{-1} z\sum_{i=1}^m \sigma(g_i)^2,$$

where $z \sim \mathsf{N}(0,1)$ is independent of $\{g_i\}_{i\leq m}$. Notice that there exists a numerical constant $C' > 0$, such that

$$\|\sigma(g)^2\|_{\Psi_1} \leq C'(\sigma(0)^2 + L^2).$$

Then by Lemma B.1, there exist numerical constants $c, C > 0$, such that

$$\mathbb{P}\bigg(\bigg|\frac{1}{m}\sum_{i=1}^m \sigma(g_i)^2 - \mathbb{E}_{g\sim\mathsf{N}(0,1)}\big[\sigma(g)^2\big]\bigg| \geq 1\bigg) \leq C\exp\Big(-\frac{cm}{L^4+\sigma(0)^4}\Big),$$
$$\mathbb{P}\big(|z| \geq \eta_3\big) \leq C\exp(-c\eta_3^2),$$

which concludes the proof of the lemma.

Acknowledgments. We thank anonymous reviewers for carefully reading the manuscript and giving helpful suggestions.

Funding. This work was supported by the NSF through award DMS-2031883, the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, the NSF grant CCF-2006489, the ONR grant N00014-18-1-2729.

References

[1] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*, Curran Associates, 2019

[2] P. Bartlett, S. Bubeck, and Y. Cherapanamjeri. Adversarial examples in multi-layer random ReLU networks. In *Advances in Neural Information Processing Systems 34*, pp. 9241–9252, Curran Associates, 2021

[3] P. Bartlett, A. Montanari, and A. Rakhlin, Deep learning: A statistical viewpoint. *Acta Numer.* **30** (2021), 87–201 Zbl 07674571 MR 4295218

[4] S. Bubeck, Y. Cherapanamjeri, G. Gidel, and R. Tachet des Combes, A single gradient step finds adversarial examples on random two-layers neural networks, pp. 10081–10091, *Advances in Neural Information Processing Systems 34*, Curran Associates, 2021

[5] S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn, Adversarial examples from computational constraints. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 831–840, Proceedings of Machine Learning Research 97, PMLR, 2019

[6] T. A. Bui, T. Le, Q. Tran, H. Zhao, and D. Phung. A unified Wasserstein distributional robustness framework for adversarial training. 2022, arXiv:2202.13437

[7] N. Carlini and D. Wagner, Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, IEEE, 2018

[8] A. Daniely and H. Shacham, Most ReLU networks suffer from ℓ_2 adversarial perturbations. *Advances in Neural Information Processing Systems 33*, pp. 6629–6636, Curran Associates, 2020

[9] A. Fawzi, H. Fawzi, and O. Fawzi, Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems 31*, Curran Associates, 2018

[10] J. Gilmer, L. Metz, F. Faghri, S. S Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, Adversarial spheres. 2018, arXiv:1801.02774

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2014, arXiv:1412.6572

[12] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems 32*, Curran Associates, 2019

[13] J. Kos, I. Fischer, and D. Song, Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 36–42, IEEE, 2018

[14] A. Kurakin, I. Goodfellow, and S. Bengio, Adversarial machine learning at scale. 2016, arXiv:1611.01236

[15] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent. *J. Stat. Mech. Theory Exp.* (2020), no. 12, article no. 124002 Zbl 07330523 MR 4241355

[16] S. Mei, T. Misiakiewicz, and A. Montanari, Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.* **59** (2022), 3–84 Zbl 07522503 MR 4412180

[17] A. Montanari and Y. Zhong, The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *Ann. Statist.* **50** (2022), no. 5, 2816–2847 Zbl 07628842 MR 4500626

[18] R. M. Neal, Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53, Lect. Notes Stat. 118, Springer, New York, 1996

[19] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, Association for Computing Machinery, New York, NY, 2017

[20] C. Qin, J. Martens, S. Gowal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, and P. Kohli, Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems 32*, Curran Associates, 2019

[21] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, Curran Associates, 2019

[22] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, Are adversarial examples inevitable? 2018, arXiv:1809.02104

[23] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks. 2013, arXiv:1312.6199

[24] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, Ensemble adversarial training: Attacks and defenses. 2017, arXiv:1705.07204

[25] A. W. van der Vaart, *Asymptotic statistics*. Camb. Ser. Stat. Probab. Math. 3, Cambridge University Press, Cambridge, 1998 Zbl 0910.62001 MR 1652247

[26] G. Vardi, G. Yehudai, and O. Shamir, Gradient methods provably converge to non-robust networks. 2022, arXiv:2202.04347

[27] E. Wong, L. Rice, and J. Z. Kolter, Fast is better than free: Revisiting adversarial training. 2020, arXiv:2001.03994

[28] X. Yuan, P. He, Q. Zhu, and X. Li, Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **30** (2019), no. 9, 2805–2824 MR 4001274

Received 15 April 2022; revised 22 January 2023.

Andrea Montanari
Department of Electrical Engineering and Department of Statistics, Stanford University, Stanford, CA 94305, USA; montanari@stanford.edu

Yuchen Wu
Department of Statistics, Stanford University, Stanford, CA 94305, USA; wuyc14@stanford.edu