Benign Overfitting in Adversarial Training of Neural Networks

Yunjuan Wang [1] Kaibo Zhang [1] Raman Arora [1]

Abstract

Benign overfitting is the phenomenon wherein none of the predictors in the hypothesis class can achieve perfect accuracy (i.e., non-realizable or noisy setting), but a model that interpolates the training data still achieves good generalization. A series of recent works aim to understand this phenomenon for regression and classification tasks using linear predictors as well as two-layer neural networks. In this paper, we study such a benign overfitting phenomenon in an adversarial setting. We show that under a distributional assumption, interpolating neural networks found using adversarial training generalize well despite inference-time attacks. Specifically, we provide convergence and generalization guarantees for adversarial training of two-layer networks (with smooth as well as non-smooth activation functions) showing that under moderate ℓ_2 norm perturbation budget, the trained model has near-zero robust training loss and near-optimal robust generalization error. We support our theoretical findings with an empirical study on synthetic and real-world data.

1. Introduction

Neural networks have been widely used in real-world applications, achieving state-of-the-art performance on various tasks such as image classification and speech recognition. Despite their tendency to be over-parameterized and capable of interpolating the training data with significant label noise, neural networks perform surprisingly well on previously unseen test data. This seemingly contradicts the classical learning theory where overfitting to the training data would typically hinder with generalization. Such phenomena, known as *benign overfitting* (Bartlett et al., 2020), is technically characterized by the following conditions: (1) the trained classifier perfectly fits the noisy training data, achieving zero training error; (2) no classifier in the related hypothesis class can achieve near-zero generalization error; (3) the trained classifier achieves near-optimal generalization error. Several recent works seek to unravel the mystery of benign overfitting in various settings, including training linear models (Bartlett et al., 2020; Chatterji et al., 2022), kernel methods (Belkin et al., 2018; Liang & Rakhlin, 2020; Mei & Montanari, 2022) and training of neural networks (Frei et al., 2022; Cao et al., 2022).

While models based on neural networks have been tremendously successful, they are highly vulnerable to small, nearly imperceptible, albeit strategic, perturbation of data. These perturbations, called adversarial examples, are abundant and easy to find computationally (Bubeck et al., 2021). The potential of such adversarial attacks to substantially degrade the performance of an otherwise well-performing model has been a source of significant concern regarding deployment of machine learning models in real-world systems. It is no surprise, then, that developing algorithms that can provably defend against such attacks and are guaranteed to improve the robustness of machine learning has gained tremendous traction in recent years.

One of the most prominent empirical defense algorithms against inference-time attacks is the adversarial training method of Madry et al. (2018). Adversarial training proceeds by simulating attacks as part of training – generating adversarial examples from (clean) training examples and using them to train a neural network. We can view adversarial training as a two-player game, wherein the learner seeks to minimize their error on the training set while an adversary strives to maximize the error by crafting small strategic corruptions of the input training examples. Several empirical studies show that by using adversarial training or its variants (Zhang et al., 2019; Wang et al., 2020), the learner returns a model that is more resilient to perturbations in the input space (Madry et al., 2018; Shafahi et al., 2019b; Dong et al., 2020; Pang et al., 2021).

In contrast to the benign overfitting phenomenon that occurs in the standard (clean) setting, Sanyal et al. (2020) identified a sufficient condition on the data distribution that hurts robust generalization when the classifier perfectly fits the noisy label data. This "*robust overfitting*" phenomenon was also confirmed by Rice et al. (2020) showing that on sev-

[1] Department of Computer Science, Johns Hopkins University, Baltimore, USA. Correspondence to: Yunjuan Wang <ywang509@jhu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

eral real-world datasets, the robust test loss increases after the first learning rate decay while the robust training loss keeps decreasing throughout training. This naturally begs the question whether the modern wisdom of training neural networks to zero training loss also extends to adversarial settings, or in other words, if *benign overfitting can occur in adversarial training*. Chen et al. (2023) took a first step in studying benign overfitting for adversarially trained linear models and provide empirical results for both linear classifiers and neural networks. They acknowledged that it is nontrivial to generalize the analysis to neural networks and leave it for future work.

In this paper, we resolve the open question posed by Chen et al. (2023), asserting that benign overfitting can also occur in adversarially trained neural networks under certain data distributions. Our key contributions are as follows.

1. Given training data generated from a mixture distribution with label noise, we establish convergence guarantees for adversarial training of two-layer neural networks showing that robust training loss can be driven to zero, thereby robustly interpolating the noisy training data. We consider the hypothesis class given by two-layer neural networks. We consider smooth as well as non-smooth activation functions. Furthermore, we do not make any assumption about the robust realizability of data.
2. We provide generalization guarantees on both the clean test error and the robust test error, demonstrating that they simultaneously achieve the near-optimal standard and adversarial risk. In particular, for a moderately large network we show that for ℓ_2 norm-bounded additive adversarial attacks, if the perturbation budget is not too large, the robust test error approximates the label noise rate.
3. We validate our theoretical results with experiments on both synthetic and real-world datasets.

1.1. Related Work

Benign Overfitting. A significant body of recent works has delved into understanding why predictors that interpolate noisy training data can still achieve a good generalization performance, with a particular emphasis on linear models, e.g., linear regression (Bartlett et al., 2020; Hastie et al., 2022; Zou et al., 2021b; Chatterji et al., 2022; Koehler et al., 2021), sparse regression (Wang et al., 2022a; Chatterji & Long, 2022), logistic regression (Chatterji & Long, 2021; Wang et al., 2021), ridge regression (Tsigler & Bartlett, 2020), and kernel methods (Belkin et al., 2018; Liang et al., 2020; Liang & Rakhlin, 2020; Mei & Montanari, 2022).

For nonlinear model such as neural networks, analyzing benign overfitting becomes much more challenging. There has been some progress toward addressing this challenge. Frei et al. (2022) provided a first such guarantee for finite-width neural networks trained on logistic loss for data drawn from a Gaussian mixture model. Concurrently, Cao et al. (2022) characterized the generalization guarantees of two-layer convolutional neural networks, assuming that input data is a sum of a label-dependent signal patch and a label-independent noise patch. While the works above consider a smooth activation function, follow-up studies by Kou et al. (2023) and Xu & Gu (2023) extended the results to SGD for training neural networks with non-smooth activations (e.g., ReLU). Recently, Zhu et al. (2023) further extended these findings to deep neural networks in the lazy regime.

Robust Overfitting. Numerous works focus on mitigating overfitting in adversarial settings following the work of Rice et al. (2020). These include approaches that employ heuristic ideas, such as early stopping, adding regularization, adapting cyclic learning rate schedules (Rice et al., 2020), and smoothing the logits or weights during training (Chen et al., 2021), among others (Pang et al., 2021; Huang et al., 2020; Dong et al., 2022). Thw works of Xiao et al. (2022); Clarysse et al. (2022); Fu & Wang (2023) provide some theoretical justification for these practical approaches. Donhauser et al. (2021) and Dong et al. (2021) implicate memorization – neither work provides any theoretical results to support their claim. However, follow-up work by Li & Li (2023) considers a patch data distribution with a meaningful signal patch embedded in noisy patches–they show that the ability of a model class to memorize spurious features (noisy patches) leads to overfitting. More recently, Li et al. (2022) argued that robust generalization may require exponentially large models.

Robust Generalization Guarantees. A standard technical tool for establishing generalization bounds is that of uniform convergence. Several works build on this idea to give generalization guarantees for the robust loss, by analyzing Rademacher complexity (Yin et al., 2019; Khim & Loh, 2018; Awasthi et al., 2020), VC dimension (Cullina et al., 2018; Montasser et al., 2020), or the covering number (Balda et al., 2019; Mustafa et al., 2022; Li & Telgarsky, 2023), of the hypothesis class or utilizing PAC Bayesian analysis (Viallard et al., 2021; Xiao et al., 2023) and margin-theoretic analysis (Farnia et al., 2018). However, by definition, these guarantees rely on bounding the generalization gap, i.e., the difference between the empirical and expected error, of all hypothesis in the hypothesis class simultaneously. As such, uniform convergence bounds are unable to explain the benign overfitting phenomenon, wherein the empirical and expected errors of an interpolating predictor are not close to each other.

Computational Guarantees. The statistical guarantees based on uniform convergence fail to explain benign overfitting. It is natural then to rely on a more direct (e.g., trajectory-based) analysis of the output of the training algorithm. However, a good theoretical understanding of why and when adversarial training succeeds remains elusive. Much of the recent work (Charles et al., 2019; Li et al., 2020; Zou et al., 2021a; Chen et al., 2023) has focused on studying adversarial training of linear models wherein the adversarial examples are given in a simple closed-form expression – this simplifies the problem greatly reducing it to standard training. Of special relevance to us in this body of results, is the work of Chen et al. (2023) who claim to demonstrate benign overfitting for linear models; yet, they fail to show that the model returned by adversarial training in the setting they consider has small robust training error, making their claim questionable.

Adversarial training of neural networks was analyzed by Gao et al. (2019) and further improved by Zhang et al. (2020); however, both of these works focus on ensuring convergence of the training procedure and do not provide generalization guarantees on robust loss. This gap has been addressed in very recent work by Li & Telgarsky (2023). However, the work of Li & Telgarsky (2023), and the prior work all focus on the lazy training regime, which, unfortunately, has been proven to be at odds with robustness (Wang et al., 2022b). Finally, Allen-Zhu & Li (2022) present an analysis of adversarial training when initialized using a network returned by standard (clean) training instead of random initialization. Mianjy & Arora (2023) provide an end-to-end analysis of adversarial training beyond the NTK setting with a variant of adversarial training that involves using a slightly different (reflected) loss for the inner loop maximization problem (for finding an attack vector as part of adversarial training), yet, their results are limited to robustly realizable distributions, which cannot justify benign overfitting as there is no noise in their setting.

2. Preliminaries

Notation. Throughout the paper, we denote scalars, vectors, and matrices with lowercase italics, lowercase bold, and uppercase bold Roman letters, respectively; e.g., u, u, and U. We use $[m]$ to denote the set $\{1, 2, \ldots, m\}$ and use both $\|\cdot\|$ and $\|\cdot\|_2$ for ℓ_2-norm. Given a matrix $\mathrm{U} = [\mathrm{u}_1, \ldots, \mathrm{u}_m] \in \mathbb{R}^{d\times m}$, we use $\|\mathrm{U}\|_F$ and $\|\mathrm{U}\|_2$ to represent the Frobenius norm and spectral norm, respectively. We use $\mathcal{B}_2(\mathrm{u}, \alpha)$ to denote the ℓ_2 ball centered at $\mathrm{u} \in \mathbb{R}^d$ of radius α. We use the standard O-notation ($\mathcal{O}$, Θ and Ω).

2.1. Problem Setup

We focus on binary classification and denote the input space and label space as $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{\pm 1\}$, respectively. We assume that the data are drawn from a noisy mixture data distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ that, along with its variants, has been studied in several recent works (Chatterji & Long, 2021; Cao et al., 2021; Frei et al., 2022). Formally, we consider the following data distribution.

Definition 2.1 (Data Distribution)**.** Let $\mathcal{D}_{\text{clust}}$ be a λ-strongly log-concave distribution over $\mathbb{R}^d$ for some $\lambda > 0$. We assume that $\mathcal{D}_{\text{clust}} = \mathcal{D}_{\text{clust}}^{(1)} \times \cdots \times \mathcal{D}_{\text{clust}}^{(d)}$ is a product distribution whose marginals are all mean-zero with the sub-Gaussian norm at most one. We further assume that $\mathbb{E}_{\xi\sim\mathcal{D}_{\text{clust}}}[\|\xi\|^2] \geq \kappa d$ holds for some $0 < \kappa < 1$. Let $\mathcal{D}_c$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. We first draw a sample $(\mathrm{x}_c, y_c) \sim \mathcal{D}_c$ by sampling $y_c \in \{\pm 1\}$ uniformly at random, sampling $\xi \sim \mathcal{D}_{\text{clust}}$, and setting $\mathrm{x}_c = y_c\mu + \xi$. Given a noise rate $\beta > 0$, we define our true data distribution $\mathcal{D}$ to be any distribution over $\mathcal{X} \times \mathcal{Y}$ such that the marginal distribution of $\mathcal{D}$ and $\mathcal{D}_c$ on $\mathcal{X}$ are the same, and the total variation distance between the two distributions is bounded by β, i.e., $d_{\text{TV}}(\mathcal{D}_c, \mathcal{D}) \leq \beta$.

The standard coupling lemma states that given two distributions $\mathcal{D}$ and $\mathcal{D}_c$ over the same domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, there exists a joint distribution over $\mathcal{Z} \times \mathcal{Z}$ such that the marginals along the projections $(z, z') \mapsto z$ and $(z, z') \mapsto z'$ are $\mathcal{D}$ and $\mathcal{D}_c$, respectively. Given that the marginal on $\mathcal{X}$ for $\mathcal{D}$ and $\mathcal{D}_c$ are the same (see the definition above), this implies that for $(\mathrm{x}, y) \sim \mathcal{D}$, $(\mathrm{x}_c, y_c) \sim \mathcal{D}_c$, $\mathrm{P}(\mathrm{x} = \mathrm{x}_c) = 1$ and $\mathrm{P}(y \neq y_c) \leq \beta$. The definition above includes two settings: 1) Independent label flip, where for each sample, label y is obtained by flipping y_c with probability at most β, independent of how other labels are generated; 2) Non-independent label flip, where there exists potential correlations between labels y. A yet another special instance that has been studied extensively in the adversarial learning literature is that of Gaussian distribution (Javanmard et al., 2020; Dobriban et al., 2020; Dan et al., 2020) which is a special case of the data generative model above for $\beta = 0$.

Hypothesis Class. We focus on learning two-layer neural networks defined as: $f(\mathrm{x}; \mathrm{W}) := \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \phi(\langle \mathrm{w}_s, \mathrm{x}\rangle)$ where m is an even integer representing the number of hidden nodes and $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function. The weight matrix at the bottom layer is denoted as $\mathrm{W} = [\mathrm{w}_1, \ldots, \mathrm{w}_m] \in \mathbb{R}^{d\times m}$ and the weight vector at the top layer by $\mathrm{a} = [a_1, \ldots, a_m] = [1, \ldots, 1, -1, \ldots, -1] \in \mathbb{R}^m$. The top layer weight vector a is kept fixed throughout the training process. The weight vectors at the bottom layer are initialized randomly as $\mathrm{w}_s^0 \sim \mathsf{N}(0, \omega_{\text{init}}^2 \mathrm{I})$, for $s \in \{1, \ldots, \frac{m}{2}\}$, and setting $\mathrm{w}_s^0 = \mathrm{w}_{s-\frac{m}{2}}^0$ for $s \in \{\frac{m}{2} + 1, \ldots, m\}$. This ensures symmetry at initialization and yields $f(\mathrm{x}; \mathrm{W}^0) = 0$ for all x. This symmetric initialization technique is commonly used in related work (Langer, 2021; Bartlett et al., 2021; Montanari & Zhong, 2022) and we employ here for analytical purposes.

Training Data. We are given a training data of size n sampled i.i.d. from the noisy data distribution, $\mathcal{S} =$

$\{(\mathrm{x}_i, y_i)\}_{i=1}^{n} \sim \mathcal{D}$. Let $\mathcal{C}$ denote the set of indices of training data corresponding to the clean labels; i.e., for $i \in \mathcal{C}$, we have that $(\mathrm{x}_i, y_i) \sim \mathcal{D}_c$; similarly, let $\mathcal{N}$ denote the indices corresponding to noisy labels; i.e., $(\mathrm{x}_i, -y_i) \sim \mathcal{D}_c \ \ \forall i \in \mathcal{N}$.

Loss Function. The 0-1 loss of a predictor $f(\cdot, \mathrm{W})$ on a data point (x, y) is defined as $\ell^{0/1}((\mathrm{x}, y); \mathrm{W}) = \mathbb{1}\left(y f(\mathrm{x}; \mathrm{W}) \leq 0\right)$, where $\mathbb{1}(\cdot)$ is the indicator function. For computational reasons, as is typical, we use the logistic loss, denoted $\ell(z) = \log\left(1 + \exp\left(-z\right)\right)$, to train the two-layer neural networks. The population and the empirical loss w.r.t. $\ell(\cdot)$ are denoted as $L(\mathrm{W}) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}} \ell(y f(\mathrm{x}; \mathrm{W}))$ and $\widehat{L}(\mathrm{W}) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(\mathrm{x}_i; \mathrm{W}))$.

Robust Loss. We consider ℓ_2 norm-bounded adversarial attacks with a perturbation budget of size $\alpha > 0$. The set of all such perturbations for an input example $\mathrm{x} \in \mathcal{X}$ is represented by $\mathcal{B}_2(\mathrm{x}, \alpha)$. This threat model motivates minimizing the robust 0-1 loss defined as $\ell_{\text{rob}}^{0/1}((\mathrm{x}, y); \mathrm{W}) = \max_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha)} \mathbb{1}(y f(\tilde{\mathrm{x}}; \mathrm{W}) \leq 0)$. The population and empirical risk w.r.t. the 0-1 loss and the robust 0-1 loss, respectively, are denoted as $L^{0/1}$, $\widehat{L}^{0/1}$, $L_{\text{rob}}^{0/1}$, and $\widehat{L}_{\text{rob}}^{0/1}$. Analogously, the population and empirical robust risk w.r.t. the (surrogate) logistic loss $\ell(\cdot)$ are defined as:

$$L_{\text{rob}}(\mathrm{W}) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}} \max_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha)} \ell(y f(\tilde{\mathrm{x}}; \mathrm{W}))$$

$$\widehat{L}_{\text{rob}}(\mathrm{W}) := \frac{1}{n}\sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W})).$$

Note that we are ultimately interested in bounding the 0-1 loss and its robust variant.

Algorithm 1 Gradient Descent-based Adversarial Training

```
Input: Step size $\eta$, perturbation budget per sample $\alpha$. Number of iterations $T$.
  Initialize $\mathrm{W}^0$ randomly.
  for $t = 0, \ldots, T-1$ do
    for $i = 1, \ldots, n$ do
      $\tilde{\mathrm{x}}_i^t = \arg\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t))$.
    end for
    Update $\mathrm{W}^{t+1} = \mathrm{W}^t - \frac{\eta}{n}\sum_{i=1}^{n} \nabla \ell(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t))$
  end for
  return: $\mathrm{W}^T$
```

Adversarial Training. The gradient descent-based adversarial training algorithm is presented in Algorithm 1. We denote the adversarial training example for some input x_i given model parameter W^t, at round t as $\tilde{\mathrm{x}}_i^t = \arg\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)) = \arg\min_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)$. A bi-product of our initialization is that $f(\cdot; \mathrm{W}^0)$ is the zero function at initialization (see the paragraph titled “Hypothesis Class”). Therefore, at $t = 0$, all perturbations of all training data fare equally, i.e., perturbing data does not increase the training loss. Therefore, for simplicity, we simply choose to not perturb the training data at iteration $t = 0$[1]. The proposed symmetric initialization and simple modification of the training procedure at $t = 0$ is instrumental in proving Lemma 4.1, which further yields tight results.

We make the following assumptions on the problem setup. Specifically, we consider a high dimensional setting where the dimension d is much larger than the number of training samples n, as stated below in Assumption (1). Such a regime is popular in biomedical settings where the data comes from limited patient information such as MRI or DNA sequence.

Assumption 1. Let $\delta \in (0, 1/2)$. We assume that there exists a positive constant C such that the following holds: (1) The dimension satisfies $d \geq C \max\{\|\mu\|^2 n, n^2 \left(\log\left(n/\delta\right) + \alpha^2\right)\}$. (2) noise rate $\beta \in [0, 1/C]$. (3) Initialization variance satisfies $\omega_{\text{init}}\sqrt{md} \leq \eta$. (4) Step size $\eta \leq (Cd^2)^{-1}$. (5) The number of samples satisfies $n \geq C \log\left(m/\delta\right)$. (6) Adversarial perturbation $\alpha \leq \|\mu\|$.

Assumption (3) requires a small initialization to ensure that the first step of adversarial training dominates the behavior of the neural network, pushing it beyond the lazy training regime. Such initialization technique has also been introduced in previous work (Ba et al., 2019; Xing et al., 2021). Given that the objective of adversarial training is to achieve a classifier that is robust against small input perturbations imperceptible to human eyes, Assumption (6) is reasonable as it imposes a mild constraint on the attack strength. Finally, we note that we can relax Assumption (1) to $d \geq C \max\{\|\mu\|^2 n, n^2 \log\left(n/\delta\right)\}$, thereby removing the dependence on α (see discussion in Section 4.1). We work with the assumption above to keep our arguments and proofs relatively simple and accessible.

3. Main Result

In this section, we present our main result providing theoretical guarantees for adversarial training of neural networks. We assume that the underlying distribution is the noisy mixture distribution described in Section 2.1. Further, we consider network architectures with both smooth and non-smooth activation functions – while we show identical results for both cases, we need slightly different assumptions for the two. Therefore, we first separately describe each setting before presenting a unified result.

[1] Due to this simple modification, we can allow perturbation budget α to be as large as $\|\mu\|$ (see Assumption (6)). On the other hand, if we allow for non-zero perturbation at $t = 0$, we will need $\alpha \leq c\,\|\mu\|$ for some $c \in [0, 1)$.

3.1. Smooth Activation Function

Here we consider a strictly increasing, 1-Lipschitz, H-smooth activation function that is approximately homogeneous with $\phi(0) = 0$. Formally, there exists $\gamma, H > 0, 0 \leq \zeta < 1, c_1 \geq 0, c_2 \geq 0$ such that $0 < \gamma \leq \phi'(z) \leq 1$, $\phi'(z)$ is H-Lipschitz, and $|\phi'(z) \cdot z - \phi(z)| \leq c_1 + c_2 |z|^\zeta, \forall z \in \mathbb{R}$. Smooth activation functions have been extensively studied both theoretically and empirically (Liu & Di, 2021; Biswas et al., 2022). One example of such an activation function that satisfies our condition is the smoothed Leaky ReLU activation (Frei et al., 2022) defined as follows:

$$\phi_{\text{SLReLU}}(z) = \begin{cases} z - \frac{1-\gamma}{4H}, & z \geq \frac{1}{H} \\ \frac{1-\gamma}{4} H z^2 + \frac{1+\gamma}{2} z, & |z| \leq \frac{1}{H} \\ \gamma z - \frac{1-\gamma}{4H}, & z \leq -\frac{1}{H} \end{cases}. \quad (1)$$

However, we do need an additional assumption on top of what (Frei et al., 2022) require. In particular, we assume that $\phi'(z)z$ and $\phi(z)$ are close to each other. We argue that this is a mild assumption, and holds trivially for standard ReLU and Leaky ReLU, with $c_1 = c_2 = 0$. For $\phi_{\text{SLReLU}}(z)$, of (Frei et al., 2022), the assumption holds with $\zeta = 0$ with $c_1 = \frac{1-\gamma}{4H}$, and $c_2 = 0$. The reason we need this additional assumption is because the neural networks with $\phi_{\text{SLReLU}}(z)$ activation function are no longer homogeneous. Consequently, without the assumption, we end up with terms in the upper bound on the empirical robust risk that depends on the Frobenius norm of the weight matrix (see Section 4.3 for more details).

3.2. Non-smooth activation function

Here, we consider a more practical setting where the activation function is no longer smooth. We consider a homogeneous non-smooth activation function that satisfies the following properties.

$$\begin{aligned} &\phi(0) = 0, \phi'(z)z = \phi(z), \forall z \in \mathbb{R}; \\ &0 \leq \phi'(z) \leq 1, \forall z \in \mathbb{R}; \\ &\exists \gamma \in (0, 1], \text{s.t.} \phi'(z) \geq \gamma, \forall z > 0. \end{aligned}$$

This includes ReLU and Leaky ReLU activation functions.

3.3. Theoretical Guarantees

Our main result establishes benign overfitting in adversarially trained neural networks. In particular, we show that adversarial training converges to neural networks with zero robust training loss and with standard (clean) test error close to the noise rate. Furthermore, for small attack strength, α, the robust test error also converges to the noise rate. Formally, we show the following.

Theorem 3.1. Let $\varepsilon > 0, \delta \in (0, 1/2)$. Let $\kappa \in (0, 1)$ and $\lambda > 0$ as given in Definition 2.1. We consider the following regimes and parameter settings for smooth and non-smooth activations functions, respectively.

Smooth Activation. Let ϕ be a γ-leaky H-smooth activation with $0 \leq \zeta < 1$. Set $\bar{T} = \left(\frac{35+8\sqrt{m/d^3}}{\gamma\|\mu\|\eta\varepsilon}\right)^{\frac{2}{1-\zeta}}$. We assume that there exists some constant $C > 0$ such that Assumption 1 holds, (A1) $d \leq \|\mu\|^4 / C$, and (A2) $\|\mu\|^2 \geq C \log(n/\delta)$.

Non-smooth Activation. Let ϕ be a non-smooth activation with $\gamma \in (0, 1]$. Set $\bar{T} = \Omega\left(\frac{1}{\|\mu\|^2\gamma^2\eta\varepsilon^2}\right)$. We assume that there exists some constant $C > 0$ such that Assumption 1 holds, (B1) $m \geq C \log(n/\delta)$, and (B2) $\|\mu\|^2 \geq C \max\left\{\sqrt{\frac{d}{n}\log(md/n\delta)}, \log(n/\delta)\right\}$.

Then, there exists a constant $c > 0$ such that after running Algorithm 1 for $T \geq \bar{T}$ iterations, we have that with probability at least $1 - 2\delta$ over the random initialization and the draw of an i.i.d. sample of size n, the following holds:

1. The robust training loss satisfies $\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \varepsilon$, the robust training error satisfies $\widehat{L}^{0/1}_{\text{rob}}(\mathrm{W}^T) = 0$.
2. The clean test error satisfies
$$L^{0/1}(\mathrm{W}^T) \leq \beta + 2\exp\left(-\frac{c\lambda n \|\mu\|^4}{C^2 d}\right).$$
3. For $\frac{\alpha}{\|\mu\|} \leq \frac{1}{C}\sqrt{\frac{n\|\mu\|^2}{d}}$, the robust test error satisfies
$$L^{0/1}_{\text{rob}}(\mathrm{W}^T) \leq \beta + 2\exp\left(-c\lambda \|\mu\|^2 \left(\frac{\|\mu\|}{C}\sqrt{\frac{n}{d}} - \frac{\alpha}{\|\mu\|}\right)^2\right).$$

3.4. Discussion

Theorem 3.1 shows that adversarially-trained neural networks can interpolate the noisy training data. In fact, the trained network correctly classifies all training data after the first step of adversarial training. For generalization guarantees, Theorems 3.1 suggests an interesting interplay between the parameters d, n, and $\|\mu\|$. Importantly, when $n \gg \tilde{\Omega}\left(\frac{d}{\|\mu\|^4}\right)$, it ensures a small clean test error. Furthermore, when $n \gg \tilde{\Omega}\left(\frac{d(1+\alpha)^2}{\|\mu\|^4}\right)$[2], the robust test error is also guaranteed to be small. This result aligns with the literature suggesting that adversarial robustness requires more data (Schmidt et al., 2018). Notably, the clean test error obtained through the adversarial training algorithm shares the same bound as that derived through gradient descent (Frei et al., 2022; Xu & Gu, 2023), even when the perturbation size α is as large as the signal size $\|\mu\|$. For the robust generalization error, the constraint on the perturbation can

[2] If we fix $\frac{\alpha}{\|\mu\|}$ to be a constant, then $n \gg \tilde{\Omega}\left(\frac{d}{\|\mu\|^2}\right)$ guarantees the robust test error to be small. This is verified in Section 5.

be the same scale as the signal size; i.e. $\alpha \leq \mathcal{O}(\|\mu\|)$ when $d = \Theta(n \|\mu\|^2)$. It is worth noting that the robust test error decreases as n/d increases or as the attack strength $\frac{\alpha}{\|\mu\|}$ decreases, which is consistent with the findings in previous literature (Schmidt et al., 2018; Shafahi et al., 2019a).

For non-smooth activation functions, Assumption (B1) is a relatively mild constraint on the network width. Assumption (B2) is slightly more stringent compared to Assumption (A2). However, it is worth noting that in the clean setting, the minimax generalization error is at least $\mathcal{O}(\exp(-\min(\|\mu\|^2, n\|\mu\|^4/d)))$ (Giraud & Verzelen, 2019), implying that (B2) is unavoidable up to logarithmic factors if we desire a classifier with good generalization.

Next, we provide a lower bound on the robust test error that is independent of the algorithm as well as hypothesis class.

Theorem 3.2. We consider independent label flip with probability β. Let $p(\mathrm{x})$ be the density function of $\mathcal{D}_{\text{clust}}$. For any given classifier $f(\cdot;\mathrm{W})$, when $\alpha < \|\mu\|$, we have $L_{\text{rob}}^{0/1}(\mathrm{W}) \geq \beta + \frac{1-2\beta}{4}\int_{\mathbb{R}^d} \min\{p(\xi), p(\xi + \mathrm{v})\} d\xi$, where $\mathrm{v} = 2(1 - \alpha/\|\mu\|)\mu$. When $\alpha \geq \|\mu\|$, the robust test error satisfies $L_{\text{rob}}^{0/1}(\mathrm{W}) \geq 0.5$.

Consider the special instance of when $\mathcal{D}_{\text{clust}}$ is standard Gaussian. Theorem 3.2 recovers the optimal risk in Dobriban et al. (2020) up to a scaling factor when $\beta = 0$. Moreover, the upper bound on the robust test error (denoted as UBD) that we provide in Theorem 3.1 and the lower bound (denoted as LBD) in Theorem 3.2 satisfy the following: $(\text{UBD} - \beta) = (\text{LBD} - \beta)^{\mathcal{O}(n\|\mu\|^2/d)}$. Our upper bound roughly matches the lower bound when $\frac{n\|\mu\|^2}{d} = \Omega(1)$.

Overfitting with Adversarial Training. While our result may, at first, seem in conflict with the robust overfitting phenomenon that observed by recent empirical studies, we note that there is actually **no contradiction** with this empirical observation as we consider a specific data-generative model and a bound on the size of the adversarial perturbation during adversarial training. Indeed, recent empirical studies by (Dong et al., 2021) and (Yu et al., 2022) confirm that small α prevents adversarial training from overfitting. Furthermore, (Xing et al., 2022) explored the phase transition between standard training and adversarial training and showed that the optimization trajectories in the two settings are close to each other when α is small. One interesting future direction is to justify the generalization guarantee for moderately large attack strength $\frac{\alpha}{\|\mu\|}$.

Comparison with Theoretical Works. Several recent works focus on giving convergence and generalization guarantees for adversarial training (Gao et al., 2019; Zhang et al., 2020; Mianjy & Arora, 2023; Li & Telgarsky, 2023); here we compare and contrast our work with each of these. The work of (Gao et al., 2019) prove convergence for a modified algorithm for adversarial training wherein the iterates are projected onto a norm ball to ensure that the network weights stay close to initialization. However, they further need to assume that a robust network exists in the vicinity of the initialization. Such an assumption has been shown to be invalid in a recent work (Wang et al., 2022b). In a related work, (Zhang et al., 2020) provide a fine-grained convergence analysis for datasets that are well-separated. More recently, (Li & Telgarsky, 2023) give convergence and generalization guarantees for adversarial training of shallow networks with early stopping. Unfortunately, all of the aforementioned works are limited to the lazy regime (aka, the NTK setting) which has been shown to be at odds with adversarial robustness (Wang et al., 2022b). (Mianjy & Arora, 2023) were the first to provide both convergence and generalization guarantees beyond the NTK regime, yet their analysis was restricted to robust realizable data distributions.

Our work stands out from prior work in several ways. First, we study the standard adversarial training algorithm commonly used in practice. Second, we give convergence guarantees for adversarial training on non-separable data, unlike other works that make restrictive assumptions regarding linear separability and robust realizability. Finally, our results hold for neural networks that can be trained for arbitrary many iterations allowing $\|\mathrm{W}^t\|$ to go to infinity, i.e., beyond the NTK regime.

4. Proof Sketch

We begin by providing some intuition for our proof. We show that when the perturbation size is not large ($\alpha \leq \|\mu\|$), the trajectory of the adversarial training remains close to that of the standard training. Furthermore, given a good initialization of the neural network the dynamics of the training algorithm can be shown to be nearly linear. We also leverage a result from high dimensional probability, that the training data we draw is (nearly) separable even though the underlying data distribution is non-separable. We show that both of these events happen with high probability and establish what we refer to as a “good” run of the algorithm and are central to our proof.

Next, we formalize this intuition and provide a brief proof sketch of our main result. We focus primarily on neural networks with smooth activation function (i.e., Theorem 3.1) and note the differences in the analysis when extending the result to the non-smooth activation functions. In our analysis, we borrow many ideas from (Frei et al., 2022) and (Xu & Gu, 2023). However, the extension is not straightforward and our focus in this section is on highlighting the technical challenges we overcome and the key insights we utilized in our analysis. We also identify several non-rigorous arguments and present a discussion regarding technical improvements over (Frei et al., 2022); we defer them to Appendix B.1 due to space limitations. For detailed proofs, we refer the reader to the Appendix.

4.1. Properties of Adversarial Training Examples

For convergence and generalization guarantees, Assumption (6) allows the perturbation α to be as large as $\|\mu\|$. This requires a fine-grained analysis of the properties of the adversarial examples generated during the training process, as characterized in the following lemma.

Lemma 4.1. $\forall t \in \mathbb{N}, i \in [n]$, the distance between $\tilde{\mathrm{x}}_i^t$ and $\mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ satisfies $\mathrm{dist}(\tilde{\mathrm{x}}_i^t, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) \leq \min\left\{\omega_{\text{init}}\sqrt{md}/\eta, \alpha\right\}$.

Essentially, a smaller initialization on the model weight ω_{init} leads to a shorter distance between the generated adversarial examples to the linear subspace spanned by the training data $\{\mathrm{x}_i\}_{i=1}^n$. Notably, $\mathrm{dist}(\tilde{\mathrm{x}}_i^t, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) \leq \min\{1, \alpha\}$ due to Assumption (3). This helps us control the size of $\left|\langle \mu, \tilde{\mathrm{x}}_i^t \rangle\right| = \mathcal{O}(\|\mu\|^2)$, independently of α when α is relatively large.

We can leverage another property of $\tilde{\mathrm{x}}_i^t$ to relax Assumption (1) to $d \geq C \max\{\|\mu\|^2 n, n^2 \log(n/\delta)\}$. In particular, we show that the angle between the direction of the additive adversarial perturbation for each training examples, i.e., $y_i(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)$, and the direction of the combined training data (i.e., $-\sum_{k=1}^n y_k \mathrm{x}_k$) is small. This allows us to control the size of $\langle \tilde{\mathrm{x}}_i^t, \tilde{\mathrm{x}}_j^t \rangle$ for all $i, j \in [n]$. Both of these properties we discuss above are crucial to our analysis and proofs.

4.2. Generalization Guarantee

As a proof strategy we seek to get an upper bound on the robust test error in terms of a lower bound on the normalized expected conditional margin. This follows using a concentration argument given that $\mathcal{D}_{\text{clust}}$ is λ-strongly log-concave.

Lemma 4.2. Suppose that $\mathbb{E}_{(\mathrm{x}, y_c) \sim \mathcal{D}_c}[y_c f(\mathrm{x}; \mathrm{W}) | y_c = \bar{y}] - \|\mathrm{W}\|_2 \alpha \geq 0$ holds for both $\bar{y} = 1$ and $\bar{y} = -1$. Then, there exists a universal constant $c > 0$ such that

$$L_{\text{rob}}^{0/1}(\mathrm{W}) \leq \beta + \sum_{\bar{y} \in \{-1, +1\}} \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x}, y_c) \sim \mathcal{D}_c}[y_c f(\mathrm{x}; \mathrm{W}) | y_c = \bar{y}]}{\|\mathrm{W}\|_2} - \alpha\right)^2\right)$$

To get a lower bound on the normalized expected conditional margin, we leverage the smoothness property of the activation function to derive a lower bound on the increment in the un-normalized margin for an independent test example (x, y).

Lemma 4.3 (Informal)**.** For some constant C_2, with high probability, we have for any $t \geq 0$ and $(\mathrm{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, there exist $\tilde{\rho}_i^t = \rho\left(\mathrm{W}^t, \tilde{\mathrm{x}}_i^t, \mathrm{x}\right) \in [\gamma^2, 1]$ such that

$$\begin{aligned} & y\left[f(\mathrm{x}; \mathrm{W}^{t+1}) - f(\mathrm{x}; \mathrm{W}^t)\right] \\ & \quad \geq \frac{\eta}{n} \sum_{i=1}^n \tilde{g}_i(\mathrm{W}^t)\left(\tilde{\xi}_i^t \left\langle y_i \tilde{\mathrm{x}}_i^t, y\mathrm{x} \right\rangle - \frac{H \|\mathrm{x}\|^2 C_2^2 d\eta}{2\sqrt{mn}}\right). \end{aligned}$$

where $\tilde{g}_i(\mathrm{W}^t) = -\ell'(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) = \frac{1}{1 + \exp\left(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right)}$.

For the non-smooth activation function, we get a similar result which we defer to the Appendix due to space constraints. Finally, we seek a positive lower bound on un-normalized expected conditional margin for model W^t by expressing it in terms of the cumulative increments of margin; i.e., showing $\mathbb{E}_{(\mathrm{x}, y_c) \sim \mathcal{D}_c | y_c = 1}[y_c f(\mathrm{x}; \mathrm{W}^t)] = \sum_{t=1}^T \mathbb{E}_{(\mathrm{x}, y_c) \sim \mathcal{D}_c | y_c = 1}[y_c f(\mathrm{x}; \mathrm{W}^t) - y_c f(\mathrm{x}; \mathrm{W}^{t-1})] + \mathbb{E}_{(\mathrm{x}, y_c) \sim \mathcal{D}_c | y_c = 1}[y_c f(\mathrm{x}; \mathrm{W}^0)]$. A positive lower bound holds trivially positive if $\langle y_i \tilde{\mathrm{x}}_i^t, y_c \mathrm{x} \rangle$ is always bounded below by some positive constant. However, due to the presence of noisy labels y_i and adversarial examples $\tilde{\mathrm{x}}_i$, $\langle y_i \tilde{\mathrm{x}}_i^t, y_c \mathrm{x} \rangle$ may be negative. Note, though, that the term $\langle y_i \tilde{\mathrm{x}}_i^t, y_c \mathrm{x} \rangle$ scales with $\tilde{g}_i(\mathrm{W}^t)$. If we can show that $\tilde{g}_i(\mathrm{W}^t)$ is of the same order across all training examples, and assume a small perturbation budget and that only a small fraction of labels are noisy, then we can mitigate the effect of the negative terms. The key lemma providing such a result by bounding the loss ratio is as follows.

Lemma 4.4 (Informal)**.** Given Assumption 1, there is an absolute constant $C_r > 0$ such that with high probability, we have for all $t \geq 0$, $\max_{i,j \in [n]} \frac{\tilde{g}_i(\mathrm{W}^t)}{\tilde{g}_j(\mathrm{W}^t)} \leq C_r$.

To see why the above holds, note that for any given $i, j \in [n]$, we have that $\frac{\tilde{g}_i(\mathrm{W}^t)}{\tilde{g}_j(\mathrm{W}^t)} \leq \max\left\{2, \frac{2\exp\left(-y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t)\right)}\right\}$, where $\tilde{\mathrm{x}}_i^t = \arg\min_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i; \alpha)} y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)$. For successive iterates we get that $\frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^{t+1}; \mathrm{W}^{t+1})\right)} \leq \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t)\right)} \cdot \frac{\exp\left(y_i f(\tilde{\mathrm{x}}_i^{t+1}; \mathrm{W}^t) - y_i f(\tilde{\mathrm{x}}_i^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t) - y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^{t+1})\right)}$. Finally, we use induction to complete the proof.

For smooth activation functions, the proof of Lemmas 4.3 and 4.4, follows by controling the term $y\left[f(\mathrm{x}; \mathrm{W}^{t+1}) - f(\mathrm{x}, \mathrm{W}^t)\right]$ via Taylor approximation. For non-smooth activation functions, we need to ensure that there exist enough neurons have positive activations at initialization as well as throughout the training process.

Lemma 4.5 (Informal)**.** Given Assumption 1 and (B2), with high probability, for all $s \in [m]$, we have $\left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i \right\rangle \geq \alpha \left\|\mathrm{w}_s^1\right\|\right\}\right| = \Theta(n)$; for all $i \in [n]$, we have $\left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i \right\rangle \geq \alpha \left\|\mathrm{w}_s^1\right\|\right\}\right| = \Theta(m)$.

4.3. Convergence Guarantee

In order to control the robust training loss, a naive approach would be to decouple the increment of the robust training loss, from iterate t to $t+1$, into two terms as follows:

$$\begin{aligned}&\widehat{L}_{\text{rob}}(\mathrm{W}^{t+1})-\widehat{L}_{\text{rob}}(\mathrm{W}^{t})\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\Big[\left(\ell(y_i f(\tilde{\mathrm{x}}_i^{t+1};\mathrm{W}^{t+1}))-\ell(y_i f(\tilde{\mathrm{x}}_i^{t};\mathrm{W}^{t+1}))\right)\\&\qquad+\left(\ell(y_i f(\tilde{\mathrm{x}}_i^{t};\mathrm{W}^{t+1}))-\ell(y_i f(\tilde{\mathrm{x}}_i^{t};\mathrm{W}^{t}))\right)\Big].\end{aligned}$$

The second term can be controlled by the smoothness property of the loss function. The first term, unfortunately, is upper bounded by $\left\|\mathrm{W}^{t+1}\right\|\left\|\tilde{\mathrm{x}}_i^{t+1}-\tilde{\mathrm{x}}_i^{t}\right\|$, and the robust training loss hence inevitably depends on the norm of iterates $\left\|\mathrm{W}^{t+1}\right\|$ if no additional assumptions are made. This poses a problem if we do not constrain the model weights within a bounded domain, as $\left\|\mathrm{W}^{t}\right\|$ may tend to infinity as the number of epochs increases. To mitigate this issue, we instead control the robust training loss via the norm of the iterates. Specifically, we first show that $\widehat{L}_{\text{rob}}(\mathrm{W}^{T}) \leq \frac{2}{T}\sum_{t=0}^{T-1} G_{\text{rob}}(\mathrm{W}^{t})$ where $G_{\text{rob}}(\mathrm{W}) := \frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)} -\ell'(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}))$; this holds due to a property of the loss $\ell(\cdot)$ (see the Appendix for more details). We then bound $G_{\text{rob}}(\mathrm{W}^{t})$ by a constant scaling of $\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^{t}),\mathrm{V}\right\rangle$, where $\mathrm{V}\in\mathbb{R}^{m\times d}$ is a matrix with row $\mathrm{v}_s = a_s\mu/\left\|\mu\right\|$. We achieve this result using Lemma 4.4 and the fact that only a small fraction labels are noisy. Given $\sum_{t=0}^{T-1}\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^{t}),\mathrm{V}\right\rangle = \left\langle \mathrm{W}^{T},\mathrm{V}\right\rangle - \left\langle \mathrm{W}^{0},\mathrm{V}\right\rangle \leq \left\|\mathrm{W}^{T}\right\|_F + \left\|\mathrm{W}^{0}\right\|$, the only thing we need to prove is that the growth rate of $\left\|\mathrm{W}_T\right\|$ is smaller than $\mathcal{O}(T)$. This property holds for both smooth activation functions that satisfy our construction and non-smooth activation functions such as ReLU and Leaky ReLU.

5. Experiments

In this section, we present a simple empirical study on a synthetic dataset to support our theoretical results. We follow the generative model in Section 2 to synthesize a dataset with independent label flips when generating y from y_c. We set $\mu = \left\|\mu\right\|_2[1,0,0,\ldots,0]^\top$, $\beta = 0.1$, and generate $n = 100$ training samples and 2K test samples with the noise vector sampled from the standard multivariate Gaussian distribution, $\xi \sim \mathcal{N}(0,\mathrm{I})$. We train a two-layer ReLU network with width 1K. We use the default initialization in PyTorch and train the network applying full-batch gradient-descent based adversarial training using logistic loss for 1K iterations. We use PGD attack to generate adversarial examples with attack strength $\alpha/\left\|\mu\right\|$ and attack stepsize $\alpha/(5\left\|\mu\right\|)$ for 20 iterations. The outer minimization is trained using an initial learning rate of 0.1 with decay by 10 after training for every 500 iterations. We note that adversarial training achieves 100% robust training accuracy. We estimate the robust test accuracy using the same PGD attack. We consider settings with varying dimension d and attack strength $\frac{\alpha}{\|\mu\|}$.

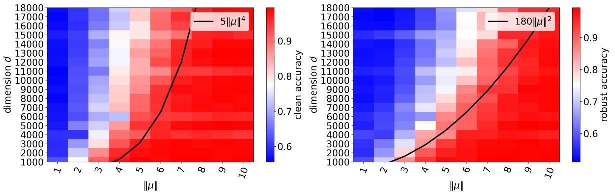

Figure 1. Clean test accuracy (left) / robust test accuracy (right) as a function of signal size $\|\mu\|$ and dimension d, for a fixed perturbation ratio $\alpha/\left\|\mu\right\| = 0.1$.

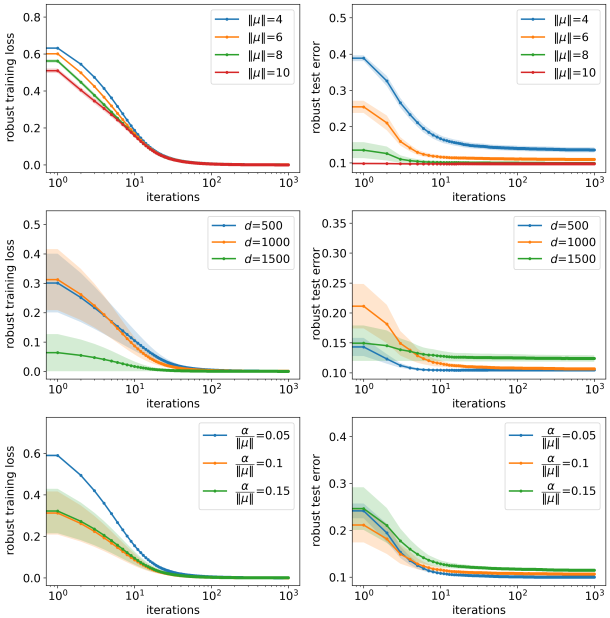

Figure 2. Robust training loss (left) / robust test error (right) as a function of training iterations. Top row: fix $d = 1000$, $\frac{\alpha}{\|\mu\|} = 0.1$. Middle row: fix $\|\mu\| = 5.0$, $\frac{\alpha}{\|\mu\|} = 0.1$. Bottom row: fix $\|\mu\| = 5.0$, $d = 1000$. Each curve is averaged over 10 runs and shaded regions show standard error.

For our first experiment, we fix the perturbation ratio $\frac{\alpha}{\|\mu\|} = 0.1$, and vary the value of the signal strength $\|\mu\|$ from 1 to 10 and the dimension d from 1K to 18K. We show the results in Figure 1 as a heat map of clean accuracy and robust accuracy averaged over ten independent random runs. We observe a phase transition for both clean accuracy and robust accuracy at the value of dimension d around $\mathcal{O}(\|\mu\|^4)$ for clean accuracy and $\mathcal{O}(\|\mu\|^2)$ for robust accuracy. This is consistent with the main theorems (see discussion in Section 3.4).

For our next experiment, we plot the robust training loss and robust test error as a function of the number of training iterations in Figure 2. For the top row, we fix $d = 1000$, $\frac{\alpha}{\|\mu\|} = 0.1$, and vary the signal size $\|\mu\| \in [4, 6, 8, 10]$; for the middle row, we fix $\|\mu\| = 5.0$, $\frac{\alpha}{\|\mu\|} = 0.1$, vary dimension $d \in [500, 1000, 1500]$; for the bottom row, we fix $\|\mu\| = 5.0, d = 1000$, vary attack rate $\frac{\alpha}{\|\mu\|} \in [0.05, 0.1, 0.15]$. We observe that the robust training loss goes to zero while the robust test error converges to the label noise rate of 0.1. Furthermore, smaller $\|\mu\|$, larger d, and larger $\frac{\alpha}{\|\mu\|}$ all lead to worse robust test error, which is consistent with our theory.

We observe the same trends on MNIST dataset even though the data generative assumptions are no longer valid; we defer a detailed discussion to the Appendix.

6. Conclusion

In this paper, we show benign overfitting in adversarial training of two-layer neural networks under a noisy mixture data distribution. Specifically, we show that under ℓ_2 norm perturbations, the robust training loss converges to zero while the robust generalization error is near-optimal. Our work suggests several promising future directions. Our results assume a generative model with a structured log-concave data distribution. It is natural to explore whether our findings can be extended to more general data distributions. Another interesting direction is to investigate whether our results generalize to the setting where the data dimension and the number of training samples have the same scale. Finally, we note that our main result only partially characterizes the phase transition from small to large test errors for small and large attack strengths, respectively. An important next step is to provide generalization guarantees for attacks of moderate strength and to explore the relationship between the perturbation size, signal size, dimension, and the number of training samples.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This research was supported, in part, by DARPA GARD award HR00112020004 and NSF CAREER award IIS-1943251.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441. PMLR, 2020.

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., and Zhang, T. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2019.

Balda, E. R., Behboodi, A., Koep, N., and Mathar, R. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2018.

Biswas, K., Kumar, S., Banerjee, S., and Pandey, A. K. Smooth Maximum Unit: Smooth Activation Function for Deep Networks using Smoothing Maximum Technique. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 794–803, 2022.

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

Cao, Y., Gu, Q., and Belkin, M. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign Overfitting in Two-layer Convolutional Neural Networks. *arXiv preprint arXiv:2202.06526*, 2022.

Charles, Z., Rajput, S., Wright, S., and Papailiopoulos, D. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*, 2019.

Chatterji, N. S. and Long, P. M. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *The Journal of Machine Learning Research*, 22 (1):5721–5750, 2021.

Chatterji, N. S. and Long, P. M. Foolish crowds support benign overfitting. *The Journal of Machine Learning Research*, 23(1):5448–5459, 2022.

Chatterji, N. S., Long, P. M., and Bartlett, P. L. The interplay between implicit bias and benign overfitting in two-layer linear networks. *The Journal of Machine Learning Research*, 23(1):12062–12109, 2022.

Chen, J., Cao, Y., and Gu, Q. Benign overfitting in adversarially robust linear classification. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021.

Clarysse, J., Hörrmann, J., and Yang, F. Why adversarial training can hurt robust accuracy. *arXiv preprint arXiv:2203.02006*, 2022.

Cullina, D., Bhagoji, A. N., and Mittal, P. PAC-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Dong, C., Liu, L., and Shang, J. Label noise in adversarial training: A novel perspective to study robust overfitting. *Advances in Neural Information Processing Systems*, 35: 17556–17567, 2022.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331, 2020.

Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H., and Zhu, J. Exploring memorization in adversarial training. In *International Conference on Learning Representations*, 2021.

Donhauser, K., Tifrea, A., Aerni, M., Heckel, R., and Yang, F. Interpolation can hurt robust generalization even when there is no noise. *Advances in Neural Information Processing Systems*, 34:23465–23477, 2021.

Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Frei, S., Chatterji, N. S., and Bartlett, P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Fu, S. and Wang, D. Theoretical analysis of robust overfitting for wide DNNs: An NTK approach. *arXiv preprint arXiv:2310.06112*, 2023.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Giraud, C. and Verzelen, N. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Huang, L., Zhang, C., and Zhang, H. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33:19365–19376, 2020.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pp. 2034–2078. PMLR, 2020.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign Overfitting for Two-layer ReLU Networks. *arXiv preprint arXiv:2303.04145*, 2023.

Langer, S. Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function. *Journal of Multivariate Analysis*, 182:104695, 2021.

Li, B. and Li, Y. Why clean generalization and robust overfitting both happen in adversarial training. *arXiv preprint arXiv:2306.01271*, 2023.

Li, B., Jin, J., Zhong, H., Hopcroft, J., and Wang, L. Why robust generalization in deep learning is difficult: Perspective of expressive power. *Advances in Neural Information Processing Systems*, 35:4370–4384, 2022.

Li, J. D. and Telgarsky, M. On achieving optimal adversarial test error. In *International Conference on Learning Representations*, 2023.

Li, Y., Fang, E., Xu, H., and Zhao, T. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. 2020.

Liang, T., Rakhlin, A., and Zhai, X. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pp. 2683–2711. PMLR, 2020.

Liu, X. and Di, X. TanhExp: A Smooth Activation Function with High Convergence Speed for Lightweight Neural Networks. *IET Computer Vision*, 15(2):136–150, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Mianjy, P. and Arora, R. Robustness guarantees for adversarially trained neural networks. *Advances in neural information processing systems*, 2023.

Montanari, A. and Zhong, Y. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

Montasser, O., Hanneke, S., and Srebro, N. Reducing adversarially robust learning to non-robust PAC learning. *Advances in Neural Information Processing Systems*, 33: 14626–14637, 2020.

Mustafa, W., Lei, Y., and Kloft, M. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pp. 16174–16196. PMLR, 2022.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Sanyal, A., Dokania, P. K., Kanade, V., and Torr, P. H. How benign is benign overfitting? *arXiv preprint arXiv:2007.04028*, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019a.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019b.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Viallard, P., VIDOT, E. G., Habrard, A., and Morvant, E. A pac-bayes analysis of adversarial robustness. *Advances in Neural Information Processing Systems*, 34:14421–14433, 2021.

Wang, G., Donhauser, K., and Yang, F. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10572–10602. PMLR, 2022a.

Wang, K., Muthukumar, V., and Thrampoulidis, C. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Wang, Y., Ullah, E., Mianjy, P., and Arora, R. Adversarial robustness is at odds with lazy training. *Advances in Neural Information Processing Systems*, 35:6505–6516, 2022b.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. Stability analysis and generalization bounds of adversarial training. *arXiv preprint arXiv:2210.00960*, 2022.

Xiao, J., Sun, R., and Luo, Z.-Q. Pac-bayesian adversarially robust generalization bounds for deep neural networks. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Xing, Y., Song, Q., and Cheng, G. On the generalization properties of adversarial training. In *International Conference on Artificial Intelligence and Statistics*, pp. 505–513. PMLR, 2021.

Xing, Y., Song, Q., and Cheng, G. Phase transition from clean training to adversarial training. In *Advances in Neural Information Processing Systems*, 2022.

Xu, X. and Gu, Y. Benign overfitting of non-smooth neural networks beyond lazy training. *International Conference on Artificial Intelligence and Statistics*, pp. 11094–11117, 2023.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094. PMLR, 2019.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, Y., Plevrakis, O., Du, S. S., Li, X., Song, Z., and Arora, S. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *Advances in Neural Information Processing Systems*, 33: 679–688, 2020.

Zhu, Z., Liu, F., Chrysos, G., Locatello, F., and Cevher, V. Benign overfitting in deep neural networks under lazy training. In *International Conference on Machine Learning*, pp. 43105–43128. PMLR, 2023.

Zou, D., Frei, S., and Gu, Q. Provable robustness of adversarial training for learning halfspaces with noise. In *International Conference on Machine Learning*, pp. 13002–13011. PMLR, 2021a.

Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. Benign overfitting of constant-stepsize SGD for linear regression. In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021b.

Supplementary Material

A. Additional Experiments

For the synthetic dataset, we also run an additional experiment. We fix the signal size $\|\mu\| = 5.0$, vary dimension d from 500 to 6K and perturbation ratio $\frac{\alpha}{\|\mu\|}$ from 0.05 to 0.45. Figure 3 plots the robust accuracy as a heat map averaged over ten independent runs. Our findings indicate that, increasing the dimension leads to a smaller perturbation ratio required to achieve the same level of robust test accuracy.

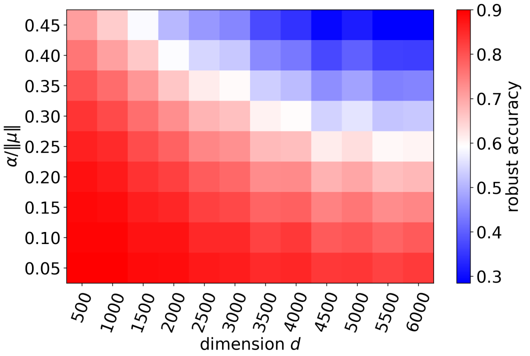

Figure 3. Robust test accuracy on synthetic dataset as a function of d and $\frac{\alpha}{\|\mu\|}$ for a fixed $\|\mu\| = 5$.

In order to see if our results extend beyond the generative data model we consider in this paper, we run the same set of experiments as above on the MNIST dataset. MNIST is a dataset of 28×28 greyscale handwritten digits. We extract examples corresponding to images of the digits '0' and '1', resulting in 12,665 training examples and 2,115 test examples. We view the input image as a vector input of size d and normalize the data to ensure that the ℓ_2 norm of each input vector is equal to $\|\mu\|$. A random subset of size $n = 100$ is used for training. We do not introduce any label noise; i.e., $\beta = 0$. We train a two-layer ReLU network with width 1K using the same training procedure as for the experiments on the synthetic data.

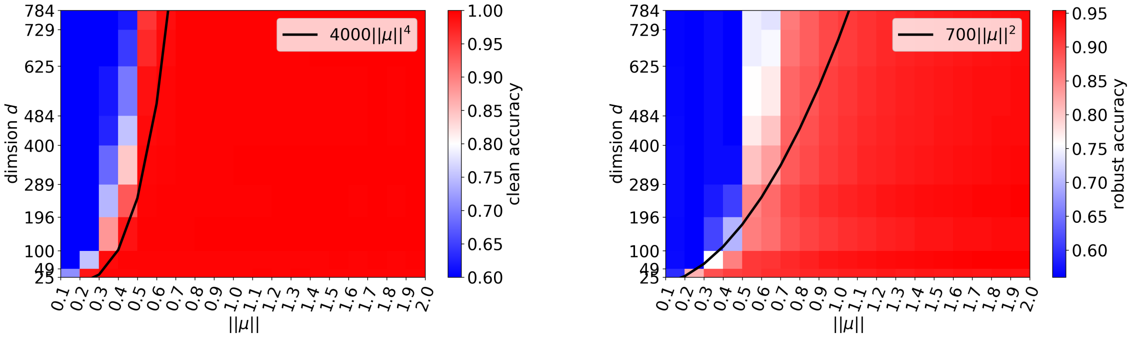

Figure 4. Clean test accuracy (left) / robust test accuracy (right) on MNIST dataset as a function of signal size $\|\mu\|$ and dimension d, for a fixed perturbation ratio $\alpha/\ \|\mu\| = 0.3$

The perturbation ratio is set to $\frac{\alpha}{\|\mu\|} = 0.3$, the signal size $\|\mu\|$ is varied from 0.1 to 2.0. We downsample the images by different factors to simulate data with dimension d ranging between 25 and 784. We plot the heat map for both the clean accuracy and the robust accuracy averaged over ten independent random runs in Figure 4. We observe a phase transition in both subplots at the value of dimension d around $\mathcal{O}(\|\mu\|^4)$ for clean accuracy and $\mathcal{O}(\|\mu\|^2)$ for robust accuracy. This confirms that even when the data distribution deviates from a Gaussian mixture model, the result in Theorem 3.1 is still indicative of an interesting relationship between d, $\|\mu\|$, $\frac{\alpha}{\|\mu\|}$ and n.

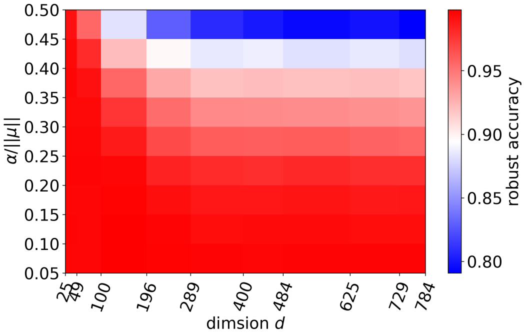

Figure 5. Robust test accuracy (right) as a function of dimension d and perturbation ratio $\frac{\alpha}{\|\mu\|}$, for a fixed signal size $\|\mu\| = 5.0$ on MNIST

As for the MNIST, for a second set of experiments, we fix the signal size to $\|\mu\| = 5.0$ and vary the dimension d from 25 to 784 and perturbation ratios $\frac{\alpha}{\|\mu\|}$ from 0.05 to 0.45. The resulting heat map of robust accuracy averaged over five independent runs is presented in Figure 5. We see a similar trend as in Figure 3.

We also plot the robust training loss and robust test error as a function of the number of training iterations in Figure 6. For the left column, we fix $d = 784$, $\frac{\alpha}{\|\mu\|} = 0.1$, and vary the signal size $\|\mu\| \in [3, 5, 7, 10]$; for the middle column, we fix

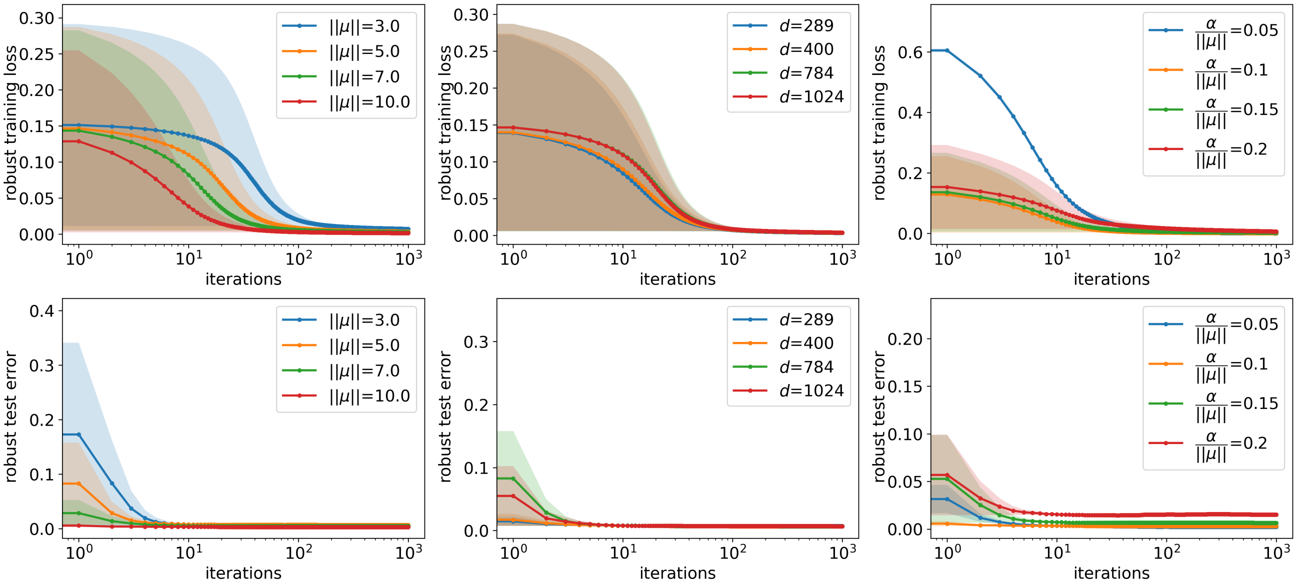

Figure 6. Robust training loss (top) / robust test error (bottom) on MNIST dataset as a function of training iterations. Left column: fix $d = 784$, $\frac{\alpha}{\|\mu\|} = 0.1$. Middle column: fix $\|\mu\| = 5.0$, $\frac{\alpha}{\|\mu\|} = 0.1$. Right column: fix $\|\mu\| = 10.0, d = 784$. Each curve is averaged over 5 runs and shaded regions show standard error.

$\|\mu\| = 5.0$, $\frac{\alpha}{\|\mu\|} = 0.1$, vary dimension $d \in [289, 400, 784, 1024]$; for the right column, we fix $\|\mu\| = 10.0, d = 784$, vary attack rate $\frac{\alpha}{\|\mu\|} \in [0.05, 0.1, 0.15, 0.2]$. We observe that both the robust training loss and the robust test error goes to zero.

B. Missing Proofs

We start by introducing some important notations that will be used throughout our proof. We find the negative derivative of the logistic loss to be useful in our discussion; we denote it as $g(z) := -\ell'(z) = 1/(1 + \exp(z))$. Note that $g(\cdot)$ is non-negative and decreasing and can serve as a surrogate for the 0-1 loss. More importantly, we can check that finding adversarial examples that maximize $\ell(\cdot)$ is equivalent to maximizing $g(\cdot)$, i.e., $\operatorname{argmax}_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W})) = \operatorname{argmax}_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} g(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}))$. For simplicity, we denote $\tilde{\ell}_i(\mathrm{W})$ and $\tilde{\mathrm{g}}_i(\mathrm{W})$ to represent $\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}))$ and $\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} g(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}))$, respectively. The empirical risk and the robust empirical risk w.r.t. the surrogate loss $g(\cdot)$ are denoted as

$$\widehat{G}(\mathrm{W}) := \frac{1}{n}\sum_{i=1}^{n} g(y_i f(\mathrm{x}_i; \mathrm{W})), \quad \widehat{G}_{\mathrm{rob}}(\mathrm{W}) := \frac{1}{n}\sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} g(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W})).$$

B.1. Missing Proofs in Section 3.1

Improvements over (Frei et al., 2022). We have identified two non-rigorous arguments in the proof of (Frei et al., 2022) and addressed them in our analysis. The first issue arises in the Lemma 4.1 of (Frei et al., 2022), where the concentration inequality for the Lipschitz function class (Equation (2) in (Frei et al., 2022)) is applied. However, the expectation should be taken with respect to x instead of (x, y). To resolve this, we introduce Lemma B.2, conditioning on the label y_c, and apply the concentration argument twice. The second issue is found in the proof of Lemma 4.11, Equation 24 in (Frei et al., 2022), specifically in the calculation of $\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[\xi_i \langle y_i \mathrm{x}_i, y_c \mathrm{x}\rangle]$. In their analysis, the expectation is taken only over $\langle y_i \mathrm{x}_i, y_c \mathrm{x}\rangle$, but it should also consider the dependence of ξ_i on x. In our analysis, presented in Lemma B.14, we provide a careful treatment of this expression, incorporating the additional assumption that $d \leq \frac{\|\mu\|^4}{C}$.

Theorem B.1. Let $\varepsilon > 0, \delta \in (0, 1/2)$. $\kappa \in (0, 1)$ and $\lambda > 0$ are defined in Definition 2.1. Let ϕ be a γ-leaky H-smooth activation with $0 \leq \zeta < 1$. Set $\bar{T} = \left(\frac{35+8\sqrt{m/d^3}}{\gamma\|\mu\|\eta\varepsilon}\right)^{\frac{2}{1-\zeta}}$. There exists some constant $C > 0$ such that Assumption 1 and the following holds: (A1) The dimension satisfies $d \leq \|\mu\|^4 / C$. (A2) The signal size satisfies $\|\mu\|^2 \geq C \log(n/\delta)$. Then there exists a constant $c > 0$ such that after running Algorithm 1 for $T \geq \bar{T}$ iterations, we have that with probability at least $1 - 2\delta$ over the random initialization and the draw of an i.i.d. sample of size n, the following holds:

1. The robust training loss satisfies $\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \varepsilon$, the robust training error satisfies $\widehat{L}^{0/1}_{\text{rob}}(\mathrm{W}^T) = 0$.
2. The clean test error satisfies
$$L^{0/1}(\mathrm{W}^T) \leq \beta + 2\exp\Big(-\frac{c\lambda n \|\mu\|^4}{C^2 d}\Big).$$
3. For $\frac{\alpha}{\|\mu\|} \leq \frac{1}{C}\sqrt{\frac{n\|\mu\|^2}{d}}$, the robust test error satisfies
$$L^{0/1}_{\text{rob}}(\mathrm{W}^T) \leq \beta + 2\exp\Big(-c\lambda \|\mu\|^2 \Big(\frac{\|\mu\|}{C}\sqrt{\frac{n}{d}} - \frac{\alpha}{\|\mu\|}\Big)^2\Big).$$

Proof of Theorem B.1. By Lemma B.3 and Lemma B.4, a good run occurs with probability at least $1 - 2\delta$. The robust training loss bound is proved in Lemma B.15. For the robust loss, we apply Lemma B.2 with Lemma B.14, which give us with probability at least $1 - 2\delta$,

$$\begin{aligned}
L^{0/1}_{\text{rob}}(\mathrm{W}^T) &= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}[\exists \tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y \neq \mathrm{sign}(f(\tilde{\mathrm{x}};\mathrm{W}^T))] \\
&\leq \beta + \exp\left(-c\lambda\Big(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c = -1]}{\|\mathrm{W}\|_2} - \alpha\Big)^2\right) \\
&\qquad + \exp\left(-c\lambda\Big(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c = 1]}{\|\mathrm{W}\|_2} - \alpha\Big)^2\right) \\
&\leq \beta + 2\exp\left(-c\lambda\left(\frac{\gamma^2\sqrt{n}}{32C_2\sqrt{d}}\|\mu\|^2 - \alpha\right)^2\right) \\
&\leq \beta + 2\exp\left(-c\lambda\left(\frac{\sqrt{n}}{C\sqrt{d}}\|\mu\|^2 - \alpha\right)^2\right), && \text{(Choose } C \geq \tfrac{32C_2}{\gamma^2}\text{)}
\end{aligned}$$

where the last line holds for $\frac{\alpha}{\|\mu\|} \leq \frac{1}{C}\sqrt{\frac{n\|\mu\|^2}{d}}$, so that $\frac{\sqrt{n}}{C\sqrt{d}}\|\mu\|^2 - \alpha \geq 0$.

Similar for the standard loss, applying Lemma B.2 gives us

$$\begin{aligned}
L^{0/1}(\mathrm{W}^T) &\leq \beta + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = 1]}{\|\mathrm{W}\|_2}\right)^2\right) \\
&\qquad + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = -1]}{\|\mathrm{W}\|_2}\right)^2\right) \\
&\leq \beta + 2\exp\left(-c\lambda\left(\frac{\gamma^2\sqrt{n}}{32C_2\sqrt{d}}\|\mu\|^2\right)^2\right) && \text{(Lemma B.14)} \\
&\leq \beta + 2\exp\left(-c\lambda\left(\frac{\sqrt{n}}{C\sqrt{d}}\|\mu\|^2\right)^2\right) && \text{(Choose } C \geq \tfrac{32C_2}{\gamma^2}\text{)} \\
&= \beta + 2\exp\left(-\frac{c\lambda n\|\mu\|^4}{C^2 d}\right).
\end{aligned}$$

□

The proof of Theorem 3.1 builds upon a sequence of Lemmas, which we show below. Lemma B.2 bound the robust test error by the normalized expected conditional margin via a concentration argument.

Lemma B.2. Suppose that $\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=\bar{y}] - \|\mathrm{W}\|_2\,\alpha \geq 0$ holds for both $\bar{y}=1$ and $\bar{y}=-1$. Then, there exists a universal constant $c>0$ such that

$$L_{\text{rob}}^{0/1}(\mathrm{W}) \leq \beta + \sum_{\bar{y}\in\{-1,+1\}} \exp\Big(-c\lambda\Big(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=\bar{y}]}{\|\mathrm{W}\|_2} - \alpha\Big)^2\Big),$$

$$L^{0/1}(\mathrm{W}) \leq \beta + \sum_{\bar{y}\in\{-1,+1\}} \exp\Big(-c\lambda\Big(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=\bar{y}]}{\|\mathrm{W}\|_2}\Big)^2\Big).$$

Proof of Lemma B.2. We have

$$\begin{aligned}
L_{\text{rob}}^{0/1}(\mathrm{W}) &= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}\left[\exists\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y\neq \operatorname{sign}(f(\tilde{\mathrm{x}};\mathrm{W}))\right] \\
&= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}\left[\exists\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } yf(\tilde{\mathrm{x}};\mathrm{W})\leq 0\right] \\
&\leq \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\exists\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y_c f(\tilde{\mathrm{x}};\mathrm{W})\leq 0\right] \\
&= \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W})\leq 0\right].
\end{aligned}$$

For any $\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)$, we have

$$\begin{aligned}
|y_c f(\mathrm{x};\mathrm{W}) - y_c f(\tilde{\mathrm{x}};\mathrm{W})| &= \frac{1}{\sqrt{m}}\left|\sum_{s=1}^m a_s\left[\phi(\langle \mathrm{w}_s,\mathrm{x}\rangle) - \phi(\langle \mathrm{w}_s,\tilde{\mathrm{x}}\rangle)\right]\right| \\
&\leq \frac{1}{\sqrt{m}}\sum_{s=1}^m |a_s|\,|\langle \mathrm{w}_s,\mathrm{x}-\tilde{\mathrm{x}}\rangle| && (\phi \text{ is 1-Lipschitz}) \\
&\leq \frac{1}{\sqrt{m}}\sqrt{\sum_{s=1}^m a_s^2}\sqrt{\sum_{s=1}^m \langle \mathrm{w}_s,\mathrm{x}-\tilde{\mathrm{x}}\rangle^2} && \text{(Cauchy-Schwartz)} \\
&= \|\mathrm{W}(\mathrm{x}-\tilde{\mathrm{x}})\| \\
&\leq \|\mathrm{W}\|_2\,\alpha. && \text{(By the definition of the spectral norm)}
\end{aligned}$$

Since $\mathcal{D}_{\text{clust}}$ is λ-strongly log concave, and $y_c f(\mathrm{x};\mathrm{W})$ is $\|\mathrm{W}\|_2$-Lipschitz, there is an absolute constant $\bar{c}>0$ such that for any $q\geq 1$, $\|y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}[y_c f(\mathrm{x};\mathrm{W})]\|_{L^q} \leq \bar{c}\,\|\mathrm{W}\|_2\sqrt{q/\lambda}$. Therefore, there is an absolute constant $c>0$ such that for any $t\geq 0$, for fixed $y_c=1$ (same for $y_c=-1$), we have

$$\mathrm{P}\left(y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}[y_c f(\mathrm{x};\mathrm{W})] \leq -t\right) \leq \exp\left(-c\lambda\left(\frac{t}{\|\mathrm{W}\|_2}\right)^2\right). \tag{2}$$

where the expectation is w.r.t. x. Choose $t = \mathbb{E}[y_c f(\mathrm{x};\mathrm{W})] - \|\mathrm{W}\|_2\,\alpha \geq 0$, we have

$$\begin{aligned}
&\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W})\leq 0 \middle| y_c=1\right) \\
&= \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\Bigg(y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W})\middle| y_c=1\right] \leq y_c f(\mathrm{x};\mathrm{W}) \\
&\qquad - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W})\middle| y_c=1\right] - \min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W}) \Bigg| y_c=1\Bigg) \\
&\leq \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\Bigg(y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W})\middle| y_c=1\right] \leq \|\mathrm{W}\|_2\,\alpha \\
&\qquad - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W})\middle| y_c=1\right] \Bigg| y_c=1\Bigg)
\end{aligned}$$

$$\leq \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W})\Big|y_c=1\right]-\left\|\mathrm{W}\right\|_2\alpha}{\left\|\mathrm{W}\right\|_2}\right)^2\right).$$

Consider both $y_c = 1$ and $y_c = -1$ gives us

$$\begin{aligned}
L_{\text{rob}}^{0/1}(\mathrm{W}) &\leq \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right)\\
&= \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\Big|y_c=-1\right)\cdot\mathrm{P}(y_c=1)\\
&\quad + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_c f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\Big|y_c=-1\right)\cdot\mathrm{P}(y_c=-1)\\
&\leq \beta + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})\Big|y_c=1]-\left\|\mathrm{W}\right\|_2\alpha}{\left\|\mathrm{W}\right\|_2}\right)^2\right)\\
&\quad + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})\Big|y_c=-1]-\left\|\mathrm{W}\right\|_2\alpha}{\left\|\mathrm{W}\right\|_2}\right)^2\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&L^{0/1}(\mathrm{W})\\
&= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}[y f(\mathrm{x};\mathrm{W}) \leq 0]\\
&\leq \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W}) \leq 0\right]\\
&\leq \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W}) \leq 0|y_c=1\right] + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W}) \leq 0|y_c=-1\right]\\
&\leq \beta + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=1] \leq -\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=1]|y_c=1\right]\\
&\quad + \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[y_c f(\mathrm{x};\mathrm{W}) - \mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=-1] \leq -\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})|y_c=-1]|y_c=-1\right]\\
&\leq \beta + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})\big|y_c=1]}{\left\|\mathrm{W}\right\|_2}\right)^2\right) + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W})\big|y_c=-1]}{\left\|\mathrm{W}\right\|_2}\right)^2\right).
\end{aligned}$$

□

Now we only need to derive a lower bound on the normalized expected conditional margin. Below is a series of structural results that leads us to our destination. Lemma B.3 and B.4 are the properties of initialized network weights as well as the generated data.

Lemma B.3. Under Assumption 1, (A1) and (A2), there is a universal constant $C_0 > 1$ such that with probability at least $1 - \delta/2$ over the random initialization,

$$\frac{1}{2}\omega_{\text{init}}^2 d \leq \left\|\mathrm{w}_s^0\right\|_2^2 \leq \frac{3}{2}\omega_{\text{init}}^2 d, \forall s\in[m]; \left\|\mathrm{W}^0\right\|_2 \leq C_0\omega_{\text{init}}(\sqrt{m}+\sqrt{d})$$

Proof of lemma B.3. For any fixed s, note that $\left\|\mathrm{w}_s^0\right\|_2^2$ is a ω_{init}^2-multiple of a chi-squared random variable with d degrees of freedom. By concentration of the χ^2 distribution, for any $t\in(0,1]$,

$$\mathrm{P}\left(\left|\frac{1}{d\omega_{\text{init}}^2}\left\|\mathrm{w}_s^0\right\|_2^2 - 1\right| \geq t\right) \leq 2\exp\left(-dt^2/8\right).$$

In particular, if we choose $t = 1/2$, with probability at least $1 - 2\exp\left(-d/32\right)$, we have

$$\frac{1}{2}\omega_{\text{init}}^2 d \le \left\|\mathrm{w}_s^0\right\|_2^2 \le \frac{3}{2}\omega_{\text{init}}^2 d.$$

Applying a union bound, with probability at least $1 - 2m\exp\left(-d/32\right)$, we have

$$\frac{1}{2}\omega_{\text{init}}^2 d \le \left\|\mathrm{w}_s^0\right\|_2^2 \le \frac{3}{2}\omega_{\text{init}}^2 d, \forall s \in [m].$$

Note that

$$\begin{aligned}
&1 - 2m\exp\left(-d/32\right) \\
&\ge 1 - 2\delta\exp\left(n/C - d/32\right) && \text{(Assumption (5))} \\
&\ge 1 - 2\delta\exp\left(-d/64\right) && (d \ge 64n \text{ from Assumption (1), Assumption (A2) and } C \text{ sufficiently large}) \\
&\ge 1 - \delta/4. && (d \ge 192 \text{ from Assumption (1), Assumption (A2) and } C \text{ sufficiently large})
\end{aligned}$$

Therefore $\frac{1}{2}\omega_{\text{init}}^2 d \le \left\|\mathrm{w}_s^0\right\|_2^2 \le \frac{3}{2}\omega_{\text{init}}^2 d, \forall s \in [m]$ holds with probability at least $1 - \delta/4$.
For the spectral norm, since the entries of $[\mathrm{w}_1, \dots, \mathrm{w}_{\frac{m}{2}}]/\omega_{\text{init}}$ are i.i.d. standard normal variables, by Theorem 4.4.5 in (Vershynin, 2018), there exists a universal constant $c > 0$ such that for any $u \ge 0$, with probability at least $1 - 2\exp\left(-u^2\right)$, we have

$$\left\|[\mathrm{w}_1, \dots, w_{\frac{m}{2}}]\right\|_2 \le c\omega_{\text{init}}(\sqrt{m/2} + \sqrt{d} + u).$$

In particular, taking $u = \sqrt{\log\left(8/\delta\right)}$, we have with probability at least $1 - \delta/4$, $\left\|[\mathrm{w}_1, \dots, \mathrm{w}_{\frac{m}{2}}]\right\|_2 \le c\omega_{\text{init}}(\sqrt{m/2} + \sqrt{d} + \sqrt{\log\left(8/\delta\right)})$. Since $\left\|\mathrm{W}^0\right\|_2 = \sqrt{2}\left\|[\mathrm{w}_1, \dots, \mathrm{w}_{\frac{m}{2}}]\right\|_2$ holds by symmetric initialization, and $\sqrt{d} \ge \sqrt{C\log\left(1/\delta\right)} \ge \sqrt{\log\left(8/\delta\right)}$ by Assumption (1) and C sufficiently large, we are done with the spectral norm. □

Lemma B.4. Let $(\mathrm{x}_i, y_i) \sim \mathcal{D}, \forall i \in [n]$, where $\mathrm{x}_i = y_i^c\mu + \xi_i$, $\mathrm{P}(y_i^c \ne y_i) \le \beta$. Given $0 < \kappa < 1$ in Definition 2.1, there exists $C_1 = \frac{10}{\kappa} > 1$ such that for large enough C, with probability at least $1 - \delta$ over $\mathcal{D}^n$, the following hold

(C1) $\forall i \in [n]$, $\frac{\kappa d}{2} \le \|\xi_i\|^2 \le (3 + \frac{\kappa}{2})d$, $d/C_1 \le \|\mathrm{x}_i\|^2 \le C_1 d$; $\forall \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_i' \in \mathcal{B}_2(\mathrm{x}_i, \alpha)$, $d/(4C_1) \le \left(\sqrt{d/C_1} - \alpha\right)^2 \le \|\tilde{\mathrm{x}}_i\|^2 \le \left(\sqrt{C_1 d} + \alpha\right)^2 \le 4C_1 d$, $\langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_i' \rangle \ge \left(\sqrt{d/C_1} - \alpha\right)^2 \ge d/(4C_1)$.

(C2) $\forall i \ne j \in [n], |\langle \xi_i, \xi_j \rangle| \le C_1\left(\sqrt{d\log\left(n/\delta\right)}\right), |\langle \mathrm{x}_i, \mathrm{x}_j \rangle| \le C_1\left(\|\mu\|^2 + \sqrt{d\log\left(n/\delta\right)}\right)$.

$\forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \tilde{\mathrm{x}}_j \in \mathcal{B}_2(\mathrm{x}_j, \alpha), |\langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_j \rangle| \le C_1\left(\|\mu\|^2 + \sqrt{d\log\left(n/\delta\right)}\right) + 2\alpha\sqrt{C_1 d} + \alpha^2$,

$|\langle \mathrm{x}_i, \tilde{\mathrm{x}}_j \rangle| \le C_1\left(\|\mu\|^2 + \sqrt{d\log\left(n/\delta\right)}\right) + \alpha\sqrt{C_1 d}$.

(C3) $\forall z_1, z_2, \dots, z_n \in \mathbb{R}, \left\|\sum_{i=1}^n z_i\xi_i\right\|^2 \le 4d\sum_{i=1}^n z_i^2$.

(C4) $\forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \forall z_1, z_2, \dots, z_n \in \mathbb{R}, \frac{d}{8C_1} \cdot \sum_{i=1}^n z_i^2 \le \left\|\sum_{i=1}^n z_i\tilde{\mathrm{x}}_i\right\|^2 \le 8C_1 d \cdot \sum_{i=1}^n z_i^2$.

(C5) $\forall i \in \mathcal{C}, \left|\langle \mu, y_i\mathrm{x}_i \rangle - \|\mu\|^2\right| \le \|\mu\|^2/2, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \frac{1}{2}\|\mu\|^2 - \|\mu\|\alpha \le \langle \mu, y_i\tilde{\mathrm{x}}_i \rangle \le \frac{3}{2}\|\mu\|^2 + \|\mu\|\alpha$.

(C6) $\forall i \in \mathcal{N}, \left|\langle \mu, y_i\mathrm{x}_i \rangle + \|\mu\|^2\right| \le \|\mu\|^2/2, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), -\frac{3}{2}\|\mu\|^2 - \|\mu\|\alpha \le \langle \mu, y_i\tilde{\mathrm{x}}_i \rangle \le -\frac{1}{2}\|\mu\|^2 + \|\mu\|\alpha$.

(C7) The number of noisy samples satisfies $|\mathcal{N}|/n \le \beta + \sqrt{\frac{2}{C}}$.

Proof of Lemma B.4. The proof is a simple extension of Lemma 13 in (Chatterji & Long, 2021). For statement (C1), $\frac{\kappa d}{2} \le \|\xi_i\|^2 \le (3+\frac{\kappa}{2})d$ and $d/C_1 \le \|\mathrm{x}_i\|^2 \le C_1 d$ follows directly from the proof of Lemma 19 in (Chatterji & Long, 2021). Since

$$\begin{aligned} \|\mathrm{x}_i - \tilde{\mathrm{x}}_i\| &\le \alpha \le \|\mu\| && \text{(Assumption (6))} \\ &\le \frac{1}{2}\sqrt{\frac{d}{C_1}} && (d \ge C\|\mu\|^2 n \text{ from Assumption (1) with sufficiently large } C) \\ &\le \frac{1}{2}\|\mathrm{x}_i\|, \end{aligned}$$

$d/(4C_1) \le \left(\sqrt{d/C_1}-\alpha\right)^2 \le \|\tilde{\mathrm{x}}_i\|^2 \le \left(\sqrt{C_1 d}+\alpha\right)^2 \le 4C_1 d$ holds. Because $\left\|\mathrm{x}_i - \tilde{\mathrm{x}}_i'\right\| \le \alpha \le \frac{1}{2}\|\mathrm{x}_i\|$ also holds, through some simple calculation, we have $\left\langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_i'\right\rangle \ge (\|\mathrm{x}_i\|-\alpha)^2 \ge \left(\sqrt{d/C_1}-\alpha\right)^2 \ge d/(4C_1)$.

For statement (C2), $|\langle \xi_i, \xi_j\rangle| \le C_1\left(\sqrt{d\log(n/\delta)}\right)$ and $|\langle \mathrm{x}_i, \mathrm{x}_j\rangle| \le C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right)$ follows directly from the proof of Lemma 20 in (Chatterji & Long, 2021).

$$\begin{aligned} |\langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_j\rangle| &= |\langle \mathrm{x}_i, \mathrm{x}_j\rangle + \langle \tilde{\mathrm{x}}_i - \mathrm{x}_i, \mathrm{x}_j\rangle + \langle \mathrm{x}_i, \tilde{\mathrm{x}}_j - \mathrm{x}_j\rangle + \langle \tilde{\mathrm{x}}_i - \mathrm{x}_i, \tilde{\mathrm{x}}_j - \mathrm{x}_j\rangle| \\ &\le |\langle \mathrm{x}_i, \mathrm{x}_j\rangle| + \alpha\|\mathrm{x}_i\| + \alpha\|\mathrm{x}_j\| + \|\tilde{\mathrm{x}}_i - \mathrm{x}_i\| \cdot \|\tilde{\mathrm{x}}_j - \mathrm{x}_j\| \\ &\le C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right) + 2\alpha\sqrt{C_1 d} + \alpha^2. \end{aligned}$$

$$\begin{aligned} |\langle \mathrm{x}_i, \tilde{\mathrm{x}}_j\rangle| &= |\langle \mathrm{x}_i, \mathrm{x}_j\rangle + \langle \mathrm{x}_i, \tilde{\mathrm{x}}_j - \mathrm{x}_j\rangle| \\ &\le |\langle \mathrm{x}_i, \mathrm{x}_j\rangle| + \alpha\|\mathrm{x}_i\| \\ &\le C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right) + \alpha\sqrt{C_1 d}. \end{aligned}$$

The statement (C3) holds since

$$\begin{aligned} \left\|\sum_{i=1}^n z_i \xi_i\right\|^2 &= \sum_{i=1}^n z_i^2\|\xi_i\|^2 + 2\sum_{i<j} z_i z_j \langle \xi_i, \xi_j\rangle \\ &\le \sum_{i=1}^n z_i^2(3+\frac{\kappa}{2})d + 2C_1\sum_{i<j}\frac{z_i^2+z_j^2}{2}\sqrt{d\log(n/\delta)} && \text{((C1), (C2))} \\ &\le \sum_{i=1}^n z_i^2(3+\frac{\kappa}{2})d + 2C_1\sum_{i<j}\frac{z_i^2+z_j^2}{2}\frac{d}{\sqrt{C}n} && \text{(Assumption (1))} \\ &\le (3+\frac{\kappa}{2}+\frac{C_1}{\sqrt{C}})d\sum_{i=1}^n z_i^2 \le 4d\sum_{i=1}^n z_i^2. && (C \text{ sufficiently large}) \end{aligned}$$

The statement (C4) holds since

$$\begin{aligned} \left\|\sum_{i=1}^n z_i \tilde{\mathrm{x}}_i\right\|^2 &= \sum_{i=1}^n z_i^2\|\tilde{\mathrm{x}}_i\|^2 + 2\sum_{i<j} z_i z_j \langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_j\rangle \\ &\le \sum_{i=1}^n z_i^2 4C_1 d + 2\sum_{i<j}\frac{z_i^2+z_j^2}{2}\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right) && \text{((C1), (C2))} \\ &\le \left(4C_1 d + n\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\right)\sum_{i=1}^n z_i^2 \\ &\le 8C_1 d\sum_{i=1}^n z_i^2 && \text{(Assumption (1), } C \text{ sufficiently large)} \end{aligned}$$

and

$$
\begin{aligned}
\left\|\sum_{i=1}^{n} z_i \tilde{\mathrm{x}}_i\right\|^2 &= \sum_{i=1}^{n} z_i^2 \left\|\tilde{\mathrm{x}}_i\right\|^2 + 2\sum_{i<j} z_i z_j \left\langle \tilde{\mathrm{x}}_i, \tilde{\mathrm{x}}_j \right\rangle \\
&\geq \sum_{i=1}^{n} z_i^2 \frac{d}{4C_1} - 2\sum_{i<j} \frac{z_i^2 + z_j^2}{2} \left(C_1 \left(\|\mu\|^2 + \sqrt{d \log(n/\delta)} \right) + 2\alpha\sqrt{C_1 d} + \alpha^2 \right) && \text{((C1), (C2))} \\
&\geq \left(\frac{d}{4C_1} - n \left(C_1 \left(\|\mu\|^2 + \sqrt{d \log(n/\delta)} \right) + 2\alpha\sqrt{C_1 d} + \alpha^2 \right) \right) \sum_{i=1}^{n} z_i^2 \\
&\geq \frac{d}{8C_1} \sum_{i=1}^{n} z_i^2. && \text{(Assumption (1), } C \text{ sufficiently large)}
\end{aligned}
$$

Statement (C5) and statement (C6) follow similarly from the proof of Lemma 21 and 22 in (Chatterji & Long, 2021) and combining the fact that

$$
\left| \langle \mu, y_i \mathrm{x}_i \rangle - \langle \mu, y_i \tilde{\mathrm{x}}_i \rangle \right| \leq \|\mu\| \cdot \|\mathrm{x}_i - \tilde{\mathrm{x}}_i\| \leq \|\mu\| \, \alpha.
$$

The last statement follows from Hoeffding's inequality:

$$
\begin{aligned}
&\mathrm{P}(|\mathcal{N}|/n - \beta > \sqrt{\frac{2}{C}}) \\
&\leq e^{-2n(\sqrt{\frac{2}{C}})^2} \\
&\leq e^{-2C \log(1/\delta)(\frac{2}{C})} && \text{(Assumption (5))} \\
&= \delta^4 \\
&\leq \delta/6. && (\delta < 0.5)
\end{aligned}
$$

□

Definition B.5. If the events in Lemma B.3 and Lemma B.4 occur, let us say that we have a good run.

Lemma B.3 and Lemma B.4 show that a good run occurs with probability at least $1 - 2\delta$. In the following we assume a good run occurs. Lemma B.6 leverages the smoothness property of activation function and derive the result via Taylor approximation. Lemma B.7 characterizes the relationship between $\widehat{L}_{\text{rob}}(\cdot)$ and $\widehat{G}_{\text{rob}}(\cdot)$. Lemma B.8 further derives the bounds on the gradient norm given adversarial training example, as well as the pairwise correlations of the gradients given different adversarial training examples. These are standard results that have been derived by (Frei et al., 2022), and we simply extend them for adversarial training scenario.

Lemma B.6 (Lemma 4.5 in (Frei et al., 2022))**.** For an H-smooth activation ϕ and any $\mathrm{W}, \mathrm{V} \in \mathbb{R}^{m \times d}$, and $\mathrm{x} \in \mathbb{R}^d$, we have

$$
\left| f(\mathrm{x}; \mathrm{W}) - f(\mathrm{x}; \mathrm{V}) - \langle \nabla f(\mathrm{x}; \mathrm{V}), \mathrm{W} - \mathrm{V} \rangle \right| \leq \frac{H \|\mathrm{x}\|^2}{2\sqrt{m}} \|\mathrm{W} - \mathrm{V}\|_2^2 .
$$

Lemma B.7. Let $C_1 > 1$ be the constant from Lemma B.4. For an H-smooth activation ϕ and any $\mathrm{W}, \mathrm{V} \in \mathbb{R}^{m \times d}$, on a good run it holds that

$$
\frac{1}{\sqrt{C_1 d} + \alpha} \left\| \nabla \widehat{L}_{\text{rob}}(\mathrm{W}) \right\|_F \leq \widehat{G}_{\text{rob}}(\mathrm{W}) \leq \widehat{L}_{\text{rob}}(\mathrm{W}) \wedge 1.
$$

Proof of Lemma B.7. Since ϕ is 1-Lipschitz, we have $\forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)$,

$$
\|\nabla f(\tilde{\mathrm{x}}_i; \mathrm{W})\|_F^2 = \frac{1}{m} \sum_{s=1}^{m} \|a_s \phi'(\langle \mathrm{w}_s, \tilde{\mathrm{x}}_i \rangle) \tilde{\mathrm{x}}_i\|^2 \leq \left(\sqrt{C_1 d} + \alpha \right)^2 . \tag{3}
$$

For $\forall i \in [n]$, choose $\tilde{\mathrm{x}}_i = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}))$ so that $\tilde{\ell}_i(\mathrm{W}) = \ell(y_i f(\tilde{\mathrm{x}}_{;} \mathrm{W}))$, $\tilde{g}_i(\mathrm{W}) = g(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}))$, we have

$$\begin{aligned}
\left\| \nabla \widehat{L}_{\text{rob}}(\mathrm{W}) \right\|_F &= \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}) y_i \nabla f(\tilde{\mathrm{x}}_i; \mathrm{W}) \right\|_F \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}) \left\| \nabla f(\tilde{\mathrm{x}}_i; \mathrm{W}) \right\|_F && \text{(Jensen's inequality)} \\
&\leq \frac{\sqrt{C_1 d} + \alpha}{n} \sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}) && \text{(Equation (3))} \\
&\leq \frac{\sqrt{C_1 d} + \alpha}{n} \sum_{i=1}^{n} \min\left(\tilde{\ell}_i(\mathrm{W}), 1 \right) && \text{(By the definition of } \tilde{g}_i(\mathrm{W}) \text{ and } \tilde{\ell}_i(\mathrm{W})\text{)} \\
&\leq \left(\sqrt{C_1 d} + \alpha \right) \left(\widehat{L}_{\text{rob}}(\mathrm{W}) \wedge 1 \right). && \text{(Jensen's inequality)}
\end{aligned}$$

□

Lemma B.8. Let $C_1 > 1$ be the constant from Lemma B.4. For a γ-leaky, H-smooth activation ϕ, on a good run, for any $i, j \in [n], i \neq j, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \forall \tilde{\mathrm{x}}_i' \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \forall \tilde{\mathrm{x}}_j \in \mathcal{B}_2(\mathrm{x}_j, \alpha)$, we have

$$\begin{aligned}
\left| \langle \nabla f(\mathrm{x}_i, \mathrm{W}), \nabla f(\tilde{\mathrm{x}}_j, \mathrm{W}) \rangle \right| &\leq C_1 \left(\|\mu\|^2 + \sqrt{d \log{(n/\delta)}} \right) + \alpha \sqrt{C_1 d}, \\
\left| \langle \nabla f(\tilde{\mathrm{x}}_i, \mathrm{W}), \nabla f(\tilde{\mathrm{x}}_j, \mathrm{W}) \rangle \right| &\leq C_1 \left(\|\mu\|^2 + \sqrt{d \log{(n/\delta)}} \right) + 2\alpha \sqrt{C_1 d} + \alpha^2.
\end{aligned}$$

Moreover, for any $i \in [n]$ and any $\mathrm{W} \in \mathbb{R}^{m \times d}$, we have

$$\begin{aligned}
\left(\sqrt{d/C_1} - \alpha \right)^2 \gamma^2 &\leq \left\| \nabla f(\tilde{\mathrm{x}}_i; \mathrm{W}) \right\|_F^2 \leq \left(\sqrt{C_1 d} + \alpha \right)^2, \\
\left(\sqrt{d/C_1} - \alpha \right)^2 \gamma^2 &\leq \left| \langle \nabla f(\tilde{\mathrm{x}}_i, \mathrm{W}), \nabla f(\tilde{\mathrm{x}}_i', \mathrm{W}) \rangle \right| \leq \left(\sqrt{C_1 d} + \alpha \right)^2.
\end{aligned}$$

Proof of Lemma B.8. The proof is similar as Lemma 4.7 in (Frei et al., 2022).

$$\langle \nabla f(\mathrm{x}, \mathrm{W}), \nabla f(\mathrm{y}, \mathrm{W}) \rangle = \frac{1}{m} \langle \mathrm{x}, \mathrm{y} \rangle \sum_{s=1}^{m} \phi'(\langle \mathrm{w}_s, x \rangle) \phi'(\langle \mathrm{w}_s, y \rangle).$$

Therefore,

$$\left| \langle \nabla f(\mathrm{x}, \mathrm{W}), \nabla f(\mathrm{y}, \mathrm{W}) \rangle \right| = \frac{1}{m} \left| \langle \mathrm{x}, \mathrm{y} \rangle \right| \sum_{s=1}^{m} \phi'(\langle \mathrm{w}_s, x \rangle) \phi'(\langle \mathrm{w}_s, y \rangle) \in [\gamma^2 \left| \langle \mathrm{x}, \mathrm{y} \rangle \right|, \left| \langle \mathrm{x}, \mathrm{y} \rangle \right|].$$

Thus, the first two inequalities follow from Lemma B.4 (C2). The last two inequalities follow from Lemma B.4 (C1). □

Lemma B.9 plays a crucial role in our analysis. It demonstrates that the margin increases with each epoch of adversarial training, given any adversarial examples. More importantly, it proves the loss g is at the same scale across all adversarial training examples.

Lemma B.9. For a γ-leaky, H-smooth activation ϕ, there is a constant $C_r = \frac{64 C_1 \left(\sqrt{C_1} + 0.5 \sqrt{\frac{1}{C_1}} \right)^2}{\gamma^2}$ such that on a good run, provided $C > 1$ is sufficiently large, we have for all $t \geq 0$,

$$\begin{aligned}
y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^{t+1}) \geq y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^t) \geq 0, \forall \tilde{\mathrm{x}}_k \in \mathcal{B}_2(\mathrm{x}_k, \alpha), \forall k \in [n], \\
\max_{i,j \in [n]} \frac{g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t))} \leq \frac{16 \left(\sqrt{C_1 d} + \alpha \right)^2}{\gamma^2 \left(\sqrt{d/C_1} - \alpha \right)^2} \leq C_r,
\end{aligned}$$

where $\tilde{\mathrm{x}}_i^t = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} \ell(y_i f(\tilde{\mathrm{x}}; \mathrm{W}^t))$, $\tilde{\mathrm{x}}_j^t = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_j, \alpha)} \ell(y_j f(\tilde{\mathrm{x}}; \mathrm{W}^t))$.

Proof of Lemma B.9. By Fact A.2 in (Frei et al., 2022), we have

$$\frac{g(x)}{g(y)} \le \max\left(2, 2\frac{\exp(-x)}{\exp(-y)}\right)$$

holds for any $x, y \in \mathbb{R}$, so

$$\max_{i,j\in[n]} \frac{g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t))} \le \max\left(2, 2 \cdot \max_{i,j\in[n]} \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^t; \mathrm{W}^t)\right)}\right).$$

In the remainder of the proof we will show that the ratio of the exponential losses is bounded. We will prove it by induction. Since a good run occurs, all the events in Lemma B.3 and Lemma B.4 occurs. In particular, we have $\left\|\mathrm{W}^0\right\|_2 \le C_0\omega_{\text{init}}(\sqrt{m}+\sqrt{d})$ and $\left\|\tilde{\mathrm{x}}_i^0\right\| \le \sqrt{C_1 d} + \alpha$. Note that at initialization, we have $|f(\tilde{\mathrm{x}}_i^0; \mathrm{W}^0)| = 0$. For any $\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \tilde{\mathrm{x}}_j \in \mathcal{B}_2(\mathrm{x}_j, \alpha)$, consider $t = 0$, we have

$$\max_{i,j\in[n]} \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^0; \mathrm{W}^0)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^0; \mathrm{W}^0)\right)} = 1 \le \frac{8\left(\sqrt{C_1 d} + \alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2}.$$

Assume the result holds at time t and consider the case $t+1$. For simplicity we only consider the exponential ratio for the first sample and the second sample, and denote $A_t := \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^t; \mathrm{W}^t)\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^t; \mathrm{W}^t)\right)}$. Then $A_t \le \frac{8\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}$. Fix $k \in [n]$, consider $\forall \tilde{\mathrm{x}}_k \in \mathcal{B}(\mathrm{x}_k, \alpha)$, define $\tilde{\rho}_i^t = \frac{1}{m}\sum_{s=1}^m \phi'\left(\left\langle \mathrm{w}_s^{(t)}, \tilde{\mathrm{x}}_k \right\rangle\right) \phi'\left(\left\langle \mathrm{w}_s^{(t)}, \tilde{\mathrm{x}}_i^t \right\rangle\right) \in [\gamma^2, 1]$. We first need to show that $y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^{t+1}) \ge y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^t)$.

$$\begin{aligned}
& y_k\left[f(\tilde{\mathrm{x}}_k; \mathrm{W}^{t+1}) - f(\tilde{\mathrm{x}}_k; \mathrm{W}^t)\right] \\
& \ge y_k\left[\left\langle \nabla f(\tilde{\mathrm{x}}_k; \mathrm{W}^t), \mathrm{W}^{t+1} - \mathrm{W}^t \right\rangle\right] - \frac{H\left\|\tilde{\mathrm{x}}_k\right\|^2}{2\sqrt{m}}\left\|\mathrm{W}^{t+1} - \mathrm{W}^t\right\|_2^2 && \text{(for } y \in \{\pm 1\}\text{, apply Lemma B.6)} \\
& = y_k \eta\left[\left\langle \nabla f(\tilde{\mathrm{x}}_k; \mathrm{W}^t), \frac{1}{n}\sum_{i=1}^n \tilde{g}_i(\mathrm{W}^t) y_i \nabla f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right\rangle\right] - \frac{H\left\|\tilde{\mathrm{x}}_k\right\|^2 \eta^2}{2\sqrt{m}}\left\|\nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_2^2 \\
& \ge \eta\left[\frac{1}{n}\sum_{i=1}^n \tilde{g}_i(\mathrm{W}^t)\left\langle y_k \nabla f(\tilde{\mathrm{x}}_k; \mathrm{W}^t), y_i \nabla f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)\right\rangle\right] - \frac{H\left(\sqrt{C_1 d} + \alpha\right)^4 \eta^2}{2\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) && \text{(Lemma B.7)} \\
& = \frac{\eta}{n}\left(\tilde{g}_k(\mathrm{W}^t)\tilde{\rho}_k^t\left\langle \tilde{\mathrm{x}}_k^t, \tilde{\mathrm{x}}_k\right\rangle + \sum_{i\ne k}\tilde{g}_i(\mathrm{W}^t)\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y_k\tilde{\mathrm{x}}_k\right\rangle\right) - \frac{H\left(\sqrt{C_1 d} + \alpha\right)^4 \eta^2}{2\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
& \ge \frac{\eta}{n}\left[\tilde{g}_k(\mathrm{W}^t)\left(\gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2 - \frac{\max_j \tilde{g}_j(\mathrm{W}^t)}{\tilde{g}_k(\mathrm{W}^t)}\sum_{i\ne k}\left(C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right) + 2\alpha\sqrt{C_1 d} + \alpha^2\right)\right)\right] && \text{(Lemma B.4)} \\
& \quad - \frac{H\left(\sqrt{C_1 d} + \alpha\right)^4 \eta^2}{2\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
& \ge \frac{\eta}{n}\left[\tilde{g}_k(\mathrm{W}^t)\left(\gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2 - C_r n\left(C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right) + 2\alpha\sqrt{C_1 d} + \alpha^2\right)\right)\right] \\
& && \text{(By induction, } \tfrac{\max_j \tilde{g}_j(\mathrm{W}^t)}{\tilde{g}_k(\mathrm{W}^t)} \le C_r\text{)} \\
& \quad - \frac{H\left(\sqrt{C_1 d} + \alpha\right)^4 \eta^2}{2\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
& \ge \frac{\eta\gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2}{2n}\tilde{g}_k(\mathrm{W}^t) - \frac{H\left(\sqrt{C_1 d} + \alpha\right)^4 \eta^2}{2\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
& && \text{(this line holds with large enough } C \text{ via Assumption (1))}
\end{aligned}$$

$$
\begin{aligned}
&\geq \eta \widehat{G}_{\text{rob}}(\mathrm{W}^t)\left(\frac{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}{2nC_r}-\frac{H\left(\sqrt{C_1 d}+\alpha\right)^4\eta}{2\sqrt{m}}\right) && \text{(By induction, } \tilde{g}_k(\mathrm{W}^t)\geq \tfrac{1}{C_r}\widehat{G}_{\text{rob}}(\mathrm{W}^t)\text{)}\\
&\geq \eta \widehat{G}_{\text{rob}}(\mathrm{W}^t)\left(\frac{\gamma^2\left(\frac{1}{2}\sqrt{d/C_1}\right)^2}{2nC_r}-\frac{H\left(2\sqrt{C_1 d}\right)^4\eta}{2\sqrt{m}}\right) && \text{(Assumption (6) and Assumption (1))}\\
&\geq \eta \widehat{G}_{\text{rob}}(\mathrm{W}^t)\left(\frac{\gamma^2\left(\frac{1}{2}\sqrt{d/C_1}\right)^2}{2nC_r}-\frac{H\left(2\sqrt{C_1 d}\right)^4}{2Cd^2}\right) && \text{(Assumption (4))}\\
&\geq 0, && (4)
\end{aligned}
$$

where the last line holds from Assumption (1) with sufficiently large C.

Now we are back to prove the upper bound of the exponential ratio A_t. We have

$$
\begin{aligned}
A_{t+1} &= \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t+1})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t+1};\mathrm{W}^{t+1})\right)}\\
&= \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t};\mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t})\right)}\cdot\frac{\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t};\mathrm{W}^{t})-y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t+1})\right)}{\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t})-y_2 f(\tilde{\mathrm{x}}_2^{t+1};\mathrm{W}^{t+1})\right)}\\
&\leq A_t\cdot\frac{\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t})-y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t+1})\right)}{\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t})-y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t+1})\right)} && \text{(By the definition of } \tilde{\mathrm{x}}_1^t,\tilde{\mathrm{x}}_2^{t+1}\text{)}\\
&= A_t\cdot\frac{\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t})-y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t}-\frac{\eta}{n}\sum_{i=1}^n\nabla\ell(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)))\right)}{\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t})-y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t}-\frac{\eta}{n}\sum_{i=1}^n\nabla\ell(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)))\right)}\\
&\leq A_t\cdot\frac{\exp\left(-\frac{\eta}{n}\sum_{i=1}^n y_1 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)}{\exp\left(-\frac{\eta}{n}\sum_{i=1}^n y_2 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)} && \text{(Lemma B.6)}\\
&\qquad\cdot\exp\left(\frac{2H\left(\sqrt{C_1 d}+\alpha\right)^2\eta^2}{\sqrt{m}}\left\|\frac{1}{n}\sum_{i=1}^n\nabla\ell(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\right\|^2\right)\\
&= A_t\cdot\exp\Big(-\frac{\eta}{n}g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t)\right\rangle\\
&\qquad\qquad+\frac{\eta}{n}g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t)\right\rangle\Big)\\
&\qquad\cdot\frac{\exp\left(-\frac{\eta}{n}\sum_{i\neq 1} y_1 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)}{\exp\left(-\frac{\eta}{n}\sum_{i\neq 2} y_2 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)}\\
&\qquad\cdot\exp\left(\frac{2H\left(\sqrt{C_1 d}+\alpha\right)^2\eta^2}{\sqrt{m}}\left\|\frac{1}{n}\sum_{i=1}^n\nabla\ell(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\right\|^2\right),
\end{aligned}
$$

where the first inequality holds since $\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t)\right)\geq\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t};\mathrm{W}^t)\right)$, $\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t+1};\mathrm{W}^{t+1})\right)\leq\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t+1})\right)$ by the definition of $\tilde{\mathrm{x}}_1^t,\tilde{\mathrm{x}}_2^{t+1}$.

We next bound each of the above term separately. For the first term, we have

$$
\begin{aligned}
&\exp\left(-\frac{\eta}{n}g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t)\right\rangle+\frac{\eta}{n}g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\left\langle\nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t)\right\rangle\right)\\
&=\exp\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\eta}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))}\left\langle\nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t)\right\rangle-\left\langle\nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t),\nabla f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t)\right\rangle\right)\right)\\
&\leq\exp\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\gamma^2\eta}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))}\left(\sqrt{d/C_1}-\alpha\right)^2-\frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2}\right)\right) && \text{(Lemma B.8)}
\end{aligned}
$$

$$= \exp\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t; \mathrm{W}^t))\eta\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} - \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\right)\right).$$

For the second term, we have

$$\begin{aligned}
&\frac{\exp\left(-\frac{\eta}{n}\sum_{i\neq 1} y_1 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle \nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t), \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)}{\exp\left(-\frac{\eta}{n}\sum_{i\neq 2} y_2 y_i g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left\langle \nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t), \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right)}\\
&\leq \exp\Bigg(\frac{\eta}{n}\sum_{i\neq 1} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left|\left\langle \nabla f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^t), \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right|\\
&\qquad + \frac{\eta}{n}\sum_{i\neq 2} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left|\left\langle \nabla f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^t), \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right|\Bigg)\\
&\leq \exp\left(\frac{2\eta}{n}\sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left(C_1\left\|\mu\right\|^2 + C_1\sqrt{d\log\left(n/\delta\right)} + 2\alpha\sqrt{C_1 d}+\alpha^2\right)\right). \qquad \text{(Lemma B.8)}
\end{aligned}$$

For the third term, we have

$$\begin{aligned}
&\exp\left(\frac{2H\left(\sqrt{C_1 d}+\alpha\right)^2\eta^2}{\sqrt{m}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\ell(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\right\|^2\right)\\
&\leq \exp\left(\frac{2H\left(\sqrt{C_1 d}+\alpha\right)^4\eta^2}{\sqrt{m}}\cdot\frac{1}{n}\sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\right) \qquad \text{(Lemma B.7)}\\
&\leq \exp\left(\frac{\eta}{n}\sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\right). \qquad \text{(Large enough } C \text{ for assumption (4))}
\end{aligned}$$

Combining the above results gives us that

$$\begin{aligned}
A_{t+1} \leq A_t \exp&\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t; \mathrm{W}^t))\eta\gamma^2\left(\sqrt{\frac{d}{C_1}}-\alpha\right)^2}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} - \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{\frac{d}{C_1}}-\alpha\right)^2}\right)\right)\\
\cdot \exp&\left(\frac{2\eta}{n}\sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left(C_1\left\|\mu\right\|^2 + 2C_1\sqrt{d\log\left(n/\delta\right)} + 2\alpha\sqrt{C_1 d}+\alpha^2\right)\right).
\end{aligned}$$

Now consider the following two cases. If $\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} \leq \frac{2\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}$, then we have

$$\begin{aligned}
A_{t+1} &\leq A_t\cdot\exp\left(\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\eta(\sqrt{C_1 d}+\alpha)^2}{n}\right) \qquad (g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\geq 0)\\
&\quad\cdot\exp\left(\frac{2\eta}{n}\sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\left(C_1\left\|\mu\right\|^2 + 2C_1\sqrt{d\log\left(n/\delta\right)} + 2\alpha\sqrt{C_1 d}+\alpha^2\right)\right)\\
&\leq A_t\cdot\exp\left(\frac{\eta\left(\sqrt{C_1 d}+\alpha\right)^2}{n}\right)\exp\left(2\eta\left(C_1\left\|\mu\right\|^2 + 2C_1\sqrt{d\log\left(n/\delta\right)} + 2\alpha\sqrt{C_1 d}+\alpha^2\right)\right)\\
&\qquad\qquad (0\leq g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))\leq 1)\\
&\leq \frac{2g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))}\exp\left(\frac{\eta\left(\sqrt{C_1 d}+\alpha\right)^2}{n}\right)\exp\left(2\eta\left(C_1\left\|\mu\right\|^2 + 2C_1\sqrt{d\log\left(n/\delta\right)} + 2\alpha\sqrt{C_1 d}+\alpha^2\right)\right)\\
&\qquad\qquad (\tfrac{1}{2}\exp\left(-z\right)\leq g(z)\leq\exp\left(-z\right), \forall z\geq 0; \text{Equation (4)})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{4\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2} \exp\left(2\eta\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2 + \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{2n}\right)\right) \\
&\le \frac{8\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2},
\end{aligned}
$$

where the last line holds from Assumption (4) with sufficiently large C so that the following holds

$$
\begin{aligned}
&2\eta\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2 + \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{2n}\right) \\
&= \frac{\eta\left(\sqrt{C_1 d}+\alpha\right)^2}{n} + 2\eta\left(C_1\|\mu\|^2+\alpha^2\right) + 4\eta\left(C_1\sqrt{d\log(n/\delta)}+\alpha\sqrt{C_1 d}\right) \\
&\le \frac{(2\sqrt{C_1 d})^2}{Cd^2 n} + \frac{2(2C_1\|\mu\|^2)}{Cd^2} + \frac{4(C_1\sqrt{\frac{d^2}{Cn^2}}+C_1 d)}{Cd^2} \qquad \text{(Assumption (1), (4), (6) with sufficiently large } C\text{)} \\
&\le \frac{1}{8}.
\end{aligned}
$$

Otherwise, $\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} > \frac{2\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}$, then we have

$$
\begin{aligned}
A_{t+1} &\le A_t \cdot \exp\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\eta\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} - \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\right)\right) \\
&\quad \cdot \exp\left(\frac{2\eta g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))}{n}\sum_{i=1}^{n}\frac{g(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2;\mathrm{W}^t))}\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2\right)\right) \\
&\le A_t \cdot \exp\left(-\frac{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\eta\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}{n}\left(\frac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} - \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\right)\right) \\
&\quad \cdot \exp\left(2\eta g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2\right)\max\left\{2, \frac{16\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\right\}\right) \\
&\le A_t \cdot \exp\Bigg(-g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))\eta\Bigg(\frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{n} \\
&\qquad - \frac{32\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2\right)\Bigg)\Bigg) \\
&\qquad \text{(Assumption that } \tfrac{g(y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t))}{g(y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t))} > \tfrac{2\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\text{)} \\
&\le A_t \le \frac{8\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2},
\end{aligned}
$$

where the last line holds from Assumption (6), Assumption (1) with C being sufficiently large that

$$
\frac{32\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\left(C_1\|\mu\|^2 + 2C_1\sqrt{d\log(n/\delta)} + 2\alpha\sqrt{C_1 d} + \alpha^2\right)
$$

$$
\begin{aligned}
&\le \frac{32\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\frac{1}{2}\sqrt{d/C_1}\right)^2}\left(C_1\frac{d}{Cn}+4C_1\sqrt{d}\sqrt{\log\left(n/\delta\right)+\alpha^2}+\|\mu\|^2\right)\\
&\le \frac{32\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\frac{1}{2}\sqrt{d/C_1}\right)^2}\left(C_1\frac{d}{Cn}+4C_1\sqrt{\frac{d^2}{Cn^2}}+\frac{d}{Cn}\right)\\
&\le \frac{\left(\sqrt{C_1 d}+\alpha\right)^2}{n}.
\end{aligned}
$$

We have shown through induction that

$$
\max_{i,j\in[n]}\frac{g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t))}\le \frac{16\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}.
$$

By Assumption (6) and Assumption (1), we know $\alpha \le \|\mu\| \le \frac{1}{2}\sqrt{d/C_1}$. Therefore,

$$
\max_{i,j\in[n]}\frac{g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t))}\le \frac{16\left(\sqrt{C_1 d}+\alpha\right)^2}{\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}\le \frac{64C_1\left(\sqrt{C_1}+0.5\sqrt{\frac{1}{C_1}}\right)^2}{\gamma^2}=C_r.
$$

□

With Lemma B.9, we can characterize a property of the adversarial training example $\tilde{\mathrm{x}}_i^t$:during training the perturbed data $\tilde{\mathrm{x}}_i^t$ is close to the linear subspace $\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\}$ in $\mathbb{R}^d$.

Lemma B.10. $\forall t\in\mathbb{N}$ and $i\in[n]$, the distance between $\tilde{\mathrm{x}}_i^t$ and $\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\}$ satisfies $\mathrm{dist}(\tilde{\mathrm{x}}_i^t,\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\})\le \min\left\{\frac{\omega_{\text{init}}\sqrt{md}}{\eta},\alpha\right\}$.

Proof of Lemma B.10. We define $C_d=\frac{\omega_{\text{init}}\sqrt{md}}{\eta}$ for simplicity. The upper bound α is obvious because the perturbation size is α. Now we look at C_d. We prove the result via induction. Consider time $t=0$, from the symmetric initialization, for any given x, we have $f(\mathrm{x};\mathrm{W}^0)=0$ is a constant function. Therefore, for any given training data x_i, generating the adversarial examples by adding any perturbations on x_i cannot increase the training loss. For simplicity, we consider the algorithm runs standard GD at time $t=0$; i.e. no adversarial training examples are generated for the first step, the adversarial training process starts at $t\ge 1$. This gives us that $\mathrm{dist}(\tilde{\mathrm{x}}_i^0,\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\})=\mathrm{dist}(\mathrm{x}_i,\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\})=0\le C_d$. Suppose we have $\mathrm{dist}(\tilde{\mathrm{x}}_i^s,\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\})\le C_d$ holds for any $1\le s\le t-1$, and we will now prove the result for t.

Recall $\tilde{\mathrm{x}}_k^t=\mathrm{argmax}_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x}_k,\alpha)}\ \ell(y_k f(\tilde{\mathrm{x}};\mathrm{W}^t))$. We decompose $\tilde{\mathrm{x}}_k^t=\tilde{\mathrm{x}}_{k,\parallel}^t+\tilde{\mathrm{x}}_{k,\perp}^t$, where $\tilde{\mathrm{x}}_{k,\parallel}^t\in\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\}$ and $\tilde{\mathrm{x}}_{k,\perp}^t\perp\mathrm{span}\{\mathrm{x}_1,\dots,\mathrm{x}_n\}$. Assume $\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2>C_d$, and we will prove via contradiction.

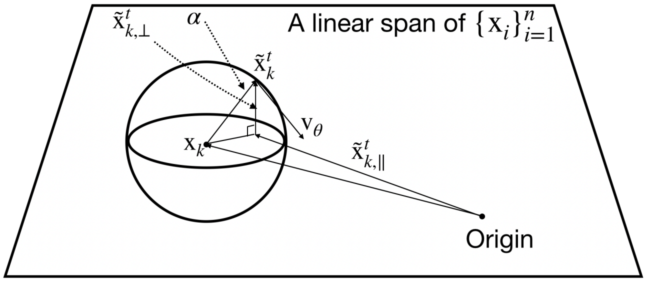
$\tilde{\mathrm{x}}_{k,\perp}^t$
α
A linear span of $\{\mathrm{x}_i\}_{i=1}^n$
$\tilde{\mathrm{x}}_k^t$
v_θ
x_k
$\tilde{\mathrm{x}}_{k,\parallel}^t$
Origin

Figure 7. Graphical illustration of adversarial examples and corresponding vectors that used in the proof of Lemma B.10.

As the loss function is monotonically decreasing, $\tilde{\mathrm{x}}_k^t = \mathrm{argmin}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_k, \alpha)} y_k f(\tilde{\mathrm{x}}; \mathrm{W}^t)$. As a result, there is no feasible direction that is also a descent direction. Here we construct directions $\mathrm{v}_\theta = -\tilde{\mathrm{x}}_{k,\perp}^t - \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i)$ for every $\theta \in \mathbb{R}$ that satisfies $0 < \theta < \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{(\alpha^2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2) \cdot 8C_1 dn}}$. We have that

$$\begin{aligned}
\left\langle \tilde{\mathrm{x}}_k^t - \mathrm{x}_k, \mathrm{v}_\theta \right\rangle &= \left\langle \tilde{\mathrm{x}}_{k,\perp}^t + \tilde{\mathrm{x}}_{k,\parallel}^t - \mathrm{x}_k, -\tilde{\mathrm{x}}_{k,\perp}^t - \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \left\langle \tilde{\mathrm{x}}_{k,\parallel}^t - \mathrm{x}_k, -\theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + \left\langle \tilde{\mathrm{x}}_{k,\perp}^t, -\tilde{\mathrm{x}}_{k,\perp}^t \right\rangle \\
&\leq \theta \|\tilde{\mathrm{x}}_{k,\parallel}^t - \mathrm{x}_k\|_2 \cdot \|\sum_{i=1}^n y_i \mathrm{x}_i\|_2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2 \\
&\leq \theta \sqrt{(\alpha^2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2) \cdot 8C_1 dn} - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2 < 0, && \text{(Lemma B.4 (C4))}
\end{aligned}$$

therefore v_θ are feasible directions. From the above discussion, we know that v_θ cannot be descent directions. Pick $\theta = \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{8\alpha^2 C_1 dn}}$. From the form of the classifier $y_k f(\tilde{\mathrm{x}}; \mathrm{W}^t) = y_k \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \phi(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}} \rangle)$, and combining the fact that ϕ is strictly increasing with $\phi' \in [\gamma, 1]$, we know there exists $s_0 \in [m]$ such that $y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^t, \mathrm{v}_\theta \right\rangle \geq 0$.

$$\begin{aligned}
0 &\geq y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^t, -\mathrm{v}_\theta \right\rangle \\
&= \sum_{t'=0}^{t-1} y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^{t'+1} - \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \sum_{t'=0}^{t-1} y_k a_{s_0} \left\langle \frac{\eta a_{s_0}}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k'}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \sum_{t'=0}^{t-1} \left\langle \frac{\eta}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k',\parallel}^{t'}, \theta (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&\qquad + \sum_{t'=0}^{t-1} y_k \left\langle \frac{\eta}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k',\perp}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle
\end{aligned}$$

Applying Lemma B.3 and Lemma B.4 (C4), the third term can be bounded by

$$\left| y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \right| \leq \left\| \mathrm{w}_{s_0}^0 \right\| \left(\left\| \tilde{\mathrm{x}}_{k,\perp}^t \right\| + \left\| \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\| \right) \leq 2\omega_{\text{init}} \sqrt{d} (\left\| \tilde{\mathrm{x}}_{k,\perp}^t \right\| + \theta \sqrt{8C_1 dn})$$

For the second term, we have

$$\begin{aligned}
&\sum_{t'=0}^{t-1} y_k \left\langle \frac{\eta}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k',\perp}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t \right\rangle \\
&\leq \frac{\eta}{n\sqrt{m}} \sum_{t'=0}^{t-1} \left\langle \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \tilde{\mathrm{x}}_{k',\perp}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t \right\rangle && (\phi'(\cdot) \leq 1) \\
&\leq \sum_{t'=0}^{t-1} \frac{C_d \eta}{\sqrt{m}} \widehat{G}_{\text{rob}}(\mathrm{W}^{t'}) \left\| \tilde{\mathrm{x}}_{k,\perp}^t \right\| && \text{(Lemma B.9, induction that } \|\tilde{\mathrm{x}}_{k',\perp}^{t'}\|_2 \leq C_d, \forall t' \in [t-1])
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\sum_{t'=0}^{t-1}\left\langle \frac{\eta}{n\sqrt{m}}\sum_{k'=1}^{n}\tilde{g}_{k'}(\mathrm{W}^{t'})\phi'(\left\langle \mathrm{w}_{s_0}^{t'},\tilde{\mathrm{x}}_{k'}^{t'}\right\rangle)y_{k'}\tilde{\mathrm{x}}_{k',\parallel}^{t'},\theta(\sum_{i=1}^{n}y_i\mathrm{x}_i)\right\rangle\\
&=\theta\frac{\eta}{n\sqrt{m}}\sum_{t'=0}^{t-1}\left(\sum_{k'=1}^{n}\left\langle\gamma\tilde{g}_{k'}(\mathrm{W}^{t'})\tilde{\mathrm{x}}_{k',\parallel}^{t'},\mathrm{x}_{k'}\right\rangle-\sum_{k'=1}^{n}\sum_{i\neq k'}\left\langle\tilde{g}_{k'}(\mathrm{W}^{t'})\tilde{\mathrm{x}}_{k'}^{t'},\mathrm{x}_i\right\rangle\right)\\
&\geq\theta\frac{\eta}{\sqrt{m}}\sum_{t'=0}^{t-1}\frac{\widehat{G}_{\text{rob}}(\mathrm{W}^{t'})}{C_r}\left(\frac{\gamma d}{4C_1}-n\left(C_1\left(\|\mu\|^2+\sqrt{d\log{(n/\delta)}}\right)+\alpha\sqrt{C_1d}\right)\right)\right) && \text{(Lemma B.4, B.9)}\\
&\geq\theta\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{\text{rob}}(\mathrm{W}^{t'})}{8C_1C_r\sqrt{m}} && \text{(Assumption (1))}
\end{aligned}$$

As a result, we have

$$\begin{aligned}
y_ka_{s_0}\left\langle\mathrm{w}_{s_0}^{t},-\mathrm{v}_\theta\right\rangle&\geq\theta\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{8C_1C_r\sqrt{m}}-\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2-2\omega_{\text{init}}\sqrt{d}(\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2+\theta\sqrt{8C_1dn})\\
&&&\text{($\|\tilde{\mathrm{x}}_{k',\perp}^{t'}\|_2\leq C_d$ from induction)}\\
&\geq\frac{\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2^2}{\sqrt{8\alpha^2C_1dn}}\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{8C_1C_r\sqrt{m}}-\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2-2\omega_{\text{init}}\sqrt{d}(\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2+\frac{\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2^2}{\alpha})\\
&&&\text{(plug in θ)}\\
&\geq\frac{\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2^2}{\sqrt{\alpha^2C_1dn}}\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1C_r\sqrt{m}}-\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2-2\omega_{\text{init}}\sqrt{d}\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2\\
&&&\text{($\omega_{\text{init}}\leq\frac{\eta}{\sqrt{md}}\leq\frac{\eta}{\sqrt{Cmn}}$ and C sufficiently large)}\\
&\geq\frac{C_d\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2}{\sqrt{\alpha^2C_1dn}}\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1C_r\sqrt{m}}-\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2-2\omega_{\text{init}}\sqrt{d}\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2\\
&\geq\frac{C_d\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2}{\sqrt{\alpha^2C_1dn}}\sum_{t'=0}^{t-1}\frac{\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1C_r\sqrt{m}}-5\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2\\
&&&\text{($\sum_{t'=0}^{t-1}\widehat{G}_{rob}(W^{t'})\geq\widehat{G}_{rob}(W^{0})=\frac{1}{2}$)}\\
&>0. &&\text{($d\geq Cn\alpha^2$ from Assumption (1), and C sufficiently large)}
\end{aligned}$$

This is a contradiction. Therefore, we have proved $\mathrm{dist}(\tilde{\mathrm{x}}_k^t,\mathrm{span}\{\mathrm{x}_1,\ldots,\mathrm{x}_n\})=\|\tilde{\mathrm{x}}_{k,\perp}^{t}\|_2\leq C_d$. By induction, proof is complete. □

Using Lemma B.10, we can prove a different version of Lemma B.4 (C5) and (C6) that will be used later.

Lemma B.11. $\forall i\in\mathcal{C}$, $\frac{1}{3}\|\mu\|^2\leq\left\langle\mu,y_i\tilde{\mathrm{x}}_i^t\right\rangle\leq3\|\mu\|^2$. $\forall i\in\mathcal{N}$, $-3\|\mu\|^2\leq\left\langle\mu,y_i\tilde{\mathrm{x}}_i^t\right\rangle\leq-\frac{1}{3}\|\mu\|^2$.

Proof of Lemma B.11. From Lemma B.4 (C5) and (C6), we know that $\frac{1}{2}\|\mu\|^2\leq\langle\mu,y_i\mathrm{x}_i\rangle\leq2\|\mu\|^2$ holds for all $i\in\mathcal{C}$, and $-2\|\mu\|^2\leq\langle\mu,y_i\mathrm{x}_i\rangle\leq-\frac{1}{2}\|\mu\|^2$ holds for all $i\in\mathcal{N}$. Therefore, it suffices to prove $|\left\langle\mu,y_i\tilde{\mathrm{x}}_i^t\right\rangle-\langle\mu,y_i\mathrm{x}_i\rangle|\leq\frac{1}{6}\|\mu\|^2$. We can decompose $\tilde{\mathrm{x}}_i^t-\mathrm{x}_i=(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\parallel+(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\perp$, where $(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\parallel\in\mathrm{span}\{\mathrm{x}_1,\ldots,\mathrm{x}_n\}$ and $(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\perp\perp\mathrm{span}\{\mathrm{x}_1,\ldots,\mathrm{x}_n\}$. From Lemma B.10, $\|(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\perp\|_2\leq\min\{C_d,\alpha\}\leq C_d\leq1$. For the parallel component, we can write $(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\parallel=\sum_{k=1}^{n}z_k\mathrm{x}_k$, where $z_k\in\mathbb{R}$. From Lemma B.4 (C4), $\alpha^2\geq\|\tilde{\mathrm{x}}_i^t-\mathrm{x}_i\|_2^2\geq\|(\tilde{\mathrm{x}}_i^t-\mathrm{x}_i)_\parallel\|_2^2\geq\frac{d}{8C_1}\cdot\sum_{k=1}^{n}z_k^2$. Thus, $\sqrt{\frac{8C_1n\alpha^2}{d}}\geq\sqrt{n\sum_{k=1}^{n}z_k^2}\geq\sum_{k=1}^{n}|z_k|$.

Now we can prove the statement.

$$\begin{aligned}
|\left\langle \mu, y_i\tilde{\mathrm{x}}_i^t \right\rangle - \langle \mu, y_i\mathrm{x}_i\rangle| &= |\left\langle \mu, \tilde{\mathrm{x}}_i^t - \mathrm{x}_i \right\rangle| \\
&\le |\left\langle \mu, (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|} \right\rangle| + |\left\langle \mu, (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\perp} \right\rangle| \\
&\le \sum_{k=1}^{n} |z_k| \cdot |\left\langle \mu, \mathrm{x}_k \right\rangle| + C_d\|\mu\| \\
&\le \sqrt{\frac{8C_1 n\alpha^2}{d}} \cdot 2\|\mu\|^2 + C_d\|\mu\| \\
&\le \frac{1}{6}\|\mu\|^2 . && \text{(Assumption (A2), Assumption (1) and C being sufficiently large)}
\end{aligned}$$

□

With Lemma B.9, we are able to give a tighter bound on the norm of W^t.

Lemma B.12. There is a constant $C_2 > 1$ such that

$$\left\|\mathrm{W}^t\right\|_F \le \left\|\mathrm{W}^0\right\|_F + C_2\eta\sqrt{\frac{d}{n}}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s).$$

Proof of Lemma B.12. By triangle inequality we have that

$$\begin{aligned}
\left\|\mathrm{W}^t\right\|_F &= \left\|\mathrm{W}^{t-1} - \eta\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^{t-1})\right\|_F \\
&\le \left\|\mathrm{W}^{t-1}\right\| + \left\|\eta\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^{t-1})\right\|_F \\
&\le \left\|\mathrm{W}^0\right\|_F + \eta\sum_{s=0}^{t-1}\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^s)\right\|_F . && \text{(Telescope)}
\end{aligned}$$

Consider $\tilde{\mathrm{x}}_i^s = \operatorname{argmax}_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^s))$. Then we have the following

$$\begin{aligned}
&\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^s)\right\|_F^2 \\
&= \frac{1}{n^2}\left\|\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^s)y_i\nabla f(\tilde{\mathrm{x}}_i^s;\mathrm{W}^s)\right\|_F^2 \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\left(\tilde{g}_i(\mathrm{W}^s)\right)^2\left\|\nabla f(\tilde{\mathrm{x}}_i^s;\mathrm{W}^s)\right\|_F^2 + \sum_{i\neq j} y_i y_j \tilde{g}_i(\mathrm{W}^s)\tilde{g}_j(\mathrm{W}^s)\left\langle\nabla f(\tilde{\mathrm{x}}_i^s;\mathrm{W}^s), \nabla f(\tilde{\mathrm{x}}_j^s;\mathrm{W}^s)\right\rangle\right] \\
&\le \frac{1}{n^2}\Bigg[\sum_{i=1}^{n}\left(\tilde{g}_i(\mathrm{W}^s)\right)^2\left(\sqrt{C_1 d}+\alpha\right)^2 \\
&\qquad + \sum_{i\neq j}\tilde{g}_i(\mathrm{W}^s)\tilde{g}_j(\mathrm{W}^s)\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\Bigg] && \text{(Lemma B.8)} \\
&\le \frac{1}{n^2}\max_{k\in[n]}\tilde{g}_k(\mathrm{W}^s)\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^s)\left(2C_1 d+2\alpha^2+n\left(C_1\|\mu\|^2+C_1\sqrt{d\log(n/\delta)}+2\alpha\sqrt{C_1 d}+\alpha^2\right)\right) \\
&\le \frac{5C_1 d}{n}\max_{k\in[n]}\tilde{g}_k(\mathrm{W}^s)\widehat{G}_{\text{rob}}(\mathrm{W}^s),
\end{aligned}$$

where the last line follows Assumption (1) and Assumption (6).

Applying Lemma B.9 gives us

$$\max_{k\in[n]} \tilde{g}_k(\mathrm{W}^s) \leq \frac{C_r}{n}\sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}^s) = C_r \widehat{G}_{\text{rob}}(\mathrm{W}^s).$$

Define $C_2 := \sqrt{5C_1C_r}$, then we have

$$\left\|\nabla \widehat{L}_{\text{rob}}(\mathrm{W}^s)\right\|_F \leq \sqrt{\frac{5C_1C_rd}{n}}\widehat{G}_{\text{rob}}(\mathrm{W}^s) = C_2\sqrt{\frac{d}{n}}\widehat{G}_{\text{rob}}(\mathrm{W}^s). \tag{5}$$

As a result, we have

$$\left\|\mathrm{W}^t\right\|_F \leq \left\|\mathrm{W}^0\right\|_F + C_2\eta\sqrt{\frac{d}{n}}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s). \tag{6}$$

□

Recall that our goal is to give a lower bound on the normalized expected conditional margin. We start by giving a lower bound in terms of the cumulative increments of margin given any independent test example (x, y), shown in Lemma B.13.

Lemma B.13. Let $C_2 > 1$ be the constant from Lemma B.12. For a γ-leaky, H-smooth activation ϕ, on a good run, we have for any $t \geq 0$ and $(\mathrm{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, there exist $\tilde{\rho}_i^t = \rho\left(\mathrm{W}^t, \tilde{\mathrm{x}}_i^t, \mathrm{x}\right) \in [\gamma^2, 1]$ such that

$$y\left[f(\mathrm{x};\mathrm{W}^{t+1}) - f(\mathrm{x};\mathrm{W}^t)\right] \geq \frac{\eta}{n}\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^t)\left(\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle - \frac{H\left\|\mathrm{x}\right\|^2 C_2^2 d\eta}{2\sqrt{mn}}\right)$$

Proof of Lemma B.13. Note that since a good run occurs, Lemma B.6 implies

$$\left|f(\mathrm{x};\mathrm{W}^{t+1}) - f(\mathrm{x};\mathrm{W}^t) - \left\langle \nabla f(\mathrm{x};\mathrm{W}^t), \mathrm{W}^{t+1} - \mathrm{W}^t\right\rangle\right| \leq \frac{H\left\|\mathrm{x}\right\|^2}{2\sqrt{m}}\left\|\mathrm{W}^{t+1} - \mathrm{W}^t\right\|_2^2 \tag{7}$$

Therefore, we have

$$\begin{aligned}
& y\left[f(\mathrm{x};\mathrm{W}^{t+1}) - f(\mathrm{x};\mathrm{W}^t)\right] \\
&\geq y\left[\left\langle \nabla f(\mathrm{x};\mathrm{W}^t), \mathrm{W}^{t+1} - \mathrm{W}^t\right\rangle\right] - \frac{H\left\|\mathrm{x}\right\|^2}{2\sqrt{m}}\left\|\mathrm{W}^{t+1} - \mathrm{W}^t\right\|_2^2 && \text{(for } y \in \{\pm1\}\text{, apply (7))} \\
&= y\eta\left[\left\langle \nabla f(\mathrm{x};\mathrm{W}^t), \frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^t)y_i\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right] - \frac{H\left\|\mathrm{x}\right\|^2\eta^2}{2\sqrt{m}}\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_2^2 \\
&\geq \eta\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^t)\left\langle y\nabla f(\mathrm{x};\mathrm{W}^t), y_i\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right\rangle\right] - \frac{H\left\|\mathrm{x}\right\|^2 C_2^2 d\eta^2}{2\sqrt{mn}}\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
& && \text{(Equation (5)}, \left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_2 \leq \left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F, \widehat{G}_{\text{rob}}(\mathrm{W}^t) \leq 1) \\
&= \frac{\eta}{n}\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^t)\left(\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle - \frac{H\left\|\mathrm{x}\right\|^2 C_2^2 d\eta}{2\sqrt{mn}}\right) && (\widehat{G}_{\text{rob}}(\mathrm{W}^t) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^t))
\end{aligned}$$

where the last equality follows by defining

$$\tilde{\rho}_i^t = \rho(\mathrm{W}^t, \tilde{\mathrm{x}}_i^t, \mathrm{x}) = \frac{1}{m}\sum_{s=1}^{m}\phi'\left(\left\langle \mathrm{w}_s^{(t)}, \mathrm{x}\right\rangle\right)\phi'\left(\left\langle \mathrm{w}_s^{(t)}, \tilde{\mathrm{x}}_i^t\right\rangle\right) \in [\gamma^2, 1].$$

□

Leveraging Lemma B.9 and B.13, we now formally derive a lower bound on the normalized expected conditional margin.

Lemma B.14. For a γ-leaky H-smooth activation ϕ, and for all $C > 1$ sufficiently large, on a good run, for any $t \geq 1$, we have

$$\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2} \geq \frac{\gamma^2\sqrt{n}}{32C_2\sqrt{d}}\left\|\mu\right\|^2;$$
$$\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=-1\right]}{\left\|\mathrm{W}^t\right\|_2} \geq \frac{\gamma^2\sqrt{n}}{32C_2\sqrt{d}}\left\|\mu\right\|^2.$$

where C_2 is the constant from Lemma B.12.

Lemma B.14. From Lemma B.12, we have

$$\left\|\mathrm{W}^t\right\|_F \leq \left\|\mathrm{W}^0\right\|_F + C_2\eta\sqrt{\frac{d}{n}}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s).$$

Recall the following definition

$$\tilde{\rho}_i^t = \rho(\mathrm{W}^t, \tilde{\mathrm{x}}_i^t, \mathrm{x}) = \frac{1}{m}\sum_{s=1}^{m}\phi'\left(\left\langle \mathrm{w}_s^{(t)}, \mathrm{x}\right\rangle\right)\phi'\left(\left\langle \mathrm{w}_s^{(t)}, \tilde{\mathrm{x}}_i^t\right\rangle\right) \in [\gamma^2, 1].$$

By Lemma B.11, we have

$$\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, \mu\right\rangle|y=1] \geq \begin{cases}\frac{1}{3}\gamma^2\|\mu\|^2, & i \in \mathcal{C}\\ -3\left\|\mu\right\|^2, & i \in \mathcal{N}\end{cases}$$

If $i \in \mathcal{C}$:

$$\begin{aligned}
&\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle\Big|y=1\right]\\
&=\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle+\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, \mu\right\rangle\Big|y=1\right]\\
&\geq\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\cdot\mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\geq 0)\Big|y=1\right]\\
&\quad+\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\cdot\mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle< 0)\Big|y=1\right]+\frac{1}{3}\gamma^2\|\mu\|^2\\
&\geq\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\gamma^2\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\cdot\mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\geq 0)\Big|y=1\right]\\
&\quad+\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\cdot\mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle< 0)\Big|y=1\right]+\frac{1}{3}\gamma^2\|\mu\|^2\\
&=-\frac{1-\gamma^2}{2}\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\left|\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\right|\right]+\frac{1}{3}\gamma^2\|\mu\|^2\\
&\geq-\frac{1-\gamma^2}{2}c_3\left\|\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\right\|_{\psi_2}+\frac{1}{3}\gamma^2\|\mu\|^2 && (c_3\text{ is an absolute constant})\\
&\geq-\frac{1-\gamma^2}{2}c_4\left\|y_i\tilde{\mathrm{x}}_i^t\right\|_2+\frac{1}{3}\gamma^2\|\mu\|^2 && (c_4\text{ is an absolute constant})\\
&\geq-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha)+\frac{1}{3}\gamma^2\|\mu\|^2.
\end{aligned}$$

If $i \in \mathcal{N}$:

$$\begin{aligned}
&\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle\Big|y=1\right]\\
&=\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle+\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, \mu\right\rangle\Big|y=1\right]
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t \left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \cdot \mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \geq 0)\Big| y=1\right] \\
&\qquad + \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t \left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \cdot \mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle < 0)\Big| y=1\right] - 3\left\|\mu\right\|^2 \\
&\geq \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\gamma^2 \left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \cdot \mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \geq 0)\Big| y=1\right] \\
&\qquad + \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle \cdot \mathbb{1}(\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle < 0)\Big| y=1\right] - 3\left\|\mu\right\|^2 \\
&= -\frac{1-\gamma^2}{2}\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\left|\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\right| \Big| y=1\right] - 3\left\|\mu\right\|^2 \\
&\geq -\frac{1-\gamma^2}{2}c_3\left\|\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}-\mu\right\rangle\right\|_{\psi_2} - 3\left\|\mu\right\|^2 && (c_3 \text{ is an absolute constant}) \\
&\geq -\frac{1-\gamma^2}{2}c_4\left\|y_i\tilde{\mathrm{x}}_i^t\right\|_2 - 3\left\|\mu\right\|^2 && (c_4 \text{ is an absolute constant}) \\
&\geq -\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) - 3\left\|\mu\right\|^2 .
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^{s+1}) - yf(\mathrm{x};\mathrm{W}^{s})\Big| y=1\right] \\
&\geq \frac{\eta}{n}\sum_{i=1}^{n}\tilde{g}_i(\mathrm{W}^s)\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[\tilde{\rho}_i^t\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle - \frac{H\left\|\mathrm{x}\right\|^2 C_2^2 d\eta}{2\sqrt{mn}}\Big| y=1\right] && \text{(Lemma B.13)} \\
&\geq \eta\Bigg(\frac{1}{n}\sum_{i\in\mathcal{C}}\tilde{g}_i(\mathrm{W}^s)\left(-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) + \frac{1}{3}\gamma^2\|\mu\|^2\right) \\
&\qquad + \frac{1}{n}\sum_{i\in\mathcal{N}}\tilde{g}_i(\mathrm{W}^s)\left(-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) - 3\left\|\mu\right\|^2\right) - \frac{Hc_5 dC_2^2 d\eta}{2\sqrt{mn}}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\Bigg) && (c_5 \text{ is an absolute constant}) \\
&= \eta\left(-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) + \frac{1}{3}\gamma^2\|\mu\|^2\right) \\
&\qquad \cdot\Bigg(\left(1 - \frac{Hc_5 C_2^2 d^2\eta}{\left(-(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + \frac{2}{3}\gamma^2\|\mu\|^2\right)\sqrt{mn}}\right)\widehat{G}_{\text{rob}}(\mathrm{W}^s) \\
&\qquad - \left(1 + \frac{(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + 6\left\|\mu\right\|^2}{-(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + \frac{2}{3}\gamma^2\|\mu\|^2}\right)\frac{1}{n}\sum_{i\in\mathcal{N}}\tilde{g}_i(\mathrm{W}^s)\Bigg) \\
&\geq \eta\left(-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) + \frac{1}{3}\gamma^2\left\|\mu\right\|^2\right) \\
&\qquad \cdot\Bigg(\left(1 - \frac{Hc_5 C_2^2 d^2\eta}{\left(-(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + \frac{2}{3}\gamma^2\left\|\mu\right\|^2\right)\sqrt{mn}}\right)\widehat{G}_{\text{rob}}(\mathrm{W}^s) \\
&\qquad - (\beta + \sqrt{\frac{2}{C}})C_r\left(1 + \frac{(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + 6\left\|\mu\right\|^2}{-(1-\gamma^2)c_4(\sqrt{C_1 d}+\alpha) + \frac{2}{3}\gamma^2\left\|\mu\right\|^2}\right)\widehat{G}_{\text{rob}}(\mathrm{W}^s)\Bigg) \\
&\geq \frac{\eta\widehat{G}_{\text{rob}}(\mathrm{W}^s)}{4}\left(-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1 d}+\alpha) + \frac{1}{3}\gamma^2\left\|\mu\right\|^2\right) \\
&\geq \frac{\eta\widehat{G}_{\text{rob}}(\mathrm{W}^s)\gamma^2}{16}\left\|\mu\right\|^2 ,
\end{aligned}
$$

where the second last inequality follows from $\eta \leq \frac{1}{Cd^2} \leq \frac{\frac{1}{3}\gamma^4\|\mu\|^2\sqrt{mn}}{Hc_5 C_2^2 d^2}$, $\beta \leq 1/C, d \leq \frac{\|\mu\|^4}{C} \leq \frac{\gamma^4\|\mu\|^4}{9c_4^2 C_1}, \alpha \leq \|\mu\| \leq \sqrt{C_1 d}$

and C being sufficiently large so that

$$\begin{aligned}
&\frac{Hc_5C_2^2d^2\eta}{\left(-(1-\gamma^2)c_4(\sqrt{C_1d}+\alpha)+\frac{2}{3}\gamma^2\left\|\mu\right\|^2\right)\sqrt{mn}} \\
&\le \frac{\frac{1}{3}\gamma^4\left\|\mu\right\|^2}{-(1-\gamma^2)c_4(2\sqrt{C_1d})+\frac{2}{3}\gamma^2\left\|\mu\right\|^2} \\
&\le \frac{\frac{1}{3}\gamma^4\left\|\mu\right\|^2}{-\frac{2}{3}\gamma^2(1-\gamma^2)\left\|\mu\right\|^2+\frac{2}{3}\gamma^2\left\|\mu\right\|^2} = 0.5,
\end{aligned}$$

and

$$\begin{aligned}
&(\beta+\sqrt{\frac{2}{C}})C_r\left(1+\frac{(1-\gamma^2)c_4(\sqrt{C_1d}+\alpha)+6\left\|\mu\right\|^2}{-(1-\gamma^2)c_4(\sqrt{C_1d}+\alpha)+\frac{2}{3}\gamma^2\left\|\mu\right\|^2}\right) \\
&\le (\frac{1}{C}+\sqrt{\frac{2}{C}})C_r\left(1+\frac{(1-\gamma^2)c_4(2\sqrt{C_1d})+6\left\|\mu\right\|^2}{-(1-\gamma^2)c_4(2\sqrt{C_1d})+\frac{2}{3}\gamma^2\left\|\mu\right\|^2}\right) \\
&\le (\frac{1}{C}+\sqrt{\frac{2}{C}})C_r\left(1+\frac{\frac{2}{3}\gamma^2(1-\gamma^2)\left\|\mu\right\|^2+6\left\|\mu\right\|^2}{-\frac{2}{3}\gamma^2(1-\gamma^2)\left\|\mu\right\|^2+\frac{2}{3}\gamma^2\left\|\mu\right\|^2}\right) \\
&= (\frac{1}{C}+\sqrt{\frac{2}{C}})C_r\left(1+\frac{\frac{2}{3}\gamma^2(1-\gamma^2)+6}{\frac{2}{3}\gamma^4}\right) \\
&\le 0.25.
\end{aligned}$$

The third last inequality follows from Lemma B.9 that

$$\begin{aligned}
\sum_{i\in\mathcal{N}}\tilde{g}_i(\mathrm{W}^s) &\le |\mathcal{N}|\cdot\max_i\tilde{g}_i(\mathrm{W}^s) \\
&\le \frac{|\mathcal{N}|}{n}\sum_{k=1}^{n}\max_i\tilde{g}_i(\mathrm{W}^s) \\
&\le C_r\cdot|\mathcal{N}|\cdot\widehat{G}_{\text{rob}}(\mathrm{W}^s) \\
&\le C_r(\beta+\sqrt{\frac{2}{C}})n\widehat{G}_{\text{rob}}(\mathrm{W}^s).
\end{aligned} \tag{8}$$

The last inequality follows from $d\le\frac{\|\mu\|^4}{C}\le\frac{\gamma^4\|\mu\|^4}{144c_4^2C_1}$, and $\alpha\le\|\mu\|\le\sqrt{C_1d}$ so that

$$\begin{aligned}
&-\frac{1-\gamma^2}{2}c_4(\sqrt{C_1d}+\alpha)+\frac{1}{3}\gamma^2\left\|\mu\right\|^2 \\
&\ge -\frac{1-\gamma^2}{2}c_4(2\sqrt{C_1d})+\frac{1}{3}\gamma^2\left\|\mu\right\|^2 \\
&\ge -\frac{1-\gamma^2}{12}\gamma^2\left\|\mu\right\|^2+\frac{1}{3}\gamma^2\|\mu\|^2 \\
&\ge \frac{1}{4}\gamma^2\left\|\mu\right\|^2.
\end{aligned}$$

Applying the above result gives us the following

$$\begin{aligned}
&\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2} \\
&= \frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^0)|y=1\right]+\sum_{s=0}^{t-1}\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^{s+1})-yf(\mathrm{x};\mathrm{W}^s)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2} \\
&\ge \frac{\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\eta\gamma^2}{16\left\|\mathrm{W}^t\right\|_F}\left\|\mu\right\|^2. \qquad\qquad (f(\mathrm{x};\mathrm{W}^0)=0 \text{ via symmetric initialization})
\end{aligned}$$

Note that we have

$$\widehat{G}_{\text{rob}}(\mathrm{W}^0) = \frac{1}{n}\sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}^0) = -\frac{1}{n}\sum_{i=1}^{n} \ell'(y_i f(\tilde{\mathrm{x}}_i^0; \mathrm{W}^0)) = \frac{1}{2}. \tag{9}$$

Along with Lemma B.3, Assumption (3) and Assumption (1) gives us

$$\left\|\mathrm{W}^0\right\|_F \leq 2\omega_{\text{init}}\sqrt{md} \leq 2\eta \leq \eta\sqrt{d/n}\,\widehat{G}_{\text{rob}}(\mathrm{W}^0). \tag{10}$$

Then if $\left\|\mathrm{W}^t\right\|_F \leq 2\left\|\mathrm{W}^0\right\|_F$, we have

$$\begin{aligned}
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2} &\geq \frac{\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\eta\gamma^2}{32\left\|\mathrm{W}^0\right\|_F}\left\|\mu\right\|^2 \\
&\geq \frac{\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\eta\gamma^2}{32\eta\sqrt{d/n}\,\widehat{G}_{\text{rob}}(\mathrm{W}^0)}\left\|\mu\right\|^2 && \text{(Equation (10))} \\
&\geq \frac{\sqrt{n}\gamma^2}{32\sqrt{d}}\left\|\mu\right\|^2. && \left(\textstyle\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s) \geq \widehat{G}_{\text{rob}}(\mathrm{W}^0)\right)
\end{aligned}$$

If $\left\|\mathrm{W}^t\right\|_F > 2\left\|\mathrm{W}^0\right\|_F$, by Lemma B.12, we have

$$2\left\|\mathrm{W}^0\right\|_F \leq \left\|\mathrm{W}^t\right\|_F \leq \left\|\mathrm{W}^0\right\|_F + C_2\eta\sqrt{d/n}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s).$$

Thus we have

$$\begin{aligned}
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2} &\geq \frac{\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\eta\gamma^2}{32C_2\eta\sqrt{d/n}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)}\left\|\mu\right\|^2 \\
&\geq \frac{\sqrt{n}\gamma^2}{32C_2\sqrt{d}}\left\|\mu\right\|^2.
\end{aligned}$$

Similarly, we can get

$$\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=-1\right]}{\left\|\mathrm{W}^t\right\|_2} \geq \frac{\gamma^2\sqrt{n}}{32C_2\sqrt{d}}\left\|\mu\right\|^2.$$

□

We finally provide the convergence guarantees of robust training loss in Lemma B.15.

Lemma B.15. For a γ-leaky, H-smooth activation ϕ, provided $C > 1$ is sufficiently large, then on a good run we have that,

$$\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F \geq \frac{\gamma\left\|\mu\right\|}{4}\widehat{G}_{\text{rob}}(\mathrm{W}^t)$$

Moreover, the robust training loss satisfies

$$\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \frac{35 + 8\sqrt{\frac{m}{d^3}}}{\gamma\left\|\mu\right\|\eta T^{\frac{1-\zeta}{2}}}$$

Proof of Lemma B.15. Consider $\tilde{\mathrm{x}}_i^t$ as the adversarial example given model parameter W^t; i.e., $\tilde{\mathrm{x}}_i^t = \mathrm{argmax}_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}\,\ell(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t))$. We first need to show a lower bound for $\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F =$

$\sup_{\mathrm{U}:\|\mathrm{U}\|_F=1}\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{U}\right\rangle$, and it suffices to construct a matrix V with Frobenius norm at most one such that $\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{V}\right\rangle$ is bounded from below by a positive constant. To this end, choose $\mathrm{V}\in\mathbb{R}^{m\times d}$ be the matrix with rows $\mathrm{v}_s=\frac{a_s\mu}{\|\mu\|\sqrt{m}},\forall s\in[m]$. Then $\|\mathrm{V}\|_F=1$ since $a_s=\pm 1$, and we have for any $\mathrm{W}\in\mathbb{R}^{m\times d}$,

$$\langle\nabla f(\mathrm{x}_i;\mathrm{W}),\mathrm{V}\rangle=\frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\phi'(\langle\mathrm{w}_s,\mathrm{x}_i\rangle)\langle\mathrm{v}_s,\mathrm{x}_i\rangle=\left\langle\frac{\mu}{\|\mu\|},\mathrm{x}_i\right\rangle\frac{1}{m}\sum_{s=1}^m\phi'(\langle\mathrm{w}_s,\mathrm{x}_i\rangle) \tag{11}$$

By Lemma B.4 and Lemma B.11, we have

$$\begin{cases} y_i\langle\mu,\mathrm{x}_i\rangle\geq\frac{1}{2}\|\mu\|^2, & i\in\mathcal{C}\\ |\langle\mu,\mathrm{x}_i\rangle|\leq\frac{3}{2}\|\mu\|^2, & i\in\mathcal{N}\end{cases},\begin{cases} y_i\left\langle\mu,\tilde{\mathrm{x}}_i^t\right\rangle\geq\frac{1}{3}\|\mu\|^2, & i\in\mathcal{C}\\ \left|\left\langle\mu,\tilde{\mathrm{x}}_i^t\right\rangle\right|\leq 3\|\mu\|^2, & i\in\mathcal{N}\end{cases}$$

And $\forall z,\phi'(z)\geq\gamma>0$, equation (11) implies that we have the following lower bound for any $\mathrm{W}\in\mathbb{R}^{m\times d}$,

$$y_i\langle\nabla f(\mathrm{x}_i;\mathrm{W}),\mathrm{V}\rangle\geq\begin{cases}\frac{\gamma}{2}\|\mu\|, & i\in\mathcal{C}\\ -\frac{3}{2}\|\mu\|, & i\in\mathcal{N}\end{cases},y_i\left\langle\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}),\mathrm{V}\right\rangle\geq\begin{cases}\frac{\gamma}{3}\|\mu\|, & i\in\mathcal{C}\\ -3\|\mu\|, & i\in\mathcal{N}\end{cases}$$

This allows for a lower bound on $\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{V}\right\rangle$, since

$$\begin{aligned}\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{V}\right\rangle&=\frac{1}{n}\sum_{i=1}^n\tilde{g}_i(\mathrm{W}^t)y_i\left\langle\nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t),\mathrm{V}\right\rangle\\ &\geq\frac{1}{n}\sum_{i\in\mathcal{C}}\tilde{g}_i(\mathrm{W}^t)\frac{\gamma}{3}\|\mu\|-\frac{1}{n}\sum_{i\in\mathcal{N}}\tilde{g}_i(\mathrm{W}^t)3\|\mu\|\\ &=\frac{\gamma\|\mu\|}{3}\left[\widehat{G}_{\text{rob}}(\mathrm{W}^t)-(1+\frac{9}{\gamma})\frac{1}{n}\sum_{i\in\mathcal{N}}\tilde{g}_i(\mathrm{W}^t)\right]\\ &\geq\frac{\gamma\|\mu\|}{3}\left[\widehat{G}_{\text{rob}}(\mathrm{W}^t)-(1+\frac{9}{\gamma})\cdot C_r(\beta+\sqrt{\frac{2}{C}})\widehat{G}_{\text{rob}}(\mathrm{W}^t)\right] && \text{(Equation (8))}\\ &\geq\frac{\gamma\|\mu\|}{4}\widehat{G}_{\text{rob}}(\mathrm{W}^t), && (12)\end{aligned}$$

where the last line holds by C being sufficiently large so that

$$\begin{aligned}&(1+\frac{9}{\gamma})\cdot C_r(\beta+\sqrt{\frac{2}{C}})\\ &\leq(1+\frac{9}{\gamma})\cdot C_r(\frac{1}{C}+\sqrt{\frac{2}{C}})\\ &\leq\frac{1}{4}.\end{aligned}$$

Thus we have

$$\widehat{G}_{\text{rob}}(\mathrm{W}^t)\leq\frac{4}{\gamma\|\mu\|}\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{V}\right\rangle\leq\frac{4}{\gamma\|\mu\|}\left\|\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F. \tag{13}$$

We next give an upper bound on $\left\|\mathrm{W}^t\right\|_F^2$ as follows:

$$\begin{aligned}&\left\|\mathrm{W}^{t+1}\right\|_F^2 && (14)\\ &=\left\|\mathrm{W}^t-\eta\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F^2\end{aligned}$$

$$\begin{aligned}
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \left\|\nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t)\right\|_F^2 - 2\eta \frac{1}{n}\sum_{i=1}^n \ell'(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i \left\langle \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t), \mathrm{W}^t\right\rangle \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 - 2\eta \frac{1}{n}\sum_{i=1}^n \ell'(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i \left\langle \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t), \mathrm{W}^t\right\rangle && \text{(Equation (5))} \\
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n}\sum_{i=1}^n g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i \sum_{s=1}^m \frac{a_s}{\sqrt{m}} \phi'(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle) \left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle \\
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n}\sum_{i=1}^n g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i \sum_{s=1}^m \frac{a_s}{\sqrt{m}} \phi(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle) \\
&\quad + 2\eta \frac{1}{n}\sum_{i=1}^n g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i \sum_{s=1}^m \frac{a_s}{\sqrt{m}} (\phi'(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle) \left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle - \phi(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle)) \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n}\sum_{i=1}^n g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t) \\
&\quad + 2\eta \frac{1}{n}\sum_{i=1}^n g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)) \sum_{s=1}^m \frac{1}{\sqrt{m}} \left|\phi'(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle) \left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle - \phi(\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle)\right| \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta \frac{1}{n}\sum_{i=1}^n \sum_{s=1}^m \frac{1}{\sqrt{m}} (c_1 + c_2 \left|\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle\right|^\zeta) && (g(z)z \le \tfrac{1}{3}, g(z) \le 1) \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + \frac{2\eta c_2}{\sqrt{m}} \frac{1}{n} \sum_{i=1}^n \left\|\tilde{\mathrm{x}}_i^t\right\|_2^\zeta \sum_{s=1}^m \left\|\mathrm{w}_s^t\right\|_2^\zeta \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + \frac{2\eta c_2}{\sqrt{m}} \frac{1}{n} \sum_{i=1}^n \left\|\tilde{\mathrm{x}}_i^t\right\|_2^\zeta m^{1-\frac{\zeta}{2}} \left\|\mathrm{W}^t\right\|_F^\zeta \\
&\le \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \left\|\mathrm{W}^t\right\|_F^\zeta .
\end{aligned}$$

Then we have

$$\left\|\mathrm{W}^{t+1}\right\|_F^{2-\zeta} \le (\left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \left\|\mathrm{W}^t\right\|_F^\zeta)^{\frac{2-\zeta}{2}} .$$

If $\left\|\mathrm{W}^t\right\|_F \le 1$,

$$\begin{aligned}
\left\|\mathrm{W}^{t+1}\right\|_F^{2-\zeta} &\le \left(1 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta\right)^{\frac{2-\zeta}{2}} \\
&\le \left\|\mathrm{W}^t\right\|_F^{2-\zeta} + \left(1 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta\right)^{\frac{2-\zeta}{2}} \\
&\le \left\|\mathrm{W}^t\right\|_F^{2-\zeta} + 1 + \frac{2-\zeta}{2} \cdot \left(\eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta\right) .
\end{aligned}$$

If $\left\|\mathrm{W}^t\right\|_F > 1$,

$$\begin{aligned}
\left\|\mathrm{W}^{t+1}\right\|_F^{2-\zeta} &\le \left\|\mathrm{W}^t\right\|_F^{2-\zeta} \left(1 + \frac{\eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \left\|\mathrm{W}^t\right\|_F^\zeta}{\left\|\mathrm{W}^t\right\|_F^2}\right)^{\frac{2-\zeta}{2}} \\
&\le \left\|\mathrm{W}^t\right\|_F^{2-\zeta} \left(1 + \frac{2-\zeta}{2} \cdot \frac{\eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1 \sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \left\|\mathrm{W}^t\right\|_F^\zeta}{\left\|\mathrm{W}^t\right\|_F^2}\right)
\end{aligned}$$

$$\leq \left\|\mathrm{W}^t\right\|_F^{2-\zeta} + \frac{2-\zeta}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right).$$

Combining the two cases, we have

$$\left\|\mathrm{W}^{t+1}\right\|_F^{2-\zeta} \leq \left\|\mathrm{W}^t\right\|_F^{2-\zeta} + 1 + \frac{2-\zeta}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right).$$

Summing up the above inequality, we get

$$\left\|\mathrm{W}^T\right\|_F^{2-\zeta} \leq \left\|\mathrm{W}^0\right\|_F^{2-\zeta} + T\cdot\left(1 + \frac{2-\zeta}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right)\right).$$

Thus,

$$\begin{aligned}
&\left\|\mathrm{W}^T\right\|_F \\
&\leq \left(\left\|\mathrm{W}^0\right\|_F^{2-\zeta} + T\cdot\left(1 + \frac{2-\zeta}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right)\right)\right)^{\frac{1}{2-\zeta}} \\
&\leq \left\|\mathrm{W}^0\right\|_F + T^{\frac{1}{2-\zeta}}\cdot\left(1 + \frac{2-\zeta}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right)\right)^{\frac{1}{2-\zeta}} \\
&\leq \left\|\mathrm{W}^0\right\|_F + T^{\frac{1}{2-\zeta}}\cdot\left(1 + \frac{1}{2}\cdot\left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta + 2\eta c_1\sqrt{m} + 2\eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right)\right) \\
&= \left\|\mathrm{W}^0\right\|_F + T^{\frac{1}{2-\zeta}}\cdot\left(1 + \eta^2\frac{C_2^2 d}{2n} + \frac{1}{3}\eta + \eta c_1\sqrt{m} + \eta c_2 m^{\frac{1-\zeta}{2}}(\sqrt{C_1 d}+\alpha)^\zeta\right).
\end{aligned}$$

Consider the correlation between iterates weight and V as follows:

$$\begin{aligned}
\left\langle \mathrm{W}^{t+1}, \mathrm{V}\right\rangle &= \left\langle \mathrm{W}^t - \eta\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t), \mathrm{V}\right\rangle \\
&= \left\langle \mathrm{W}^t, \mathrm{V}\right\rangle - \eta\left\langle \nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t), \mathrm{V}\right\rangle \\
&= \dots \\
&= \left\langle \mathrm{W}^0, \mathrm{V}\right\rangle - \eta\sum_{s=0}^{t}\left\langle \nabla\widehat{L}_{\text{rob}}(\mathrm{W}^s), \mathrm{V}\right\rangle. \qquad (15)
\end{aligned}$$

Recall from Lemma B.9 that $\forall t \geq 0, y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^{t+1}) \geq y_k f(\tilde{\mathrm{x}}_k; \mathrm{W}^t), \forall \tilde{\mathrm{x}}_k \in \mathcal{B}_2(\mathrm{x}_k, \alpha), \forall k \in [n]$. Then we have $y_k f(\tilde{\mathrm{x}}_k^T; \mathrm{W}^T) \geq y_k f(\tilde{\mathrm{x}}_k^T; \mathrm{W}^t)$, and therefore $\ell(y_k f(\tilde{\mathrm{x}}_k^T; \mathrm{W}^T)) \leq \ell(y_k f(\tilde{\mathrm{x}}_k^T; \mathrm{W}^t)) \leq \ell(y_k f(\tilde{\mathrm{x}}_k^t; \mathrm{W}^t))$ by definition that $\tilde{\mathrm{x}}_i^t = \arg\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} \ell\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)\right), t \leq T$. As a result, we have

$$\begin{aligned}
&\widehat{L}_{\text{rob}}(\mathrm{W}^T) \\
&= \frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^T)\right) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)\right) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} -\ell'\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t)\right) \qquad (\ell(z) \leq -2\ell'(z) \text{ when } z \geq 0, \text{ Equation (4)}) \\
&= \frac{2}{T}\sum_{t=0}^{T-1}\widehat{G}_{\text{rob}}(\mathrm{W}^t)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{8}{\gamma \|\mu\| T} \sum_{t=0}^{T-1} \left\langle -\nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t), \mathrm{V} \right\rangle && \text{(Equation (13))} \\
&= \frac{8}{\gamma \|\mu\| \eta T} \left(\left\langle \mathrm{W}^T, \mathrm{V} \right\rangle - \left\langle \mathrm{W}^0, \mathrm{V} \right\rangle \right) && \text{(Equation (15))} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(\left\|\mathrm{W}^T\right\|_F + \left\|\mathrm{W}^0\right\|_F \right) \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(2 \left\|\mathrm{W}^0\right\|_F + T^{\frac{1}{2-\zeta}} \cdot \left(1 + \eta^2 \frac{C_2^2 d}{2n} + \frac{1}{3}\eta + \eta c_1 \sqrt{m} + \eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \right) \right) \\
&\leq \frac{8}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2-\zeta}}} \left(2 \left\|\mathrm{W}^0\right\|_F + 1 + \eta^2 \frac{C_2^2 d}{2n} + \frac{1}{3}\eta + \eta c_1 \sqrt{m} + \eta c_2 m^{\frac{1-\zeta}{2}} (\sqrt{C_1 d} + \alpha)^\zeta \right) \\
&\leq \frac{8}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2-\zeta}}} \left(\omega_{\text{init}} \sqrt{6md} + \frac{4}{3} + \eta \left(\frac{C_2^2 d \eta}{2n} + (c_1 + c_2) \sqrt{m} (\sqrt{C_1 d} + \alpha) \right) \right) && \text{(Lemma B.3)} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2-\zeta}}} \left(\sqrt{6\eta} + \frac{4}{3} + \eta \left(\frac{C_2^2 d \eta}{2n} + (c_1 + c_2) \sqrt{m} (\sqrt{C_1 d} + \alpha) \right) \right) && \text{(Assumption (3))} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2-\zeta}}} \left(\sqrt{6} + \frac{4}{3} + \frac{1}{Cd^2} \left(\frac{C_2^2 d}{2nCd^2} + (c_1 + c_2) \sqrt{m} (\sqrt{C_1 d} + \alpha) \right) \right) && \text{(Assumption (4))} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2-\zeta}}} \left(\sqrt{6} + \frac{4}{3} + \frac{1}{2} + \sqrt{\frac{m}{d^3}} \right) && \text{(choose } C > \max\{1, C_1, C_2, 4(c_1+c_2)^2\} \text{ be large enough)} \\
&\leq \frac{35 + 8\sqrt{\frac{m}{d^3}}}{\gamma \|\mu\| \eta T^{\frac{1-\zeta}{2}}}.
\end{aligned}
$$

$\forall \varepsilon > 0, T \geq \left(\frac{35+8\sqrt{\frac{m}{d^3}}}{\gamma \|\mu\| \eta \epsilon} \right)^{\frac{2}{1-\zeta}}$ guarantees $\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \varepsilon$. □

For the smooth Leaky ReLU activation function of (Frei et al., 2022), we have the following result as corollary.

Corollary B.16. For the γ-leaky H-smooth ReLU activation ϕ_{SLReLU} defined in Equation (1), and for $\kappa \in (0,1), \lambda > 0$ defined in Definition 2.1. There exists some constant $C > 0$ such that Assumption 1, (A1) and (A2) hold, then we have that with probability at least $1 - 2\delta$ over the random initialization and the draws of the samples, the robust training loss satisfies

$$
\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \frac{30 + \frac{6}{\sqrt{H}} m^{\frac{1}{4}}}{\gamma \|\mu\| \sqrt{\eta} \sqrt{T}}.
$$

Proof of Corollary B.16. Here our activation function ϕ is ϕ_{SLReLU}. The definition of ϕ_{SLReLU} gives us that

$$
\phi'(z) z = \begin{cases} z = \phi(z) + \frac{1-\gamma}{4H}, & z \geq \frac{1}{H} \\ \frac{1-\gamma}{2} H z^2 + \frac{1+\gamma}{2} z = \phi(z) + \frac{1-\gamma}{4} H z^2, & |z| \leq \frac{1}{H} \\ \gamma z = \phi(z) + \frac{1-\gamma}{4H}, & z \leq -\frac{1}{H} \end{cases}.
$$

Therefore, $\phi'(z)z - \phi(z) \leq \frac{1-\gamma}{4H}$. Similar as Lemma B.15, we have $\widehat{G}_{\text{rob}}(\mathrm{W}^t) \leq \frac{4}{\gamma \|\mu\|} \left\| \nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t) \right\|_F$. In terms of the upper bound on $\left\|\mathrm{W}^t\right\|_F^2$, we have

$$
\begin{aligned}
&\left\|\mathrm{W}^{t+1}\right\|_F^2 \\
&= \left\| \mathrm{W}^t - \eta \nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t) \right\|_F^2 \\
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \left\| \nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t) \right\|_F^2 - 2\eta \frac{1}{n} \sum_{i=1}^n \ell'(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) y_i \left\langle \nabla f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t), \mathrm{W}^t \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 - 2\eta \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) y_i \left\langle \nabla f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t), \mathrm{W}^t \right\rangle && \text{(Equation (5))} \\
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n} \sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) y_i \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t \rangle) \langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t \rangle \\
&= \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n} \sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) y_i \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \phi(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t \rangle) \\
&\qquad + 2\eta \frac{1}{n} \sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) \frac{1-\gamma}{4H} \sqrt{m} && (\phi'(z)z - \phi(z) \leq \tfrac{1-\gamma}{4H}) \\
&\leq \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} \widehat{G}_{\text{rob}}(\mathrm{W}^t)^2 + 2\eta \frac{1}{n} \sum_{i=1}^{n} g(y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t)) y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t) + \eta\sqrt{m} \frac{1-\gamma}{2H} && (g(z) \leq 1) \\
&\leq \left\|\mathrm{W}^t\right\|_F^2 + \eta^2 \frac{C_2^2 d}{n} + \frac{2}{3}\eta + \eta\sqrt{m} \frac{1-\gamma}{2H} && (g(z)z \leq \tfrac{1}{3}) \\
&\leq \left\|\mathrm{W}^t\right\|_F^2 + \eta \frac{C_2^2 d}{nCd^2} + \frac{2}{3}\eta + \eta\sqrt{m} \frac{1-\gamma}{2H} && \text{(Assumption (4))} \\
&\leq \left\|\mathrm{W}^t\right\|_F^2 + \eta \left(\frac{5}{3} + \frac{1-\gamma}{2H}\sqrt{m} \right).
\end{aligned}
$$

Telescoping gives us that

$$
\left\|\mathrm{W}^T\right\|_F^2 \leq \|\mathrm{W}_0\|_F^2 + \eta \left(\frac{5}{3} + \frac{1-\gamma}{2H}\sqrt{m} \right) T.
$$

As a result, we have

$$
\begin{aligned}
\widehat{L}_{\text{rob}}(\mathrm{W}^T) &= \frac{1}{n} \sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} \ell\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^T) \right) \\
&\leq \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} \ell\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t) \right) \\
&\leq \frac{1}{T} \sum_{t=0}^{T-1} \frac{2}{n} \sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)} -\ell'\left(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t) \right) && (\ell\left(z\right) \leq -2\ell'(z) \text{ when } z \geq 0, \text{ Equation (4)}) \\
&= \frac{2}{T} \sum_{t=0}^{T-1} \widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
&\leq \frac{8}{\gamma \|\mu\| T} \sum_{t=0}^{T-1} \left\langle -\nabla \widehat{L}_{\text{rob}}(\mathrm{W}^t), \mathrm{V} \right\rangle && \text{(Equation (13))} \\
&= \frac{8}{\gamma \|\mu\| \eta T} \left(\langle \mathrm{W}^T, \mathrm{V} \rangle - \langle \mathrm{W}^0, \mathrm{V} \rangle \right) && \text{(Equation (15))} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(\left\|\mathrm{W}^T\right\|_F + \left\|\mathrm{W}^0\right\|_F \right) \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(2\left\|\mathrm{W}^0\right\|_F + \sqrt{\eta \left(\frac{5}{3} + \frac{1-\gamma}{2H}\sqrt{m} \right) T} \right) \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(\omega_{\text{init}}\sqrt{6md} + \sqrt{\eta \left(\frac{5}{3} + \frac{1-\gamma}{2H}\sqrt{m} \right) T} \right) && \text{(Lemma B.3)} \\
&\leq \frac{8}{\gamma \|\mu\| \eta T} \left(\sqrt{6}\eta + \sqrt{\eta \left(\frac{5}{3} + \frac{1-\gamma}{2H}\sqrt{m} \right) T} \right) && \text{(Lemma (3))}
\end{aligned}
$$

$$\leq \frac{30+\frac{6}{\sqrt{H}}m^{\frac{1}{4}}}{\gamma\left\|\mu\right\|\sqrt{\eta}\sqrt{T}}.$$

□

B.2. Missing Proofs in Section 3.2

Theorem B.17. Let $\varepsilon > 0, \delta \in (0,1/2)$. $\kappa \in (0,1)$ and $\lambda > 0$ are defined in Definition 2.1. Let ϕ be a non-smooth activation with $\gamma \in (0,1]$. Set $\bar{T} = \left(\frac{30}{\|\mu\|\gamma c_0\sqrt{\eta}\varepsilon}\right)^2$. There exists some constant $C > 0$ such that Assumption 1 and the following holds: (B1) The network width satisfies $m \geq C\log(n/\delta)$. (B2) $\|\mu\|^2 \geq C\max\left\{\sqrt{\frac{d}{n}\log(md/n\delta)}, \log(n/\delta)\right\}$. Then there exists a constant $c > 0$ such that after running Algorithm 1 for $T \geq \bar{T}$ iterations, we have that with probability at least $1-2\delta$ over the random initialization and the draw of an i.i.d. sample of size n, the following holds:

1. The robust training loss satisfies $\widehat{L}_{\text{rob}}(\mathrm{W}^T) \leq \varepsilon$, the robust training error satisfies $\widehat{L}^{0/1}_{\text{rob}}(\mathrm{W}^T) = 0$.
2. The clean test error satisfies
$$L^{0/1}(\mathrm{W}^T) \leq \beta + 2\exp\Big(-\frac{c\lambda n\left\|\mu\right\|^4}{C^2 d}\Big).$$
3. For $\frac{\alpha}{\|\mu\|} \leq \frac{1}{C}\sqrt{\frac{n\|\mu\|^2}{d}}$, the robust test error satisfies
$$L^{0/1}_{\text{rob}}(\mathrm{W}^T) \leq \beta + 2\exp\Big(-c\lambda\left\|\mu\right\|^2\Big(\frac{\|\mu\|}{C}\sqrt{\frac{n}{d}} - \frac{\alpha}{\|\mu\|}\Big)^2\Big).$$

Proof of Theorem B.17. This proof is similar with the proof of Theorem B.1. The robust training loss bound is proved in Lemma B.29. For the generalization guarantee, apply Lemma B.2 with Lemma B.33, with probability at least $1-2\delta$,

$$\begin{aligned}
L^{0/1}_{\text{rob}}(\mathrm{W}^T) &= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}[\exists\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y \neq \mathrm{sign}(f(\tilde{\mathrm{x}};\mathrm{W}^T))] \\
&\leq \beta + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = 1]}{\left\|\mathrm{W}^T\right\|_2} - \alpha\right)^2\right) \\
&\quad + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = -1]}{\left\|\mathrm{W}^T\right\|_2} - \alpha\right)^2\right) \\
&\leq \beta + 2\exp\left(-c\lambda\left(\frac{\sqrt{n}}{C\sqrt{d}}\left\|\mu\right\|^2 - \alpha\right)^2\right), \qquad \text{(Choose } C \geq \tfrac{16C_2}{c_9}\text{)}
\end{aligned}$$

where the last line holds for $\frac{\alpha}{\|\mu\|} \leq \frac{\|\mu\|\sqrt{n}}{C\sqrt{d}}$, so that $\frac{\sqrt{n}}{C\sqrt{d}}\left\|\mu\right\|^2 - \alpha \geq 0$.

Similarly, we have

$$\begin{aligned}
L^{0/1}(\mathrm{W}^T) &\leq \beta + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = 1]}{\left\|\mathrm{W}^T\right\|_2}\right)^2\right) \\
&\quad + \exp\left(-c\lambda\left(\frac{\mathbb{E}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}[y_c f(\mathrm{x};\mathrm{W}^T)|y_c = -1]}{\left\|\mathrm{W}^T\right\|_2}\right)^2\right) \\
&\leq \beta + 2\exp\left(-\frac{c\lambda n\left\|\mu\right\|^4}{C^2 d}\right).
\end{aligned}$$

□

The proof of Theorem B.17 builds upon a sequence of Propositions and Lemmas, which we show below.

Proposition B.18 defines a set $\mathcal{G}$ to characterize the index of noise that have large variance, and show the size of set $\mathcal{G}$ is large.

Proposition B.18. Let $\xi = [\xi^1, \dots, \xi^d]^\top$ denote the random vector sampled from $\mathcal{D}_{\text{clust}}$. Define $\mathcal{G} = \{i \in [d] | \mathbb{E}(\xi^i)^2 \geq \frac{\kappa}{2}\}$. Then the number of elements $|\mathcal{G}| \geq \frac{\kappa}{2-\kappa} d$.

Proof of Proposition B.18. Since each ξ^i has subgaussian norm at most 1, we have $2 \geq \mathbb{E} \exp\left((\xi^i)^2\right) \geq 1 + \mathbb{E}(\xi^i)^2$, so $\mathbb{E}(\xi^i)^2 \leq 1$. Suppose $|\mathcal{G}| < \frac{\kappa}{2-\kappa} d$, then we have

$$\kappa d \leq \mathbb{E} \left\|\xi\right\|^2 = \sum_{i \in \mathcal{G}} \mathbb{E}(\xi^i)^2 + \sum_{i \notin \mathcal{G}} \mathbb{E}(\xi^i)^2 \leq |\mathcal{G}| \cdot 1 + (d - |\mathcal{G}|) \cdot \frac{\kappa}{2} < d \cdot \frac{\kappa}{2} + (1 - \frac{\kappa}{2}) \frac{\kappa}{2-\kappa} d = \kappa d,$$

which is a contradiction. □

Lemma B.19 provides properties of the initialized network weights, similar to Lemma B.3 except we have an additional result that the averaged initialized weights variance that belongs to set $\mathcal{G}$ are not small. This additional result will be used in proving Lemma B.21.

Lemma B.19. There is a universal constant $C_0 > 1$ such that with probability at least $1 - 3\delta/4$ over the random initialization,

$$\frac{1}{2}\omega_{\text{init}}^2 d \leq \left\|\mathrm{w}_s^0\right\|_2^2 \leq \frac{3}{2}\omega_{\text{init}}^2 d, \forall s \in [m]; \left\|\mathrm{W}^0\right\|_2 \leq C_0 \omega_{\text{init}}(\sqrt{m} + \sqrt{d});$$
$$\frac{\sum_{i \in \mathcal{G}} (\mathrm{w}_{s,i}^0)^2}{|\mathcal{G}|} \geq \frac{1}{4} \frac{\sum_{i \in [d]} (\mathrm{w}_{s,i}^0)^2}{d}, \forall s \in [m].$$

Here $\mathcal{G}$ is defined in Proposition B.18.

Proof of Lemma B.19. The first part is proved in Lemma B.3, and it holds with probability at least $1 - \delta/2$.

For the second part, for any fixed $s \in [m]$, by concentration of the χ^2 distribution, we have

$$\mathrm{P}\left(\left|\frac{1}{|\mathcal{G}|\omega_{\text{init}}^2} \sum_{i \in \mathcal{G}} (\mathrm{w}_{s,i}^0)^2 - 1\right| \geq \frac{1}{2}\right) \leq 2 \exp\left(-|\mathcal{G}|/32\right),$$
$$\mathrm{P}\left(\left|\frac{1}{d\omega_{\text{init}}^2} \sum_{i \in [d]} (\mathrm{w}_{s,i}^0)^2 - 1\right| \geq 1\right) \leq 2 \exp\left(-d/8\right).$$

Applying a union bound over all $s \in [m]$,

$$\frac{\sum_{i \in \mathcal{G}} (\mathrm{w}_{s,i}^0)^2}{|\mathcal{G}|} \geq \frac{1}{2}\omega_{\text{init}}^2 \geq \frac{1}{4} \frac{\sum_{i \in [d]} (\mathrm{w}_{s,i}^0)^2}{d}$$

holds with probability at least

$$\begin{aligned}
&1 - 2m \exp\left(-|\mathcal{G}|/32\right) - 2m \exp\left(-d/8\right) \\
&\geq 1 - 2m \exp\left(-\frac{\kappa}{32(2-\kappa)} d\right) - 2m \exp\left(-d/8\right) && \text{(Proposition B.18)} \\
&\geq 1 - 2\delta \exp\left(n/C - \frac{\kappa}{32(2-\kappa)} d\right) - 2\delta \exp\left(n/C - d/8\right) && \text{(Assumption (5))} \\
&\geq 1 - 2\delta \exp\left(-(\frac{\kappa}{32(2-\kappa)} C \log(2) - 1/C)\right) - 2\delta \exp\left(-(C \log(2)/8 - 1/C)\right) \\
& && (d \geq Cn^2 \log(2) \text{ from Assumption (1); } C \text{ sufficiently large}) \\
&\geq 1 - \delta/4. && (C \text{ sufficiently large})
\end{aligned}$$

The proof is complete by taking a union bound over the above claims. □

The following anti-concentration inequality is helpful in proving Lemma B.21.

Proposition B.20. [Anti-concentration of subgaussian random variables] Assume X is a c-subgaussian random variable with $\mathbb{E}[X]=0, \mathbb{E}[X^2]=1$, then

$$\mathrm{P}\left(X \geq \frac{1}{10c^2}\right) \geq \frac{0.04}{220^{\frac{2}{3}} c^4}.$$

Proof of Lemma B.20. Denote $a=\frac{1}{10c^2}$, $A=220^{\frac{1}{3}}c^2$ and $B=\frac{0.04}{A^2}$. From $2 \geq \mathbb{E}\exp\left(\frac{X^2}{c^2}\right) \geq 1+\mathbb{E}\frac{X^2}{c^2}$ we know $c \geq 1$. Consider a truncated version of X defined by $\tilde{X}=X \cdot \mathbb{1}_{a \leq |X| \leq A}$. We have

$$\mathbb{E}\tilde{X}=\mathbb{E}\tilde{X}-\mathbb{E}X=-\mathbb{E}\left(X \cdot \mathbb{1}_{|X|<a}\right)-\mathbb{E}\left(X \cdot \mathbb{1}_{|X|>A}\right).$$

It is trivial that $\left|\mathbb{E}\left(X \cdot \mathbb{1}_{|X|<a}\right)\right| \leq a$. By subgaussian tail bound one may compute

$$\begin{aligned}
&\left|\mathbb{E}\left(X \cdot \mathbb{1}_{|X|>A}\right)\right| \\
&\leq \mathbb{E}\left(|X| \cdot \mathbb{1}_{|X|>A}\right) \\
&\leq \int_0^A \mathrm{P}\left(|X|>A\right) dt+\int_A^\infty \mathrm{P}\left(|X|>t\right) dt \\
&\leq 2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\int_A^\infty 2\exp\left(-\frac{t^2}{c^2}\right) dt \\
&\leq 2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right).
\end{aligned}$$

These imply

$$\left|\mathbb{E}\tilde{X}\right| \leq a+2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right).$$

Similarly,

$$\begin{aligned}
1=\mathbb{E}X^2 =&\mathbb{E}\left(X^2 \cdot \mathbb{1}_{|X|<a}\right)+\mathbb{E}\left(X^2 \cdot \mathbb{1}_{|X|>A}\right)+\mathbb{E}\tilde{X}^2 \\
&\leq a^2+2A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right)+2c^2\exp\left(-\frac{A^2}{c^2}\right)+\mathbb{E}\tilde{X}^2.
\end{aligned}$$

Thus,

$$\left|\mathbb{E}\tilde{X}^2-1\right| \leq a^2+2A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right)+2c^2\exp\left(-\frac{A^2}{c^2}\right).$$

Let $\tilde{X}_+=\max(\tilde{X},0)$ and $\tilde{X}_-=\max(-\tilde{X},0)$, then $\tilde{X}_+$ and $\tilde{X}_-$ are non-negative and $\tilde{X}=\tilde{X}_+-\tilde{X}_-$, $\tilde{X}^2=\tilde{X}_+^2+\tilde{X}_-^2$. We thus have

$$\begin{aligned}
\left|\mathbb{E}\tilde{X}_+-\mathbb{E}\tilde{X}_-\right| &\leq a+2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right), \\
\mathbb{E}\tilde{X}_+^2+\mathbb{E}\tilde{X}_-^2 &\geq 1-a^2-2A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right)-2c^2\exp\left(-\frac{A^2}{c^2}\right).
\end{aligned}$$

Now assume to the contrary that $\mathrm{P}\left(X \geq a\right)<B$. We have

$$\begin{aligned}
&\mathbb{E}\tilde{X}_+ \leq A \cdot \mathrm{P}\left(|X| \geq a\right)<A \cdot B, \\
&\mathbb{E}\tilde{X}_+^2<A^2 \cdot B, \\
&\mathbb{E}\tilde{X}_- \leq \mathbb{E}\tilde{X}_++a+2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right) \\
&\qquad <A \cdot B+a+2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right), \\
&\mathbb{E}\tilde{X}_-^2 \leq A \cdot \mathbb{E}\tilde{X}_-<A\left(A \cdot B+a+2A \cdot \exp\left(-\frac{A^2}{c^2}\right)+\frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right)\right).
\end{aligned}$$

These together imply

$$
\begin{aligned}
\mathbb{E}\tilde{X}_+^2 + \mathbb{E}\tilde{X}_-^2 &< A^2 \cdot B + A\left(A \cdot B + a + 2A \cdot \exp\left(-\frac{A^2}{c^2}\right) + \frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right)\right) \\
&< 1 - a^2 - 2A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right) - 2c^2\exp\left(-\frac{A^2}{c^2}\right),
\end{aligned}
$$

which is a contradiction. The last inequality holds since

$$
\begin{aligned}
RHS - LHS =& 1 - a^2 - 2A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right) - 2c^2\exp\left(-\frac{A^2}{c^2}\right) \\
& - A^2 \cdot B - A\left(A \cdot B + a + 2A \cdot \exp\left(-\frac{A^2}{c^2}\right) + \frac{c^2}{A}\exp\left(-\frac{A^2}{c^2}\right)\right) \\
=& 1 - a^2 - 2A^2 \cdot B - A \cdot a - 4A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right) - 3c^2\exp\left(-\frac{A^2}{c^2}\right) \\
=& 1 - \frac{1}{100c^4} - 0.08 - A \cdot a - 4A^2 \cdot \exp\left(-\frac{A^2}{c^2}\right) - 3c^2\exp\left(-\frac{A^2}{c^2}\right) \\
\geq& 1 - \frac{1}{100} - 0.08 - A \cdot a - 8A^2 / \left(\frac{A^2}{c^2}\right)^2 - 3c^2 / \frac{A^2}{c^2} \\
=& 1 - 0.09 - 1.5(22c^4a^2)^{\frac{1}{3}} = 1 - 0.09 - 1.5(0.22)^{\frac{1}{3}} > 0.
\end{aligned}
$$

This completes the proof. □

By combining Proposition B.18, Lemma B.19, and Proposition B.20, we establish the existence of enough neurons with positive activation at the first and second step of adversarial training, as stated in Lemma B.21. This lemma plays a vital role throughout the entire proof of networks with non-smooth activation functions.

Lemma B.21. Suppose the events in Lemma B.4 and Lemma B.19 hold. Given Assumption 1, (B1) and (B2), there exists a constant $c_0 > 0$ that only depends on κ such that with probability at least $1 - \delta/5$,

$$
\begin{aligned}
\forall s \in [m], \left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0\right\}\right| &\geq c_0 n; \\
\forall i \in [n], \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0\right\}\right| &\geq c_0 m. \\
\forall s \in [m], \left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i \right\rangle \geq \alpha \left\|\mathrm{w}_s^1\right\|\right\}\right| &\geq c_0 n; \\
\forall i \in [n], \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i \right\rangle \geq \alpha \left\|\mathrm{w}_s^1\right\|\right\}\right| &\geq c_0 m.
\end{aligned}
$$

Proof of Lemma B.21. Firstly, we prove the result for w_s^0, i.e., the first 2 statements. Fix any given (x_i, y_i). Recall that $\mathrm{x}_i = y_i^c \mu + \xi_i$, where y_i^c is the clean label, $\xi_i \sim \mathcal{D}_{\text{clust}}$ is the noise. Note that $\left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle = y_i^c \left\langle \mathrm{w}_s^0, \mu \right\rangle + \left\langle \mathrm{w}_s^0, \xi_i \right\rangle$. The first term is a centered Gaussian with variance $\omega_{\text{init}}^2 \left\|\mu\right\|^2$, therefore applying concentration argument gives us with probability at least $1 - \delta/20$, $\max_{s\in[m]} \left|\left\langle \mathrm{w}_s^0, \mu \right\rangle\right| \leq 4\omega_{\text{init}} \left\|\mu\right\| \sqrt{\log\left(m/\delta\right)}$. For the second term, condition on ξ_i, which is a centered Gaussian with variance $\omega_{\text{init}}^2 \left\|\xi_i\right\|^2, \forall s \in [m]$. Since $\mathrm{P}\left(\left\langle \mathrm{w}_s^0, \xi_i \right\rangle \geq \frac{\omega_{\text{init}}\left\|\xi_i\right\|}{10}\right) \geq \frac{1}{5}$, applying the Hoeffding's inequality gives us with probability at least $1 - \exp\left(-m/225\right)$, there exists a subset $J_i \in [m]$ with $|J_i| \geq m/15$ such that $\left\langle \mathrm{w}_s^0, \xi_i \right\rangle \geq \frac{\omega_{\text{init}}\left\|\xi_i\right\|}{10}$ and $a_s = y_i, \forall s \in J_i$. Conditioning on $\left\|\xi_i\right\|^2 \geq \frac{\kappa d}{2}$ obtained in Lemma B.4, we have that

$$
\begin{aligned}
\left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle &\geq -4\omega_{\text{init}} \left\|\mu\right\| \sqrt{\log\left(m/\delta\right)} + \omega_{\text{init}} \left\|\xi_i\right\|/10 \qquad \text{(Proposition B.19)} \\
&\geq \omega_{\text{init}}\left(\sqrt{\kappa d}/20 - 4\left\|\mu\right\| \sqrt{\log\left(m/\delta\right)}\right) \\
&\geq \omega_{\text{init}}\left(\sqrt{\kappa d}/20 - 4\sqrt{d}/C\right) > 0,
\end{aligned}
$$

where the last line holds via Assumption (1) and (5) for large enough C. Combining the above arguments, we have with probability at least $1 - \delta/20 - n\exp\left(-m/225\right)$,

$$
\forall i \in [n], \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0\right\}\right| \geq \frac{m}{15}.
$$

Given $m \geq C \log\left(n/\delta\right)$, the above holds with probability at least $1 - \delta/10$.

For the other statement, we can condition on w_s^0 similarly. Denote $X_i = \left\langle \frac{\mathrm{w}_s^0}{\|\mathrm{w}_s^0\|_2}, \xi_i \right\rangle$, $Y_i = \frac{X_i}{\sqrt{\mathbb{E}X_i^2}}$. It is obvious that $\mathbb{E}Y_i = 0$ and $\mathbb{E}Y_i^2 = 1$.

$$\begin{aligned}
\mathbb{E}X_i^2 =& \frac{\sum_{j\in[d]} (\mathrm{w}_{s,j}^0)^2 \mathbb{E}(\xi_i^j)^2}{\|\mathrm{w}_s^0\|_2^2} \\
\geq& \frac{\sum_{j\in\mathcal{G}} (\mathrm{w}_{s,j}^0)^2 \mathbb{E}(\xi_i^j)^2}{\sum_{j\in[d]} (\mathrm{w}_{s,j}^0)^2} && \text{(here } \mathcal{G} \text{ is defined in Proposition B.18)} \\
\geq& \frac{|\mathcal{G}| \cdot \kappa/2}{4d} \geq \frac{\kappa^2}{16 - 8\kappa}. && \text{(Lemma B.19 and Proposition B.18)}
\end{aligned}$$

From Proposition 2.6.1 in (Vershynin, 2018), there exists a universal constant c^1 such that $\|Y_i\|_{\psi_2} = \frac{\|X_i\|_{\psi_2}}{\sqrt{\mathbb{E}X_i^2}} \leq \frac{c^1\sqrt{16-8\kappa}}{\kappa}$. Applying Proposition B.20 gives us that

$$\mathrm{P}\left(X_i \geq \frac{\kappa^3}{10(c^1)^2(16-8\kappa)^{\frac{3}{2}}} \right) \geq \mathrm{P}\left(Y_i \geq \frac{\kappa^2}{10(c^1)^2(16-8\kappa)} \right) \geq 0.04(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2}).$$

Therefore, $\left\langle \mathrm{w}_s^0, \xi_i \right\rangle > 0$ holds with probability at least $0.04(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2})$. Consider y_i^c be the clean label that is uniformly distributed on $\{-1, +1\}$ and is independent of ξ_i, then we have

$$\begin{aligned}
&\mathrm{P}\left(\left\langle \mathrm{w}_s^0, \xi_i \right\rangle > 0, a_s = y_i^c | \mathrm{w}_s^0\right) \\
=& \frac{1}{2}\mathrm{P}\left(\left\langle \mathrm{w}_s^0, \xi_i \right\rangle > 0 | \mathrm{w}_s^0\right) \\
\geq& 0.02(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2}).
\end{aligned}$$

Similar with the previous part, since $\omega_{\text{init}}^2 3d/2 \geq \left\|\mathrm{w}_s^0\right\|^2 \geq \omega_{\text{init}}^2 d/2$ holds, applying Hoeffding's inequality, with probability at least $1 - \delta/20 - m\exp\left(-5 \times 10^{-5}(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2})^2 n\right)$,

$$\forall s \in [m], \left|\left\{ i \in [n] : y_i^c = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0 \right\}\right| \geq 0.015(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2})n.$$

When C is sufficiently large and $n \geq C\log\left(m/\delta\right)$ as assumed in Assumption (5), the above holds with probability at least $1 - \delta/10$. Note that $|\{i \in [n]; y_i \neq y_i^c\}| \leq \left(1/C + \sqrt{2/C}\right) n \leq 0.005(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2})n$ holds for a sufficient large C. Thus,

$$\forall s \in [m], \left|\left\{ i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0 \right\}\right| \geq 0.01(\frac{\kappa^4}{220^{\frac{2}{3}}(c^1)^4(16-8\kappa)^2})n.$$

The proof of the first two statements is complete by taking a union bound over the above two claims.
Now we are going to prove the last two statements. We consider the algorithm runs standard GD at time $t = 0$; i.e. no adversarial training examples are generated for the first step, the adversarial training process starts at $t \geq 1$. The gradient descent update gives us $\mathrm{w}_s^1 = \mathrm{w}_s^0 + \frac{\eta a_s}{2n\sqrt{m}} \sum_{k=1}^{n} \phi'(\left\langle \mathrm{w}_s^0, \mathrm{x}_k \right\rangle) y_k \mathrm{x}_k$.
$\forall s \in [m], \left|\left\{ i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0 \right\}\right| \geq c_0 n$. For these i,

$$\left\langle \mathrm{w}_s^1, \mathrm{x}_i \right\rangle - \alpha\left\|\mathrm{w}_s^1\right\| = \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle + \left\langle \frac{\eta a_s}{2n\sqrt{m}} \sum_{k=1}^{n} \phi'(\left\langle \mathrm{w}_s^0, \mathrm{x}_k \right\rangle) y_k \mathrm{x}_k, \mathrm{x}_i \right\rangle - \alpha\left\|\mathrm{w}_s^1\right\|$$

$$
\begin{aligned}
&\geq -\sqrt{\frac{3}{2}}\omega_{\text{init}}\sqrt{d}\cdot 2\sqrt{C_1 d} + \frac{\eta}{2n\sqrt{m}}\left(\gamma\frac{d}{C_1} - nC_1\left(\|\mu\|^2 + \sqrt{d\log\left(n/\delta\right)}\right)\right) \\
&\quad - \alpha\left(\sqrt{\frac{3}{2}}\omega_{\text{init}}\sqrt{d} + \frac{\eta}{2n\sqrt{m}}\sqrt{8C_1 dn}\right) && \text{(Lemma B.19 and Lemma B.4)} \\
&\geq -\sqrt{\frac{3}{2}}\omega_{\text{init}}\sqrt{d}\cdot 3\sqrt{C_1 d} + \frac{\eta}{2n\sqrt{m}}\left(\frac{\gamma}{2C_1} - \sqrt{\frac{8C_1}{C}}\right)d \\
& && \left(\alpha \leq \|\mu\| \leq \sqrt{\tfrac{d}{Cn}} \leq \sqrt{C_1 d} \text{ and Assumption (1)}\right) \\
&\geq \frac{\eta}{2n\sqrt{m}}\left(\frac{\gamma}{2C_1} - \sqrt{\frac{8C_1}{C}} - \sqrt{\frac{54C_1 n^2}{d}}\right)d && \text{(Assumption (3))} \\
&> 0. && \left(d \geq Cn^2\log\left(2\right) \text{ from Assumption (1) and C sufficiently large}\right)
\end{aligned}
$$

Therefore, $\forall s \in [m], \left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i\right\rangle \geq \alpha\left\|\mathrm{w}_s^1\right\|\right\}\right| \geq c_0 n$.
Similarly, $\forall i \in [n], \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^1, \mathrm{x}_i\right\rangle \geq \alpha\left\|\mathrm{w}_s^1\right\|\right\}\right| \geq c_0 m$. □

Definition B.22. If the events in Lemma B.4, B.19 and B.21 occur, let us say that we have a good run.

A good run occurs with probability at least $1 - 2\delta$. In the following proof, we condition on a good run occurs.

The following proposition presents several properties of the distribution $\mathcal{D}_{\text{clust}}$, which are crucial for establishing the generalization guarantees.

Proposition B.23. Assume $\xi \sim \mathcal{D}_{\text{clust}}$. Then the following holds:

(D1) For any fixed $\mathrm{v} \in \mathbb{R}^d$, for any $\bar{\delta} < 0.5$, with probability at least $1 - \bar{\delta}$ w.r.t ξ, $|\langle \mathrm{v}, \xi\rangle| \leq c_6\|\mathrm{v}\|\sqrt{\log\left(1/\bar{\delta}\right)}$, where c_6 is an absolute constant.

(D2) For any $\bar{\delta} < 0.5$, with probability at least $1 - \bar{\delta}$ w.r.t ξ, $\|\xi\|^2 \leq 9d(\log\left(1/\bar{\delta}\right))$. In particular, denote event $\mathcal{E} = \left\{\|\xi\| \geq 6\sqrt{\log\left(1/\bar{\delta}\right)d}\right\}$, we have $\mathbb{E}\left[\|\xi\|\mathbb{1}\left(\mathcal{E}\right)\right] \leq 8\bar{\delta}^4 \cdot \sqrt{d \cdot \log\left(1/\bar{\delta}\right)}$.

Proof of B.23. We first prove (D1). Note that the coordinates of ξ are independent variables with ψ_2 norm at most 1. From Hoeffding's inequality, there exists a universal constant c such that

$$
\mathrm{P}\left(|\langle \mathrm{v}, \xi\rangle| > c_6\|v\|\sqrt{\log\left(1/\bar{\delta}\right)}\right) \leq 2\exp\left(-cc_6^2\|v\|^2\log\left(1/\bar{\delta}\right)/\|v\|^2\right) = 2\bar{\delta}^{cc_6^2}.
$$

By selecting $c_6 = \sqrt{\frac{2}{c}}$, we get

$$
\mathrm{P}\left(|\langle \mathrm{v}, \xi\rangle| > c_6\|v\|\sqrt{\log\left(1/\bar{\delta}\right)}\right) \leq 2\bar{\delta}^2 \leq \bar{\delta}.
$$

Next we prove (D2). Note that the coordinates of $\xi = [\xi^1, \xi^2, \ldots, \xi^d]^\top$ are independent variables with ψ_2 norm at most 1. From Bernstein's inequality, there exists a universal constant c such that for every $t > 0$,

$$
\mathrm{P}\left(\|\xi\|^2 - \mathbb{E}\|\xi\|^2 > t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{d}, t\right\}\right).
$$

Since $2 \geq \mathbb{E}\exp\left((\xi^i)^2\right) \geq 1 + \mathbb{E}(\xi^i)^2$, we have $\mathbb{E}\|\xi\|^2 = \sum_{i=1}^d \mathbb{E}(\xi^i)^2 \leq d$.

Select $t = \max\left\{\sqrt{\frac{d\log\left(1/\bar{\delta}\right)}{c}}, \frac{\log\left(1/\bar{\delta}\right)}{c}\right\}$, then $\mathrm{P}\left(\|\xi\|^2 - \mathbb{E}\|\xi\|^2 > t\right) \leq \bar{\delta}$. Therefore, with probability at least $1 - \bar{\delta}$ w.r.t

ξ,

$$\begin{aligned}
\|\xi\|^2 \leq& \mathbb{E}\|\xi\|^2 + \max\left\{\sqrt{\frac{d\log\left(1/\bar{\delta}\right)}{c}}, \frac{\log\left(1/\bar{\delta}\right)}{c}\right\} \\
&\leq d + \sqrt{\frac{d\log\left(1/\bar{\delta}\right)}{c}} + \frac{\log\left(1/\bar{\delta}\right)}{c} \\
&\leq 9d(\log\left(1/\bar{\delta}\right)). && (d \text{ is sufficiently large since } C \text{ is sufficiently large; } \bar{\delta} < 0.5)
\end{aligned}$$

As for the last statement,

$$\begin{aligned}
&\mathbb{E}\left[\|\xi\|\,\mathbb{1}\left(\mathcal{E}\right)\right] \\
&= \int_0^\infty \mathrm{P}\left(\|\xi\|\,\mathbb{1}\left(\mathcal{E}\right) \geq t\right) dt \\
&= \int_0^{6\sqrt{\log\left(1/\bar{\delta}\right)d}} \mathrm{P}\left(\|\xi\|\,\mathbb{1}\left(\mathcal{E}\right) \geq t\right) dt + \int_{6\sqrt{\log\left(1/\bar{\delta}\right)d}}^\infty \mathrm{P}\left(\|\xi\|\,\mathbb{1}\left(\mathcal{E}\right) \geq t\right) dt \\
&\leq 6\sqrt{\log\left(1/\bar{\delta}\right)d}\cdot \mathrm{P}\left(\|\xi\|\,\mathbb{1}\left(\mathcal{E}\right) \geq 6\sqrt{\log\left(1/\bar{\delta}\right)d}\right) + \int_{6\sqrt{\log\left(1/\bar{\delta}\right)d}}^\infty \exp\left(-\frac{t^2}{9d}\right) dt \quad \left(\mathrm{P}\left(\|\xi\|^2 > 9d(\log\left(1/\bar{\delta}\right))\right) \leq \bar{\delta}\right) \\
&\leq 6\sqrt{\log\left(1/\bar{\delta}\right)d}\cdot \bar{\delta}^4 + \int_0^\infty \exp\left(-\frac{\left(t + 6\sqrt{\log\left(1/\bar{\delta}\right)d}\right)^2}{9d}\right) dt \\
&\leq 6\sqrt{\log\left(1/\bar{\delta}\right)d}\cdot \bar{\delta}^4 + \int_0^\infty \exp\left(-\frac{12t\sqrt{\log\left(1/\bar{\delta}\right)d} + 36\log\left(1/\bar{\delta}\right)d}{9d}\right) dt \\
&= 6\sqrt{\log\left(1/\bar{\delta}\right)d}\cdot \bar{\delta}^4 + \bar{\delta}^4\frac{9d}{12\sqrt{\log\left(1/\bar{\delta}\right)d}} \\
&\leq 8\bar{\delta}^4\cdot\sqrt{d\cdot\log\left(1/\bar{\delta}\right)}. && (\bar{\delta} < 0.5)
\end{aligned}$$

□

Before proceeding to the next Lemma, we define some important notations which will be used frequently later. We define

$$\lambda_i(\mathrm{x}; \mathrm{W}^t) = \frac{1}{m}\sum_{s=1}^m \frac{\phi\left(\left\langle \mathrm{w}_s^t - \eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle\right) - \phi(\langle \mathrm{w}_s^t, \mathrm{x}\rangle)}{\left\langle -\eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle}\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle),$$

so that the following holds:

$$\begin{aligned}
&yf(\mathrm{x}; \mathrm{W}^{t+1}) - yf(\mathrm{x}; \mathrm{W}^t) \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \frac{\phi\left(\left\langle \mathrm{w}_s^t - \eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle\right) - \phi(\langle \mathrm{w}_s^t, \mathrm{x}\rangle)}{\left\langle -\eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle}\left\langle -\eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), y\mathrm{x}\right\rangle \\
&= \frac{\eta}{mn}\sum_{i=1}^n\sum_{s=1}^m \frac{\phi\left(\left\langle \mathrm{w}_s^t - \eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle\right) - \phi(\langle \mathrm{w}_s^t, \mathrm{x}\rangle)}{\left\langle -\eta\nabla_{\mathrm{w}_s}\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{x}\right\rangle}\tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle)\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle \\
&= \frac{\eta}{n}\sum_{i=1}^n \tilde{g}_i(\mathrm{W}^t)\lambda_i(\mathrm{x}; \mathrm{W}^t)\left\langle y_i\tilde{\mathrm{x}}_i^t, y\mathrm{x}\right\rangle.
\end{aligned}$$

For any $t \geq 1$, We define $\mathcal{A}(t)$ as the set of pairs (i, s) such that the neurons s is active for the adversarial examples $\forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)$; i.e.,

$$\mathcal{A}(t) := \left\{(i,s) \in [n] \times [m] : \left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \right\rangle > 0, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)\right\}.$$

For notation simplicity, we also define

$$\mathcal{A}(0) := \left\{(i,s) \in [n] \times [m] : \left\langle \mathrm{w}_s^t, \mathrm{x}_i \right\rangle > 0\right\}.$$

Denote its coordinate as

$$\begin{aligned}\mathcal{A}^i(t) &: = \left\{s \in [m] : (i,s) \in \mathcal{A}(t)\right\},\\ \mathcal{A}_s(t) &: = \left\{i \in [n] : (i,s) \in \mathcal{A}(t)\right\}.\end{aligned}$$

Proposition B.24. For any $t \geq 1$ and any pair $(i, s) \in \mathcal{A}(t)$, we have $\forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)$, $\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle) \geq \gamma$. Moreover, $\lambda_i(\tilde{\mathrm{x}}_i; \mathrm{W}^t) \geq \gamma^2 \left|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)\right| / m$.

Proof. The definition of $\mathcal{A}(t)$ implies that $\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle > 0, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)$. By the definition of activation function we have $\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle) \geq \gamma$.

The definition of $\lambda_i(\tilde{\mathrm{x}}_i; \mathrm{W}^t)$ gives us the following:

$$\begin{aligned}\lambda_i(\tilde{\mathrm{x}}_i; \mathrm{W}^t) &\geq \frac{\gamma}{m} \sum_{s \in \mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)} \frac{\phi\left(\left\langle \mathrm{w}_s^{t+1}, \tilde{\mathrm{x}}_i \right\rangle\right) - \phi(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle)}{\left\langle \mathrm{w}_s^{t+1}, \tilde{\mathrm{x}}_i \right\rangle - \langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle}\\ &\geq \frac{\gamma^2}{m} \left|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)\right|.\end{aligned}$$

□

We further define $\mathcal{T} := \{(i,s) \in [n] \times [m] : y_i = a_s\}$ and similarly we denote

$$\begin{aligned}\mathcal{T}^i &: = \left\{s \in [m] : (i,s) \in \mathcal{T}\right\},\\ \mathcal{T}_s &: = \left\{i \in [n] : (i,s) \in \mathcal{T}\right\}\end{aligned}$$

Lemma B.25. On a good run we have

$$\begin{aligned}\left|\mathcal{A}^i(1) \cap \mathcal{T}^i\right| &\geq \left|\mathcal{A}^i(0) \cap \mathcal{T}^i\right| \geq c_0 m,\\ \left|\mathcal{A}_s(1) \cap \mathcal{T}_s\right| &\geq \left|\mathcal{A}_s(0) \cap \mathcal{T}_s\right| \geq c_0 n.\end{aligned}$$

Proof of Lemma B.25. The proof of Lemma B.21 implies that $\forall i \in [n], s \in [m]$,

$$\begin{aligned}\left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^1, \tilde{\mathrm{x}}_i \right\rangle > 0, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)\right\}\right| &\geq \left|\left\{i \in [n] : y_i = a_s, \left\langle \mathrm{w}_s^0, \mathrm{x}_i \right\rangle > 0\right\}\right| \geq c_0 n,\\ \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^1, \tilde{\mathrm{x}}_i \right\rangle > 0, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha)\right\}\right| &\geq \left|\left\{s \in [m] : y_i = a_s, \left\langle \mathrm{w}_s^0, \tilde{\mathrm{x}}_i \right\rangle > 0\right\}\right| \geq c_0 m.\end{aligned}$$

Combine with the definition of $\mathcal{A}$ and $\mathcal{T}$ conclude the proof. □

In the following Lemma, we will 1) prove the number of neurons with positive activation increases as the training epochs increases; 2) provide both an upper bound and a lower bound on the increment in the un-normalized margin for arbitrary adversarial training examples; 3) show the loss g is at the same scale across all adversarial training examples. An analog of Lemma B.26 is Lemma B.9 for neural networks with smooth activation functions.

Lemma B.26. On a good run, there exists a constant $C_r > 0$ that only depends on κ, γ such that for any $t \geq 0$, we have

(E1) $\mathcal{A}(t) \cap \mathcal{T} \subset \mathcal{A}(t+1) \cap \mathcal{T}$.

(E2) $\frac{\eta c_0 \gamma^2 \tilde{g}_i(\mathrm{W}^t)}{2n} \left(\sqrt{\frac{d}{C_1}} - \alpha\right)^2 \leq y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^{t+1}) - y_i f(\tilde{\mathrm{x}}_i; \mathrm{W}^t) \leq \frac{3\eta}{n}\left(C_1 d + \alpha^2\right) \tilde{g}_i(\mathrm{W}^t), \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i, \alpha), \forall i \in [n]$.

(E3) $\max_{i,j\in[n]} \frac{g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t))} \le C_r$.

Here c_0 is the constant introduced in Lemma B.21

Proof of Lemma B.26. We prove via induction. Recall $\tilde{\mathrm{x}}_i^t = \arg\max_{\tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha)} \ell(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t))$. Similar as network with smooth activation functions, we have the following

$$\max_{i,j\in[n]} \frac{g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))}{g(y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t))} \le \max\left(2, 2\cdot \max_{i,j\in[n]} \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t)\right)}\right).$$

Therefore we only need $\max\limits_{i,j\in[n]} \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^t;\mathrm{W}^t)\right)} \le C_r/2$ to hold. Note that at initialization, $\max\limits_{i,j\in[n]} \frac{\exp\left(-y_i f(\tilde{\mathrm{x}}_i^0;\mathrm{W}^0)\right)}{\exp\left(-y_j f(\tilde{\mathrm{x}}_j^0;\mathrm{W}^0)\right)} = 1 \le C_r/2$.

Without loss of generality, we choose $i=1, j=2$. Through induction, at iteration t, we have $\frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^t;\mathrm{W}^t)\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^t;\mathrm{W}^t)\right)} \le C_r/2$. Now we are proving with the following order: (E1), (E2) and (E3) for $t+1$. The $t=0$ case of (E1) is proved in Lemma B.21. For any $t \ge 1$ and $(i,s) \in \mathcal{A}(t)\cap\mathcal{T}$, we have $y_i a_s = 1$ and $\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i\rangle > 0, \forall \tilde{\mathrm{x}}_i \in \mathcal{B}_2(\mathrm{x}_i,\alpha) > 0$ by definition, we have

$$\begin{aligned}
&\left\langle \mathrm{w}_s^{t+1}, \tilde{\mathrm{x}}_i\right\rangle - \left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i\right\rangle \\
&= \frac{\eta}{n\sqrt{m}} \sum_{k=1}^n y_k a_s \tilde{g}_k(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_k^t\rangle)\left\langle \tilde{\mathrm{x}}_k^t, \tilde{\mathrm{x}}_i\right\rangle \\
&= \frac{\eta}{n\sqrt{m}} \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle)\left\langle \tilde{\mathrm{x}}_i^t, \tilde{\mathrm{x}}_i\right\rangle + \frac{\eta}{n\sqrt{m}} \sum_{k\neq i} y_k a_s \tilde{g}_k(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_k^t\rangle)\left\langle \tilde{\mathrm{x}}_k^t, \tilde{\mathrm{x}}_i\right\rangle \\
&\ge \frac{\eta}{n\sqrt{m}} \frac{\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)}{C_r}\gamma\left(\sqrt{d/C_1} - \alpha\right)^2 \\
&\qquad - \frac{\eta}{\sqrt{m}}\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)\left(C_1\left(\|\mu\|^2 + \sqrt{d\log(n/\delta)}\right) + 2\alpha\sqrt{C_1 d} + \alpha^2\right) \\
&\ge \frac{\eta}{n\sqrt{m}} \frac{\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)}{C_r}\gamma\left(\sqrt{d/C_1} - \alpha\right)^2 - \frac{\eta}{n\sqrt{m}}\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t) d\left(\frac{C_1+1}{C} + \frac{\sqrt{5}C_1}{\sqrt{C}}\right) && \text{(Assumption (1))} \\
&\ge \frac{\eta}{2n\sqrt{m}} \frac{\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)}{C_r}\gamma\left(\sqrt{d/C_1} - \alpha\right)^2 > 0. && (\alpha \le \|\mu\| \le \tfrac{1}{2}\sqrt{d/C_1};\ C \text{ sufficiently large})
\end{aligned}$$

This implies that $(i,s) \in \mathcal{A}(t+1)\cap\mathcal{T}$, therefore (E1) holds.

Next we consider the following for any $t \ge 1$

$$\begin{aligned}
&y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^{t+1}) - y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t) \\
&= \frac{\eta}{n}\sum_{k=1}^n \tilde{g}_k(\mathrm{W}^t)\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle y_k\tilde{\mathrm{x}}_k^t, y_i\tilde{\mathrm{x}}_i\right\rangle \\
&= \frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\lambda_i(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle \tilde{\mathrm{x}}_i^t, \tilde{\mathrm{x}}_i\right\rangle + \frac{\eta}{n}\sum_{k\neq i}\tilde{g}_k(\mathrm{W}^t)\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle y_k\tilde{\mathrm{x}}_k^t, y_i\tilde{\mathrm{x}}_i\right\rangle.
\end{aligned}$$

The first term gives us

$$\begin{aligned}
\frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\lambda_i(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle \tilde{\mathrm{x}}_i^t, \tilde{\mathrm{x}}_i\right\rangle &\ge \frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\lambda_i(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left(\sqrt{d/C_1} - \alpha\right)^2 \\
&\ge \frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\frac{\gamma^2}{m}\left|\mathcal{A}^i(t)\cap\mathcal{A}^i(t+1)\right|\left(\sqrt{d/C_1} - \alpha\right)^2 && \text{(Proposition B.24)} \\
&\ge \frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\frac{\gamma^2}{m}\left|\mathcal{A}^i(0)\cap\mathcal{T}^i\right|\left(\sqrt{d/C_1} - \alpha\right)^2 && \text{((E1))}
\end{aligned}$$

$$\geq \frac{\eta\gamma^2 c_0}{n}\left(\sqrt{d/C_1}-\alpha\right)^2 \tilde{g}_i(\mathrm{W}^t). \qquad \text{(Lemma B.25)}$$

On the other hand,

$$\frac{\eta}{n}\tilde{g}_i(\mathrm{W}^t)\lambda_i(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle \tilde{\mathrm{x}}_i^t, \tilde{\mathrm{x}}_i\right\rangle \leq \frac{\eta}{n}\left(\sqrt{C_1 d}+\alpha\right)^2 \tilde{g}_i(\mathrm{W}^t).$$

The second terms tells us that

$$\begin{aligned}
&\frac{\eta}{n}\left|\sum_{k\neq i}\tilde{g}_k(\mathrm{W}^t)\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle y_k\tilde{\mathrm{x}}_k^t, y_i\tilde{\mathrm{x}}_i\right\rangle\right| \\
&\leq \eta\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
&\leq \frac{\eta d\tilde{g}_i(\mathrm{W}^t)}{n}. \qquad (\widehat{G}_{\text{rob}}(\mathrm{W}^t)\leq C_r\tilde{g}_i(\mathrm{W}^t);\ \text{Assumption (1) for large enough } C)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{\eta}{n}\sum_{k\neq i}\tilde{g}_k(\mathrm{W}^t)\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\left\langle y_k\tilde{\mathrm{x}}_k^t, y_i\tilde{\mathrm{x}}_i\right\rangle \\
&\geq -\eta\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
&\geq -\eta C_r\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\tilde{g}_i(\mathrm{W}^t) \qquad (\widehat{G}_{\text{rob}}(\mathrm{W}^t)\leq C_r\tilde{g}_i(\mathrm{W}^t)) \\
&\geq -\frac{\eta\gamma^2 c_0}{2n}\left(0.5\sqrt{d/C_1}\right)^2\tilde{g}_i(\mathrm{W}^t) \qquad (\text{Assumption (1) for large enough } C) \\
&\geq -\frac{\eta\gamma^2 c_0}{2n}\left(\sqrt{d/C_1}-\alpha\right)^2\tilde{g}_i(\mathrm{W}^t).
\end{aligned}$$

Summing the two terms together we have (E2) holds for any $t\geq 1$. Now let's look at the $t=0$ case of (E2). From the proof of Lemma B.21, for any $i\in[n]$, there are at least $c_0 m$ many $s\in[m]$, such that $y_i=a_s$ and $\left\langle \mathrm{w}_s^1,\tilde{\mathrm{x}}_i\right\rangle \geq \frac{\eta}{5n\sqrt{m}}\frac{\gamma}{C_1}d$. Combining the fact that $|\left\langle \mathrm{w}_s^0,\tilde{\mathrm{x}}_i\right\rangle| \leq \sqrt{\frac{3}{2}}\omega_{\text{init}}\sqrt{d}2\sqrt{C_1 d}\leq \frac{\eta\sqrt{6C_1 d}}{\sqrt{m}}\leq \frac{\eta}{n\sqrt{m}}\frac{\sqrt{6C_1}}{\sqrt{C\log(2)}}d$ and C is sufficiently large, we know that

$$\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^0)=\frac{1}{m}\sum_{s=1}^{m}\frac{\phi\left(\left\langle \mathrm{w}_s^1,\tilde{\mathrm{x}}_i\right\rangle\right)-\phi(\left\langle \mathrm{w}_s^0,\tilde{\mathrm{x}}_i\right\rangle)}{\left\langle \mathrm{w}_s^1-\mathrm{w}_s^0,\tilde{\mathrm{x}}_i\right\rangle}\phi'(\left\langle \mathrm{w}_s^0,\tilde{\mathrm{x}}_k^0\right\rangle)\in[\frac{c_0\gamma^2}{1.1},1].$$

$$y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^1)-y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^0)=\frac{\eta}{n}\sum_{k=1}^{n}\tilde{g}_k(\mathrm{W}^0)\lambda_k(\tilde{\mathrm{x}}_i;\mathrm{W}^0)\left\langle y_k\tilde{\mathrm{x}}_k^0, y_i\tilde{\mathrm{x}}_i\right\rangle$$

Similar with the logic for $t\geq 1$, we get

$$\begin{aligned}
&y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^1)-y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^0) \\
&\geq \frac{\eta\gamma^2 c_0}{1.1n}\left(\sqrt{d/C_1}-\alpha\right)^2\tilde{g}_i(\mathrm{W}^0)-\eta C_r\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\tilde{g}_i(\mathrm{W}^0) \\
&\geq \frac{\eta\gamma^2 c_0}{2n}\left(\sqrt{d/C_1}-\alpha\right)^2\tilde{g}_i(\mathrm{W}^0)
\end{aligned}$$

and

$$\begin{aligned}
y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^1)-y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^0) &\leq \frac{\eta}{n}(\sqrt{C_1 d}+\alpha)^2\tilde{g}_i(\mathrm{W}^0)+\eta C_r\left(C_1\left(\|\mu\|^2+\sqrt{d\log(n/\delta)}\right)+2\alpha\sqrt{C_1 d}+\alpha^2\right)\tilde{g}_i(\mathrm{W}^0) \\
&\leq \frac{3\eta}{n}(C_1 d+\alpha^2)\tilde{g}_i(\mathrm{W}^0)
\end{aligned}$$

Now the proof of (E2) is complete. This implies that $y_i f(\tilde{\mathrm{x}}_i^t; \mathrm{W}^t) \geq 0, \forall t \geq 0, i \in [n]$. Applying the above gives us the following,

$$
\begin{aligned}
&\frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t+1}; \mathrm{W}^{t+1})\right)} \\
&= \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t})\right)} \cdot \frac{\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^{t}) - y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t}) - y_2 f(\tilde{\mathrm{x}}_2^{t+1}; \mathrm{W}^{t+1})\right)} \\
&\leq \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t})\right)} \cdot \frac{\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^{t}) - y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t}) - y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t+1})\right)} \\
&\leq \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t})\right)} \cdot \exp\left(\frac{3\eta}{n}\left(C_1 d + \alpha^2\right) \tilde{g}_2(\mathrm{W}^t) - \frac{\eta c_0 \gamma^2}{2n}\left(\sqrt{\frac{d}{C_1}} - \alpha\right)^2 \tilde{g}_1(\mathrm{W}^t)\right) \\
&= \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t})\right)} \cdot \exp\left(\frac{\eta c_0 \gamma^2}{2n}\left(\sqrt{\frac{d}{C_1}} - \alpha\right)^2 \tilde{g}_2(\mathrm{W}^t)\left(\frac{\frac{3\eta}{n}\left(C_1 d + \alpha^2\right)}{\frac{\eta c_0 \gamma^2}{2n}\left(\sqrt{\frac{d}{C_1}} - \alpha\right)^2} - \frac{\tilde{g}_1(\mathrm{W}^t)}{\tilde{g}_2(\mathrm{W}^t)}\right)\right),
\end{aligned}
$$

where the first inequality holds since $\exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^t)\right) \geq \exp\left(y_1 f(\tilde{\mathrm{x}}_1^{t}; \mathrm{W}^t)\right)$, $\exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t+1}; \mathrm{W}^{t+1})\right) \leq \exp\left(y_2 f(\tilde{\mathrm{x}}_2^{t}; \mathrm{W}^{t+1})\right)$ by the definition of $\tilde{\mathrm{x}}_1^t, \tilde{\mathrm{x}}_2^{t+1}$.

If $\frac{\tilde{g}_1(\mathrm{W}^t)}{\tilde{g}_2(\mathrm{W}^t)} \geq \frac{\frac{3\eta}{n}\left(C_1 d+\alpha^2\right)}{\frac{\eta c_0 \gamma^2}{2n}\left(\sqrt{d/C_1}-\alpha\right)^2} = \frac{6\left(C_1 d+\alpha^2\right)}{c_0\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}$, then $\frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t+1};\mathrm{W}^{t+1})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t+1};\mathrm{W}^{t+1})\right)} \leq \frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t};\mathrm{W}^{t})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t};\mathrm{W}^{t})\right)} \leq C_r/2$. Otherwise we have

$$
\begin{aligned}
&\frac{\exp\left(-y_1 f(\tilde{\mathrm{x}}_1^{t+1}; \mathrm{W}^{t+1})\right)}{\exp\left(-y_2 f(\tilde{\mathrm{x}}_2^{t+1}; \mathrm{W}^{t+1})\right)} \\
&\leq \frac{2\tilde{g}_1(\mathrm{W}^t)}{\tilde{g}_2(\mathrm{W}^t)} \cdot \exp\left(\frac{3\eta}{n}\left(C_1 d + \alpha^2\right) \tilde{g}_2(\mathrm{W}^t) - \frac{\eta c_0 \gamma^2}{2n}\left(\sqrt{d/C_1} - \alpha\right)^2 \tilde{g}_1(\mathrm{W}^t)\right) \\
&\leq \frac{12\left(C_1 d + \alpha^2\right)}{c_0 \gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2} \exp\left(\frac{3\eta}{n}\left(C_1 d + \alpha^2\right)\right) \qquad (\tilde{g}_i(\mathrm{W}^t) \leq 1) \\
&\leq \frac{24\left(C_1 d + \alpha^2\right)}{c_0 \gamma^2\left(\sqrt{d/C_1} - \alpha\right)^2} \leq C_r/2,
\end{aligned}
$$

where the last line holds due to $\exp\left(\frac{3\eta}{n}\left(C_1 d + \alpha^2\right)\right) \leq 2$ for $\eta \leq 1/Cd^2$ with $C \geq \frac{6C_1}{\log(2)}$ given by Assumption (4). Assumption (1) and Assumption (6) gives us that $\alpha \leq \|\mu\| \leq 0.5\sqrt{\frac{d}{C_1}}$. As a result, there exists a constant $C_r = \frac{192 C_1 (C_1 + 1/(4C_1))}{c_0 \gamma^2} \geq \frac{48\left(C_1 d+\alpha^2\right)}{c_0\gamma^2\left(\sqrt{d/C_1}-\alpha\right)^2}$ such that the Lemma statement holds, where C_1 comes from Lemma B.4. □

Similar with the smooth activation setting, we can characterize a property of the adversarial training example $\tilde{\mathrm{x}}_i^t$ using Lemma B.26: during training the perturbed data $\tilde{\mathrm{x}}_i^t$ is close to the linear subspace $\mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ in $\mathbb{R}^d$.

Lemma B.27. $\forall t \in \mathbb{N}$ and $i \in [n]$, the distance between $\tilde{\mathrm{x}}_i^t$ and $\mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ satisfies $\mathrm{dist}(\tilde{\mathrm{x}}_i^t, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) \leq \min\left\{\frac{\omega_{\text{init}}\sqrt{md}}{\eta}, \alpha\right\}$.

Proof of Lemma B.27. We define $C_d = \frac{\omega_{\text{init}}\sqrt{md}}{\eta}$ for simplicity. The upper bound α is obvious because the perturbation size is α. Now we look at C_d. We prove the result via induction. Consider time $t = 0$, from the symmetric initialization, for any given x, we have $f(\mathrm{x}; \mathrm{W}^0) = 0$ is a constant function. Therefore, for any given training data x_i, generating the adversarial examples by adding any perturbations on x_i cannot increase the training loss. For simplicity, we consider the algorithm runs standard GD at time $t = 0$; i.e. no adversarial training examples are generated for the first step, the adversarial training process starts at $t \geq 1$. This gives us that $\mathrm{dist}(\tilde{\mathrm{x}}_i^0, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) = \mathrm{dist}(\mathrm{x}_i, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) = 0 \leq C_d$. Suppose we have $\mathrm{dist}(\tilde{\mathrm{x}}_i^s, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) \leq C_d$ holds for any $0 \leq s \leq t-1$, and we will now prove the result for t.

Recall $\tilde{\mathrm{x}}_k^t = \mathrm{argmax}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_k, \alpha)} \ell(y_k f(\tilde{\mathrm{x}}; \mathrm{W}^t))$. We decompose $\tilde{\mathrm{x}}_k^t = \tilde{\mathrm{x}}_{k,\|}^t + \tilde{\mathrm{x}}_{k,\perp}^t$, where $\tilde{\mathrm{x}}_{k,\|}^t \in \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ and $\tilde{\mathrm{x}}_{k,\perp}^t \perp \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$. Assume $\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 > C_d$, and we will prove via contradiction. As the loss function is monotonically decreasing, $\tilde{\mathrm{x}}_k^t = \mathrm{argmin}_{\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_k, \alpha)} y_k f(\tilde{\mathrm{x}}; \mathrm{W}^t)$. As a result, there is no feasible direction that is also a descent direction. Here we construct directions $\mathrm{v}_\theta = -\tilde{\mathrm{x}}_{k,\perp}^t - \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i)$ for every $\theta \in \mathbb{R}$ that satisfies $0 < \theta < \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{(\alpha^2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2) \cdot 8C_1 dn}}$. We have that

$$\begin{aligned}
\left\langle \tilde{\mathrm{x}}_k^t - \mathrm{x}_k, \mathrm{v}_\theta \right\rangle &= \left\langle \tilde{\mathrm{x}}_{k,\perp}^t + \tilde{\mathrm{x}}_{k,\|}^t - \mathrm{x}_k, -\tilde{\mathrm{x}}_{k,\perp}^t - \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \left\langle \tilde{\mathrm{x}}_{k,\|}^t - \mathrm{x}_k, -\theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + \left\langle \tilde{\mathrm{x}}_{k,\perp}^t, -\tilde{\mathrm{x}}_{k,\perp}^t \right\rangle \\
&\leq \theta \|\tilde{\mathrm{x}}_{k,\|}^t - \mathrm{x}_k\|_2 \cdot \|\sum_{i=1}^n y_i \mathrm{x}_i\|_2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2 \\
&\leq \theta \sqrt{(\alpha^2 - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2) \cdot 8C_1 dn} - \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2 < 0, && \text{(Lemma B.4 (C4))}
\end{aligned}$$

therefore v_θ are feasible directions. From the above discussion, we know that v_θ cannot be descent directions. Pick $\theta = \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{8\alpha^2 C_1 dn}}$. Since sufficient neurons are activated for every $\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x}_k, \alpha)$, we know there exist at least $c_0 m$ many $s \in [m]$ such that $\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_k^t \right\rangle > 0$. From the form of the classifier $y_k f(\tilde{\mathrm{x}}; \mathrm{W}^t) = y_k \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \phi(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}} \rangle)$, and combining the fact that ϕ is strictly increasing on $(0, +\infty)$, there exists an s_0 such that $y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^t, \mathrm{v}_\theta \right\rangle \geq 0$.

$$\begin{aligned}
&0 \geq y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^t, -\mathrm{v}_\theta \right\rangle \\
&= \sum_{t'=0}^{t-1} y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^{t'+1} - \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \sum_{t'=0}^{t-1} y_k a_{s_0} \left\langle \frac{\eta a_{s_0}}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k'}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&= \sum_{t'=0}^{t-1} \left\langle \frac{\eta}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k',\|}^{t'}, \theta (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&\qquad + \sum_{t'=0}^{t-1} y_k \left\langle \frac{\eta}{n\sqrt{m}} \sum_{k'=1}^n \tilde{g}_{k'}(\mathrm{W}^{t'}) \phi'(\left\langle \mathrm{w}_{s_0}^{t'}, \tilde{\mathrm{x}}_{k'}^{t'} \right\rangle) y_{k'} \tilde{\mathrm{x}}_{k',\perp}^{t'}, \tilde{\mathrm{x}}_{k,\perp}^t \right\rangle + y_k a_{s_0} \left\langle \mathrm{w}_{s_0}^0, \tilde{\mathrm{x}}_{k,\perp}^t + \theta y_k (\sum_{i=1}^n y_i \mathrm{x}_i) \right\rangle \\
&\geq \theta \sum_{t'=0}^{t-1} \frac{c_0 \gamma \eta d \widehat{G}_{rob}(\mathrm{W}^{t'})}{8C_1 C_r \sqrt{m}} - \sum_{t'=0}^{t-1} \frac{C_d \eta}{\sqrt{m}} \widehat{G}_{rob}(\mathrm{W}^{t'}) \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 - 2\omega_{\text{init}} \sqrt{d} (\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 + \theta \sqrt{8C_1 dn}) \\
&\qquad\qquad (\|\tilde{\mathrm{x}}_{k',\perp}^{t'}\|_2 \leq C_d \text{ from induction, and sufficiently many neurons activated}) \\
&\geq \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{8\alpha^2 C_1 dn}} \sum_{t'=0}^{t-1} \frac{c_0 \gamma \eta d \widehat{G}_{rob}(\mathrm{W}^{t'})}{8C_1 C_r \sqrt{m}} - \sum_{t'=0}^{t-1} \frac{C_d \eta}{\sqrt{m}} \widehat{G}_{rob}(\mathrm{W}^{t'}) \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 - 2\omega_{\text{init}} \sqrt{d} (\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 + \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\alpha}) \qquad \text{(plug in } \theta) \\
&\geq \frac{\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2^2}{\sqrt{\alpha^2 C_1 dn}} \sum_{t'=0}^{t-1} \frac{c_0 \gamma \eta d \widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1 C_r \sqrt{m}} - \sum_{t'=0}^{t-1} \frac{C_d \eta}{\sqrt{m}} \widehat{G}_{rob}(\mathrm{W}^{t'}) \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 - 2\omega_{\text{init}} \sqrt{d} \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 \\
&\qquad\qquad (\omega_{\text{init}} \leq \tfrac{\eta}{\sqrt{md}} \leq \tfrac{\eta}{\sqrt{Cmn}} \text{ and } C \text{ sufficiently large}) \\
&\geq \frac{C_d \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2}{\sqrt{\alpha^2 C_1 dn}} \sum_{t'=0}^{t-1} \frac{c_0 \gamma \eta d \widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1 C_r \sqrt{m}} - \sum_{t'=0}^{t-1} \frac{C_d \eta}{\sqrt{m}} \widehat{G}_{rob}(\mathrm{W}^{t'}) \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 - 2\omega_{\text{init}} \sqrt{d} \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{C_d\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2}{\sqrt{\alpha^2 C_1 dn}} \sum_{t'=0}^{t-1} \frac{c_0\gamma\eta d\widehat{G}_{rob}(\mathrm{W}^{t'})}{32C_1C_r\sqrt{m}} - 5\sum_{t'=0}^{t-1}\frac{C_d\eta}{\sqrt{m}}\widehat{G}_{rob}(\mathrm{W}^{t'})\|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 \qquad \Big(\sum_{t'=0}^{t-1}\widehat{G}_{rob}(W^{t'}) \geq \widehat{G}_{rob}(W^0) = \tfrac{1}{2}\Big) \\
&> 0. \qquad (d \geq Cn\alpha^2 \text{ from Assumption (1), and C sufficiently large})
\end{aligned}
$$

This is a contradiction. Therefore, we have proved $\mathrm{dist}(\tilde{\mathrm{x}}_k^t, \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}) = \|\tilde{\mathrm{x}}_{k,\perp}^t\|_2 \leq C_d$. By induction, proof is complete. □

Similar with the smooth activation setting, we can prove a different version of Lemma B.4 (C5) and (C6) that will be used later.

Lemma B.28. $\forall i \in \mathcal{C}$, $\frac{1}{3}\|\mu\|^2 \leq \langle \mu, y_i\tilde{\mathrm{x}}_i^t\rangle \leq 3\|\mu\|^2$. $\forall i \in \mathcal{N}$, $-3\|\mu\|^2 \leq \langle \mu, y_i\tilde{\mathrm{x}}_i^t\rangle \leq -\frac{1}{3}\|\mu\|^2$.

Proof of Lemma B.28. From Lemma B.4 (C5) and (C6), we know that $\frac{1}{2}\|\mu\|^2 \leq \langle \mu, y_i\mathrm{x}_i\rangle \leq 2\|\mu\|^2$ holds for all $i \in \mathcal{C}$, and $-2\|\mu\|^2 \leq \langle \mu, y_i\mathrm{x}_i\rangle \leq -\frac{1}{2}\|\mu\|^2$ holds for all $i \in \mathcal{N}$. Therefore, it suffices to prove $|\langle \mu, y_i\tilde{\mathrm{x}}_i^t\rangle - \langle \mu, y_i\mathrm{x}_i\rangle| \leq \frac{1}{6}\|\mu\|^2$. We can decompose $\tilde{\mathrm{x}}_i^t - \mathrm{x}_i = (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|} + (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\perp}$, where $(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|} \in \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ and $(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\perp} \perp \mathrm{span}\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$. From Lemma B.10, $\|(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\perp}\|_2 \leq \min\{C_d, \alpha\} \leq C_d \leq 1$. For the parallel component, we can write $(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|} = \sum_{k=1}^n z_k\mathrm{x}_k$, where $z_k \in \mathbb{R}$. From Lemma B.4 (C4), $\alpha^2 \geq \|\tilde{\mathrm{x}}_i^t - \mathrm{x}_i\|_2^2 \geq \|(\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|}\|_2^2 \geq \frac{d}{8C_1}\cdot\sum_{k=1}^n z_k^2$. Thus, $\sqrt{\frac{8C_1n\alpha^2}{d}} \geq \sqrt{n\sum_{k=1}^n z_k^2} \geq \sum_{k=1}^n |z_k|$.
Now we can prove the statement.

$$
\begin{aligned}
|\langle \mu, y_i\tilde{\mathrm{x}}_i^t\rangle - \langle \mu, y_i\mathrm{x}_i\rangle| &= |\langle \mu, \tilde{\mathrm{x}}_i^t - \mathrm{x}_i\rangle| \\
&\leq |\langle \mu, (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\|}\rangle| + |\langle \mu, (\tilde{\mathrm{x}}_i^t - \mathrm{x}_i)_{\perp}\rangle| \\
&\leq \sum_{k=1}^n |z_k|\cdot|\langle \mu, \mathrm{x}_k\rangle| + C_d\|\mu\| \\
&\leq \sqrt{\frac{8C_1n\alpha^2}{d}}\cdot 2\|\mu\|^2 + C_d\|\mu\| \\
&\leq \frac{1}{6}\|\mu\|^2. \qquad \text{(Assumption (B2), Assumption (1) and } C \text{ being sufficiently large)}
\end{aligned}
$$

□

Lemma B.29, an analog of Lemma B.15, aims at providing the convergence guarantees of robust training loss for networks with non-smooth activation functions.

Lemma B.29. For a non-smooth homogeneous activation function ϕ, provided $C > 1$ is sufficiently large, then on a good run, the robust training loss satisfies

$$
\widehat{L}_{\mathrm{rob}}(\mathrm{W}^T) \leq \frac{30}{\|\mu\|\,\gamma c_0\sqrt{\eta T}}.
$$

where $c_0 > 1$ is the constant in Lemma B.21.

Proof of Lemma B.29. The proof is similar with Lemma B.15. We first need to show a lower bound for $\left\|\nabla\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t)\right\|_F = \sup_{\mathrm{U}:\|\mathrm{U}\|_F=1}\left\langle -\nabla\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{U}\right\rangle$, and it suffices to construct a matrix V with Frobenius norm at most one such that $\left\langle -\nabla\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{V}\right\rangle$ is bounded from below by a positive constant. To this end, choose $\mathrm{V} \in \mathbb{R}^{m\times d}$ be the matrix with rows $\mathrm{v}_s = \frac{1}{\sqrt{m}}a_s\mu/\|\mu\|, \forall s \in [m]$. Then $\|\mathrm{V}\|_F = 1$ since $a_s = \pm 1$, and we have for any W^t,

$$
\langle \nabla f(\tilde{\mathrm{x}}_i; \mathrm{W}^t), \mathrm{V}\rangle = \sum_{s=1}^m \frac{1}{\sqrt{m}}a_s\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i\rangle)\langle \mathrm{v}_s, \tilde{\mathrm{x}}_i\rangle = \left\langle \frac{\mu}{\|\mu\|}, \tilde{\mathrm{x}}_i\right\rangle\frac{1}{m}\sum_{s=1}^m \phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i\rangle).
$$

$$
\begin{aligned}
1 &\geq \frac{1}{m}\sum_{s=1}^{m} \phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle) \\
&\geq \frac{1}{m}\sum_{s\in\mathcal{A}^i(t)\cap\mathcal{T}^i} \phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle) && \text{(Only count the neurons that satisfy } \langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle > 0) \\
&\geq \frac{1}{m}\sum_{s\in\mathcal{A}^i(0)\cap\mathcal{T}^i} \phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i \rangle) \geq c_0\gamma. && \text{(Lemma B.26)}
\end{aligned}
$$

By Lemma B.4 and Lemma B.28, we have

$$
\begin{cases} y_i\langle \mu, \mathrm{x}_i\rangle \geq \frac{1}{2}\|\mu\|^2, & i\in\mathcal{C} \\ |\langle \mu, \mathrm{x}_i\rangle| \leq \frac{3}{2}\|\mu\|^2, & i\in\mathcal{N} \end{cases}, \begin{cases} y_i\langle \mu, \tilde{\mathrm{x}}_i\rangle \geq \frac{1}{3}\|\mu\|^2, & i\in\mathcal{C} \\ |\langle \mu, \tilde{\mathrm{x}}_i\rangle| \leq 3\|\mu\|^2, & i\in\mathcal{N} \end{cases}
$$

And $\forall z>0, \phi'(z)\geq\gamma>0$, so applying Lemma B.25, we have the following lower bound for any W^t,

$$
y_i\left\langle \nabla f(\mathrm{x}_i;\mathrm{W}^t), \mathrm{V}\right\rangle \geq \begin{cases} \frac{\gamma}{2}\|\mu\|\cdot c_0, & i\in\mathcal{C} \\ -\frac{3}{2}\|\mu\|, & i\in\mathcal{N} \end{cases}, y_i\left\langle \nabla f(\tilde{\mathrm{x}}_i;\mathrm{W}^t), \mathrm{V}\right\rangle \geq \begin{cases} \frac{\gamma}{3}\|\mu\|\cdot c_0, & i\in\mathcal{C} \\ -3\|\mu\|, & i\in\mathcal{N} \end{cases}
$$

Similar as Lemma B.15, we have

$$
\left\langle -\nabla \widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{V}\right\rangle \geq \frac{\gamma\|\mu\|}{4} c_0 \widehat{G}_{\mathrm{rob}}(\mathrm{W}^t).
$$

Thus we have

$$
\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t) \leq \frac{4}{\gamma\|\mu\| c_0}\left\langle -\nabla \widehat{L}_{\mathrm{rob}}(\mathrm{W}^t), \mathrm{V}\right\rangle \leq \frac{4}{\gamma\|\mu\| c_0}\left\|\nabla \widehat{L}_{\mathrm{rob}}(\mathrm{W}^t)\right\|_F.
$$

We next give an upper bound on $\|\mathrm{W}^t\|_F^2$ as follows:

$$
\begin{aligned}
&\|\mathrm{W}^{t+1}\|_F^2 && (16) \\
&= \left\|\mathrm{W}^t - \eta\nabla\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t)\right\|_F^2 \\
&= \|\mathrm{W}^t\|_F^2 + \eta^2\left\|\nabla\widehat{L}_{\mathrm{rob}}(\mathrm{W}^t)\right\|_F^2 - 2\eta\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))y_i\left\langle \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t), \mathrm{W}^t\right\rangle \\
&\leq \|\mathrm{W}^t\|_F^2 + \eta^2\frac{C_2^2 d}{n}\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)^2 - 2\eta\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))y_i\left\langle \nabla f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t), \mathrm{W}^t\right\rangle && \text{(Equation (5))} \\
&= \|\mathrm{W}^t\|_F^2 + \eta^2\frac{C_2^2 d}{n}\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)^2 + 2\eta\frac{1}{n}\sum_{i=1}^{n}g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))y_i\sum_{s=1}^{m}\frac{a_s}{\sqrt{m}}\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle)\left\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\right\rangle \\
&= \|\mathrm{W}^t\|_F^2 + \eta^2\frac{C_2^2 d}{n}\widehat{G}_{\mathrm{rob}}(\mathrm{W}^t)^2 + 2\eta\frac{1}{n}\sum_{i=1}^{n}g(y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t))y_i f(\tilde{\mathrm{x}}_i^t;\mathrm{W}^t) \\
&\leq \|\mathrm{W}^t\|_F^2 + \eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta. && (g(z)z\leq\tfrac{1}{3}, g(z)\leq 1)
\end{aligned}
$$

Telescoping gives us that

$$
\|\mathrm{W}^t\|_F^2 \leq \|\mathrm{W}^0\|_F^2 + \left(\eta^2\frac{C_2^2 d}{n} + \frac{2}{3}\eta\right)t.
$$

We apply the same argument as B.15. Recall from Lemma B.26 (E2) that $\forall t\geq 0, y_k f(\tilde{\mathrm{x}}_k;\mathrm{W}^{t+1})\geq y_k f(\tilde{\mathrm{x}}_k;\mathrm{W}^t), \forall\tilde{\mathrm{x}}_k\in\mathcal{B}_2(\mathrm{x}_k,\alpha), \forall k\in[n]$. Then we have $y_k f(\tilde{\mathrm{x}}_k^T;\mathrm{W}^T)\geq y_k f(\tilde{\mathrm{x}}_k^T;\mathrm{W}^t)$, and therefore $\ell(y_k f(\tilde{\mathrm{x}}_k^T;\mathrm{W}^T))\leq\ell(y_k f(\tilde{\mathrm{x}}_k^T;\mathrm{W}^t))\leq\ell(y_k f(\tilde{\mathrm{x}}_k^t;\mathrm{W}^t))$ by definition that $\tilde{\mathrm{x}}_i^t = \arg\max_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}\ell\left(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\right), t\leq T$. As a result, we have

$$
\widehat{L}_{\mathrm{rob}}(\mathrm{W}^T) = \frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}\ell\left(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^T)\right)
$$

$$
\begin{aligned}
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}\ell\left(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\right)\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{n}\sum_{i=1}^{n}\max_{\tilde{\mathrm{x}}_i\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}-\ell'\left(y_i f(\tilde{\mathrm{x}}_i;\mathrm{W}^t)\right) && (\ell(z)\leq -2\ell'(z) \text{ when } z\geq 0, \text{ Lemma B.26 (E2)})\\
&= \frac{2}{T}\sum_{t=0}^{T-1}\widehat{G}_{\text{rob}}(\mathrm{W}^t)\\
&\leq \frac{8}{\|\mu\|\gamma c_0 T}\sum_{t=0}^{T-1}\left\langle -\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^t),\mathrm{V}\right\rangle\\
&= \frac{8}{\|\mu\|\gamma c_0\eta T}\left(\left\langle \mathrm{W}^T,\mathrm{V}\right\rangle-\left\langle \mathrm{W}^0,\mathrm{V}\right\rangle\right) && (\text{Equation (15)})\\
&\leq \frac{8}{\|\mu\|\gamma c_0\eta T}\left(\left\|\mathrm{W}^T\right\|_F+\left\|\mathrm{W}^0\right\|_F\right)\\
&\leq \frac{8}{\|\mu\|\gamma c_0\eta T}\left(2\left\|\mathrm{W}^0\right\|_F+\sqrt{\left(\eta^2\frac{C_2^2 d}{n}+\frac{2}{3}\eta\right)T}\right)\\
&\leq \frac{8}{\|\mu\|\gamma c_0\eta T}\left(\omega_{\text{init}}\sqrt{6md}+\sqrt{\left(\frac{C_2^2 d}{nCd^2}+\frac{2}{3}\right)\eta T}\right) && (\text{Lemma B.3, Assumption (4)})\\
&\leq \frac{8}{\|\mu\|\gamma c_0\eta T}\left(\sqrt{6}\eta+\sqrt{\left(\frac{1}{nd}+\frac{2}{3}\right)\eta T}\right). && (\text{Choose } C\geq C_2^2)\\
&\leq \frac{30}{\|\mu\|\gamma c_0\sqrt{\eta T}}.
\end{aligned}
$$

Thus $\forall\varepsilon>0$, $T\geq\left(\frac{30}{\|\mu\|\gamma c_0\sqrt{\eta}\varepsilon}\right)^2$ guarantees $\widehat{L}_{\text{rob}}(\mathrm{W}^T)\leq\varepsilon$.

□

Now we switch to prove generalization guarantees. Lemma B.30 provides lower bound on both the local difference of $a_s y\left(\left\langle \mathrm{w}_s^{t+1},\mathrm{x}\right\rangle-\left\langle \mathrm{w}_s^{t},\mathrm{x}\right\rangle\right)$ and the global difference of $a_s y\left(\left\langle \mathrm{w}_s^{t},\mathrm{x}\right\rangle-\left\langle \mathrm{w}_s^{0},\mathrm{x}\right\rangle\right)$, which serves a similar purpose as Lemma B.13 in networks with smooth activation functions.

Lemma B.30. Assume $(\mathrm{x},y)\sim\mathcal{D}_c$, $\mathrm{x}=y\mu+\xi$. Fix some $t>0$. On a good run, there exist constants $c_7, C'>0, C''>0$ such that the following holds for all $s\in[m]$ and for all $\tau<t$ with probability at least $1-3(d/n)^{-11}$ w.r.t. ξ,

$$
\begin{aligned}
&a_s y\left(\left\langle \mathrm{w}_s^{\tau+1},\mathrm{x}\right\rangle-\left\langle \mathrm{w}_s^{\tau},\mathrm{x}\right\rangle\right)\\
&\qquad\geq\frac{c_7\eta}{\sqrt{m}}\left(\|\mu\|^2-C'\max_{i\in[n]}|\langle\xi_i,\xi\rangle|-C'\alpha\sqrt{d\log\left(\frac{d}{n}\right)}\right)\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}
$$

Concurrently,

$$
\begin{aligned}
&a_s y\left(\left\langle \mathrm{w}_s^{t},\mathrm{x}\right\rangle-\left\langle \mathrm{w}_s^{0},\mathrm{x}\right\rangle\right)\\
&\qquad\geq\frac{c_7\eta}{\sqrt{m}}\left(\|\mu\|^2-C''\sqrt{d\log\left(dm/n\right)/n}\right)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}
$$

Proof of Lemma B.30. Consider $\mathrm{x}_i=y_i^c\mu+\xi_i$, $\mathrm{x}=y\mu+\xi$, where y_i^c and y are the clean label. Denote $\tilde{\mathrm{x}}_i^t=\arg\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x}_i,\alpha)}\ell(y_i f(\tilde{\mathrm{x}};\mathrm{W}^t))$, $\epsilon_i^t=\tilde{\mathrm{x}}_i^t-\mathrm{x}_i$. Then we have

$$
a_s y\left(\left\langle \mathrm{w}_s^{t+1},\mathrm{x}\right\rangle-\left\langle \mathrm{w}_s^{t},\mathrm{x}\right\rangle\right)
$$

$$
\begin{aligned}
&= \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} y_i y \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left\langle y_i^c \mu + \xi_i + \epsilon_i^t, y\mu + \xi \right\rangle \\
&= \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} y_i y \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left(y_i^c y \left\|\mu\right\|^2 + y \left\langle \mu, \xi_i + \epsilon_i^t \right\rangle + y_i^c \left\langle \mu, \xi \right\rangle + \left\langle \epsilon_i^t + \xi_i, \xi \right\rangle \right) \\
&= \underbrace{\frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left(\left\|\mu\right\|^2 + y_i^c \left\langle \mu, \xi_i + \epsilon_i^t \right\rangle \right)}_{B_1(t)} \\
&\quad \underbrace{- \frac{2\eta}{n\sqrt{m}} \sum_{i \in \mathcal{N}} \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left(\left\|\mu\right\|^2 + y_i^c \left\langle \mu, \xi_i + \epsilon_i^t \right\rangle \right)}_{B_2(t)} \\
&\quad \underbrace{+ \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} y_i y_i^c y \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left\langle \mu, \xi \right\rangle}_{B_3(t)} + \underbrace{\frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} y_i y \tilde{g}_i(\mathrm{W}^t)\phi'(\langle \mathrm{w}_s^t, \tilde{\mathrm{x}}_i^t\rangle) \left\langle \xi_i + \epsilon_i^t, \xi \right\rangle}_{B_4(t)}.
\end{aligned}
$$

We now bound each of the term separately. Recall from Lemma B.4 and Lemma B.28 that $|\langle \mu, \xi_i\rangle| \leq \frac{\|\mu\|^2}{2}$ and $|\langle \mu, \epsilon_i^t\rangle| \leq \frac{\|\mu\|^2}{6}$, also lemma B.25 gives us that $|\mathcal{A}_s(t)| \geq |\mathcal{A}_s(0) \cap \mathcal{T}_s| \geq c_0 n$, together with Lemma B.26 gives us that $\sum_{i \in A_s(t)} g_i(\mathrm{W}^t) \geq \frac{n c_0 \widehat{G}_{\text{rob}}(\mathrm{W}^t)}{C_r}$. Therefore, for $B_1(t)$, we have

$$
\begin{aligned}
B_1(t) &\geq \frac{\eta\gamma}{n\sqrt{m}} \cdot \frac{\left\|\mu\right\|^2}{3} \sum_{i \in \mathcal{A}_s(t)} \tilde{g}_i(\mathrm{W}^t) && \text{(Lemma B.26)} \\
&\geq \frac{\eta c_0 \gamma}{C_r\sqrt{m}} \cdot \frac{\left\|\mu\right\|^2}{3} \widehat{G}_{\text{rob}}(\mathrm{W}^t)
\end{aligned}
$$

For $B_2(t)$, we have

$$
\begin{aligned}
|B_2(t)| &\leq \frac{\eta C_r}{\sqrt{m}} \left(\frac{1}{C} + \sqrt{\frac{2}{C}} \right) \left(3\left\|\mu\right\|^2 + 2\left\|\mu\right\| \alpha \right) \widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
&\leq \frac{\eta c_0 \gamma}{4C_r\sqrt{m}} \cdot \frac{\left\|\mu\right\|^2}{3} \widehat{G}_{\text{rob}}(\mathrm{W}^t),
\end{aligned}
$$

where the last inequality holds for large enough C and $\alpha \leq \|\mu\|$.

For $B_3(t)$, define event $\mathcal{E}_1 = \left\{ |\langle \mu, \xi\rangle| \leq c_6 \sqrt{11 \log(d/n)} \left\|\mu\right\| \right\}$. Apply Proposition B.23 (D1) gives us that $\mathrm{P}(\mathcal{E}_1) \geq 1 - (d/n)^{-11}$. Therefore conditioning on $\mathcal{E}_1$ gives us that,

$$
\begin{aligned}
|B_3(t)| &\leq \frac{\eta C_r}{\sqrt{m}} \widehat{G}_{\text{rob}}(\mathrm{W}^t) \left|\langle \mu, \xi\rangle\right| \\
&\leq \frac{c_6 \eta C_r}{\sqrt{m}} \sqrt{11 \log(d/n)} \left\|\mu\right\| \widehat{G}_{\text{rob}}(\mathrm{W}^t) \\
&\leq \frac{c_6 \eta C_r}{\sqrt{m}} \sqrt{11 \log\left(\left\|\mu\right\|^4 / C^4 \right)} \left\|\mu\right\| \widehat{G}_{\text{rob}}(\mathrm{W}^t) && \text{(Assumption (B2))} \\
&\leq \frac{\eta c_0 \gamma}{4C_r\sqrt{m}} \cdot \frac{\left\|\mu\right\|^2}{3} \widehat{G}_{\text{rob}}(\mathrm{W}^t). && \text{(Choose large enough } C\text{)}
\end{aligned}
$$

For $B_4(t)$, we have

$$
|B_4(t)| \leq \frac{\eta}{\sqrt{m}} \max_{i \in [n]} \tilde{g}_i(\mathrm{W}^t) \max_{i \in [n]} \left| \left\langle \xi_i + \epsilon_i^t, \xi \right\rangle \right| \leq \frac{\eta C_r}{\sqrt{m}} \widehat{G}_{\text{rob}}(\mathrm{W}^t) \left(\max_{i \in [n]} |\langle \xi_i, \xi\rangle| + \alpha \left\|\xi\right\| \right).
$$

Proposition B.23 (D2) gives us that with probability at least $1-(d/n)^{-11}$ w.r.t. ξ, $\|\xi\| \le 10\sqrt{d\log\left(\frac{d}{n}\right)}$. Conditioning on this event gives us that

$$|B_4(t)| \le \frac{\eta C_r}{\sqrt{m}} \widehat{G}_{\text{rob}}(\mathrm{W}^t)\left(\max_{i\in[n]}|\langle \xi_i, \xi\rangle| + 10\alpha\sqrt{d\log\left(\frac{d}{n}\right)}\right).$$

Combining $B_1(t), B_2(t), B_3(t), B_4(t)$ gives us the following:

$$\begin{aligned}
&a_s y\left(\left\langle \mathrm{w}_s^{t+1}, \mathrm{x}\right\rangle - \left\langle \mathrm{w}_s^{t}, \mathrm{x}\right\rangle\right) \\
&\ge \frac{c_7\eta}{\sqrt{m}}\left(\|\mu\|^2 - C'\max_{i\in[n]}|\langle \xi_i, \xi\rangle| - C'\alpha\sqrt{d\log\left(\frac{d}{n}\right)}\right)\widehat{G}_{\text{rob}}(\mathrm{W}^t).
\end{aligned}$$

For $\langle \mathrm{w}_s^t, \mathrm{x}\rangle - \left\langle \mathrm{w}_s^0, \mathrm{x}\right\rangle$, we need to consider the cumulative of the four terms. For $B_4(t)$ in specific, we have

$$\begin{aligned}
\left\|\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y \tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)\xi_i\right\|^2 &\le 4d\sum_{i=1}^n\left(\sum_{\tau=0}^{t-1}\tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)\right)^2 \\
&\le 4C_r^2 dn\left(\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\right)^2,
\end{aligned}$$

where the first inequality comes from applying Lemma B.4 C3, and the second inequality uses the fact that $\phi'(z) \le 1$ together with $\tilde{g}_i(\mathrm{W}^\tau) \le C_r\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)$. Thus,

$$\begin{aligned}
\left\|\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y \tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)(\xi_i+\epsilon_i^\tau)\right\| &\le 2C_r\sqrt{dn}\left(\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\right) + \alpha n\left(\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\right) \\
&\le 3C_r\sqrt{dn}\left(\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\right). \\
&\qquad (\alpha \le \|\mu\|,\ d \ge Cn\|\mu\|^2 \text{ from Assumption (1) and } C \text{ large enough})
\end{aligned}$$

Therefore,

$$\left\langle\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y \tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)(\xi_i+\epsilon_i^\tau), \xi\right\rangle = \left\|\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y \tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)(\xi_i+\epsilon_i^\tau)\right\|\left\langle \psi_s^t, \xi\right\rangle,$$

where for simplicity we define $\psi_s^t = \frac{\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y\tilde{g}_i(\mathrm{W}^\tau)\phi'(\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\rangle)(\xi_i+\epsilon_i^\tau)}{\left\|\sum_{i=1}^n\sum_{\tau=0}^{t-1} y_i y\tilde{g}_i(\mathrm{W}^\tau)\phi'\left(\left\langle \mathrm{w}_s^\tau, \tilde{\mathrm{x}}_i^\tau\right\rangle\right)(\xi_i+\epsilon_i^\tau)\right\|} \in \mathbb{R}^d$, which is independent of ξ. This gives us that

$$\left|\sum_{\tau=0}^{t-1} B_4(\tau)\right| \le \frac{3C_r\eta\sqrt{d}}{\sqrt{mn}}\left(\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\right)\max_{s\in[m]}\left|\left\langle \psi_s^t, \xi\right\rangle\right|.$$

And therefore we have

$$\begin{aligned}
&a_s y\left(\left\langle \mathrm{w}_s^{t}, \mathrm{x}\right\rangle - \left\langle \mathrm{w}_s^{0}, \mathrm{x}\right\rangle\right) \\
&= \sum_{\tau=0}^{t-1} B_1(\tau) + B_2(\tau) + B_3(\tau) + B_4(\tau) \\
&\ge \frac{c_7\eta}{\sqrt{m}}\left(\|\mu\|^2 - 3C'\sqrt{d/n}\max_{s\in[m]}\left|\left\langle \psi_s^t, \xi\right\rangle\right|\right)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}$$

Define another event $\mathcal{E}_2 = \left\{ \max_{s\in[m]} |\langle \psi_s^t, \xi \rangle| \le c_6 \sqrt{11 \log (dm/n)} \right\}$. Applying Proposition B.23 (D1) gives us that $\mathrm{P}(\mathcal{E}_2) \ge 1 - (d/n)^{-11}$. Conditioning on the above events, we know

$$\begin{aligned} & a_s y \left(\langle \mathrm{w}_s^t, \mathrm{x} \rangle - \langle \mathrm{w}_s^0, \mathrm{x} \rangle \right) \\ & \quad \ge \frac{c_7 \eta}{\sqrt{m}} \left(\|\mu\|^2 - C'' \sqrt{d \log (dm/n) / n} \right) \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau). \end{aligned}$$

Applying a union bound, the above holds with probability at least $1 - 3(d/n)^{-11}$. □

Corollary B.31. Assume $(\mathrm{x}, y) \sim \mathcal{D}_c$. Fix some $t > 0$. On a good run, the following holds for all $s \in [m]$ with probability at least $1 - 4(d/n)^{-11}$,

$$a_s y \langle \mathrm{w}_s^t, \mathrm{x} \rangle \ge \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau),$$

where c_7 and C'' come from Lemma B.30.

Proof of Corollary B.31. With proper C, Assumption (B2) gives us that $\|\mu\|^2 \ge 2C'' \sqrt{d \log (md/n) / n}$. Therefore, Lemma B.30 tells us that with probability at least $1 - 3(d/n)^{-11}$,

$$a_s y \left(\langle \mathrm{w}_s^t, \mathrm{x} \rangle - \langle \mathrm{w}_s^0, \mathrm{x} \rangle \right) \ge \frac{c_7 \eta}{2\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau).$$

Using Proposition B.23 as well as Lemma B.3, the following holds with probability at least $1 - (d/n)^{-11}$,

$$\begin{aligned} \left| \langle \mathrm{w}_s^0, \mathrm{x} \rangle \right| &= \left| \langle \mathrm{w}_s^0, y\mu + \xi \rangle \right| \\ &\le \|\mathrm{w}_s^0\| \|\mu\| + c_6 \sqrt{11 \log (dm/n)} \|\mathrm{w}_s^0\| && \text{(Proposition B.23 (D1))} \\ &\le \|\mathrm{w}_s^0\| \|\mu\| + c_6 \sqrt{\frac{11n \|\mu\|^4}{C^4 d}} \|\mathrm{w}_s^0\| && \text{(Assumption (B2))} \\ &\le \|\mathrm{w}_s^0\| \|\mu\| + c_6 \sqrt{\frac{11n \|\mu\|^4}{C^5 n \|\mu\|^2}} \|\mathrm{w}_s^0\| && \text{(Assumption (1))} \\ &\le 2 \|\mu\| \|\mathrm{w}_s^0\| && \text{(Choose sufficiently large } C) \\ &\le 4\omega_{\text{init}} \|\mu\| \sqrt{d} && \text{(Lemma B.19)} \\ &\le \frac{4\eta \|\mu\|}{\sqrt{m}} && \text{(Assumption (3))} \\ &= \frac{8\eta \|\mu\|}{\sqrt{m}} \widehat{G}_{\text{rob}}(\mathrm{W}^0) && (\tilde{g}_i(\mathrm{W}^0) = 0.5) \\ &\le \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau), && (17) \end{aligned}$$

where the last line holds from $\|\mu\|^2 \ge C \log (2)$ by Assumption (B2) and C being large enough.

Therefore we have

$$a_s y \langle \mathrm{w}_s^t, \mathrm{x} \rangle \ge \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau).$$

□

Lemma B.32. For any $(\mathrm{x}, y) \sim \mathcal{D}_c$ with $\mathrm{x} = y\mu + \xi$, on a good run, there exists some constant $c_8 > 0$ such that

$$\left|\langle \mathrm{w}_s^t, \mathrm{x}\rangle\right| \leq \frac{c_8\eta}{\sqrt{m}}\left(\|\mu\|^2 + \|\mu\|\,\|\xi\| + \sqrt{d/n}\,\|\xi\| + \|\mu\|\,\alpha + \|\xi\|\,\alpha\right)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau), \forall s \in [m].$$

Proof of Lemma B.32. Consider $\forall s \in [m]$,

$$\langle \mathrm{w}_s^t, \mathrm{x}\rangle = \langle \mathrm{w}_s^0, \mathrm{x}\rangle + \sum_{\tau=0}^{t-1}\left(\langle \mathrm{w}_s^{\tau+1}, \mathrm{x}\rangle - \langle \mathrm{w}_s^\tau, \mathrm{x}\rangle\right).$$

Decompose $\langle \mathrm{w}_s^{t+1}, \mathrm{x}\rangle - \langle \mathrm{w}_s^t, \mathrm{x}\rangle$ into $B_1(t), B_2(t), B_3(t), B_4(t)$ the same way as in Lemma B.30.

$$\begin{aligned}
|B_1(t)| &\leq \frac{\eta C_r}{\sqrt{m}}\left(\frac{3}{2}\|\mu\|^2 + \|\mu\|\,\alpha\right)\widehat{G}_{\text{rob}}(\mathrm{W}^t),\\
|B_2(t)| &\leq \frac{\eta C_r}{\sqrt{m}}\left(\beta + \sqrt{\frac{2}{C}}\right)\left(3\|\mu\|^2 + 2\|\mu\|\,\alpha\right)\widehat{G}_{\text{rob}}(\mathrm{W}^t),\\
|B_3(t)| &\leq \frac{\eta C_r}{\sqrt{m}}\widehat{G}_{\text{rob}}(\mathrm{W}^t)\,|\langle \mu, \xi\rangle| \leq \frac{\eta C_r}{\sqrt{m}}\|\mu\|\,\|\xi\|\,\widehat{G}_{\text{rob}}(\mathrm{W}^t),\\
\left|\sum_{\tau=0}^{t-1}B_4(\tau)\right| &\leq \frac{2C_r\eta\sqrt{d}}{\sqrt{mn}}\|\xi\|\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau) + \frac{\eta C_r}{\sqrt{m}}\alpha\,\|\xi\|\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau),\\
\left|\langle \mathrm{w}_s^0, \mathrm{x}\rangle\right| &\leq \left\|\mathrm{w}_s^0\right\|\cdot\|\mathrm{x}\| \leq \left\|\mathrm{w}_s^0\right\|(\|\mu\| + \|\xi\|) \leq \frac{2\eta}{\sqrt{m}}(\|\mu\| + \|\xi\|) && \text{(Assumption (3), Lemma B.19)}\\
&\leq \frac{2\eta}{\sqrt{C\log(2)}\sqrt{m}}(\|\mu\|^2 + \|\mu\|\,\|\xi\|). && \text{(Assumption (B2))}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\langle \mathrm{w}_s^t, \mathrm{x}\rangle\right| &\leq \left|\langle \mathrm{w}_s^0, \mathrm{x}\rangle\right| + \sum_{\tau=0}^{t-1}|B_1(\tau)| + \sum_{\tau=0}^{t-1}|B_2(\tau)| + \sum_{\tau=0}^{t-1}|B_3(\tau)| + \left|\sum_{\tau=0}^{t-1}B_4(\tau)\right|\\
&\leq \frac{c_8\eta}{\sqrt{m}}\left(\|\mu\|^2 + \|\mu\|\,\|\xi\| + \sqrt{d/n}\,\|\xi\| + \|\mu\|\,\alpha + \|\xi\|\,\alpha\right)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}$$

□

We finally demonstrate the lower bound on the normalized expected conditional margin, similar as Lemma B.14.

Lemma B.33. On a good run, there exists some constant $c_9 > 0$ such that

$$\begin{aligned}
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}[yf(\mathrm{x};\mathrm{W}^t)|y=1]}{\left\|\mathrm{W}^t\right\|_2} &\geq \frac{c_9\sqrt{n}}{16C_2\sqrt{d}}\|\mu\|^2;\\
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}[yf(\mathrm{x};\mathrm{W}^t)|y=-1]}{\left\|\mathrm{W}^t\right\|_2} &\geq \frac{c_9\sqrt{n}}{16C_2\sqrt{d}}\|\mu\|^2.
\end{aligned}$$

Proof of Lemma B.33. Consider the following

$$yf(\mathrm{x};\mathrm{W}^t) = \frac{1}{\sqrt{m}}\sum_{a_s=y}\phi(\langle \mathrm{w}_s^t, \mathrm{x}\rangle) - \frac{1}{\sqrt{m}}\sum_{a_s=-y}\phi(\langle \mathrm{w}_s^t, \mathrm{x}\rangle).$$

We first consider $y = 1$. Denote event $\mathcal{E}$ as the conclusion of Corollary B.31 holds, then $\mathrm{P}(\mathcal{E}) \geq 1 - 4\,(d/n)^{-11}$. Corollary B.31 indicates that

$$a_s y \left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle \geq \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) > 0.$$

Therefore, on event $\mathcal{E}$ we have

$$\begin{aligned}
y f(\mathrm{x}; \mathrm{W}^t) &= \frac{1}{\sqrt{m}} \sum_{s=1}^m \phi'(\left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle) a_s y \left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle \\
&\geq \frac{1}{\sqrt{m}} \sum_{s=1}^m \phi'(\left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle) \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) \\
&\geq \frac{1}{\sqrt{m}} \sum_{a_s = y} \phi'(\left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle) \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) \\
&\geq \frac{1}{\sqrt{m}} \frac{m}{2} \gamma \frac{c_7 \eta}{4\sqrt{m}} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) \\
&= \frac{\gamma c_7 \eta}{8} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau),
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c} \left[y f(\mathrm{x}; \mathrm{W}^t) \mathbb{1}\left(\mathcal{E}\right) \middle| y = 1 \right] \\
&\geq \mathrm{P}(\mathcal{E}) \frac{\gamma c_7 \eta}{8} \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) \\
&\geq \frac{c_9}{4} \eta \|\mu\|^2 \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau). \qquad (\mathrm{P}(\mathcal{E}) \geq 1 - 4(d/n)^{-11} \geq \tfrac{3}{4}; \text{choose sufficiently large } C)
\end{aligned}$$

We now consider on event $\mathcal{E}^c$. Using Lemma B.32 gives us that,

$$\begin{aligned}
&\left| \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c} \left[y f(\mathrm{x}; \mathrm{W}^t) \mathbb{1}\left(\mathcal{E}^c\right) \middle| y = 1 \right] \right| \\
&\leq \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c} \left[\frac{1}{\sqrt{m}} \sum_{s=1}^m \left| \left\langle \mathrm{w}_s^t, \mathrm{x} \right\rangle \right| \mathbb{1}\left(\mathcal{E}^c\right) | y = 1 \right] \\
&\leq c_8 \eta \left(\|\mu\|^2 + \|\mu\| \alpha \right) \mathrm{P}(\mathcal{E}^c | y = 1) \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau) + c_8 \eta \left(\|\mu\| + \sqrt{\frac{d}{n}} + \alpha \right) \mathbb{E}\left(\|\xi\| \, \mathbb{1}\left(\mathcal{E}^c\right) | y = 1 \right) \sum_{\tau=0}^{t-1} \widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}$$

Now we denote another event $\tilde{\mathcal{E}} = \left\{ \|\xi\| \leq 12\sqrt{d \log\left(\frac{d}{n}\right)} \right\}$, then

$$\mathbb{E}\left(\|\xi\| \, \mathbb{1}\left(\mathcal{E}^c\right) \mathbb{1}\left(\tilde{\mathcal{E}}\right) | y = 1 \right) \leq \mathbb{E}\left(12\sqrt{d \log\left(\frac{d}{n}\right)} \mathbb{1}\left(\mathcal{E}^c\right) | y = 1 \right) \leq 12\sqrt{d \log\left(\frac{d}{n}\right)} 4 (\frac{d}{n})^{-11} \leq 0.5 (\frac{d}{n})^{-8},$$

$$(d \geq \log(2)\, C n^2 \text{ from Assumption (1) with sufficiently large } C)$$

$$\mathbb{E}\left(\|\xi\| \, \mathbb{1}\left(\mathcal{E}^c\right) \mathbb{1}\left(\tilde{\mathcal{E}}^c\right) | y = 1 \right) \leq \mathbb{E}\left(\|\xi\| \, \mathbb{1}\left(\tilde{\mathcal{E}}^c\right) | y = 1 \right) \leq 8 (\frac{d}{n})^{-16} \sqrt{d} \cdot (2\sqrt{\log\left(\frac{d}{n}\right)}) \leq 0.5 (\frac{d}{n})^{-8}.$$

$$(\text{Proposition B.23 (D2)}; d \geq \log(2)\, C n^2 \text{ with sufficiently large } C)$$

Putting them back gives us that

$$
\begin{aligned}
&\left|\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)\mathbb{1}\left(\mathcal{E}^c\right)\left|y=1\right]\right|\\
&\leq c_8\eta\left(\|\mu\|^2+\|\mu\|\,\alpha\right)\mathrm{P}(\mathcal{E}^c|y=1)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)+c_8\eta\left(\|\mu\|+\sqrt{d/n}+\alpha\right)\mathbb{E}(\|\xi\|\,\mathbb{1}\left(\mathcal{E}^c\right)|y=1)\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\\
&\leq c_8\eta\left(\|\mu\|^2+\|\mu\|\,\alpha\right)4(d/n)^{-11}\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)+c_8\eta\left(\|\mu\|+\sqrt{d/n}+\alpha\right)(d/n)^{-8}\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\\
&\leq c_8\eta\left(\|\mu\|^2+\|\mu\|\,\alpha+\|\mu\|+\sqrt{d/n}+\alpha\right)(d/n)^{-8}\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\qquad\text{(Assumption (1); choose sufficiently large } C)\\
&\leq\frac{c_9\eta}{8}\|\mu\|^2\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).\quad(\|\mu\|^2\geq C\log(2)\text{ from Assumption (B2)};d\geq\log(2)Cn^2\text{; choose sufficiently large } C)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]\\
&=\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)\mathbb{1}\left(\mathcal{E}\right)|y=1\right]+\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)\mathbb{1}\left(\mathcal{E}^c\right)|y=1\right]\\
&\geq\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)\mathbb{1}\left(\mathcal{E}\right)|y=1\right]-\left|\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)\mathbb{1}\left(\mathcal{E}^c\right)|y=1\right]\right|\\
&\geq\frac{c_9\eta}{8}\|\mu\|^2\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau).
\end{aligned}
$$

Then similar as Lemma B.14, recall that $\left\|\mathrm{W}^0\right\|_F\leq 2\omega_{\text{init}}\sqrt{md}\leq 2\eta\leq\eta\sqrt{d/n}\widehat{G}_{\text{rob}}(\mathrm{W}^0)$. If $\left\|\mathrm{W}^t\right\|_F\leq 2\left\|\mathrm{W}^0\right\|_F$,

$$
\begin{aligned}
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2}&\geq\frac{c_9\eta}{16\left\|\mathrm{W}^0\right\|_F}\|\mu\|^2\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\\
&\geq\frac{c_9\eta}{16\eta\sqrt{d/n}\widehat{G}_{\text{rob}}(\mathrm{W}^0)}\|\mu\|^2\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\\
&\geq\frac{c_9\sqrt{n}}{16\sqrt{d}}\|\mu\|^2.\qquad\left(\textstyle\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s)\geq\widehat{G}_{\text{rob}}(\mathrm{W}^0)\right)
\end{aligned}
$$

If $\left\|\mathrm{W}^t\right\|_F>2\left\|\mathrm{W}^0\right\|_F$, by Lemma B.12, we have

$$
2\left\|\mathrm{W}^0\right\|_F\leq\left\|\mathrm{W}^t\right\|_F\leq\left\|\mathrm{W}^0\right\|_F+C_2\eta\sqrt{d/n}\sum_{s=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^s).
$$

Thus,

$$
\begin{aligned}
\frac{\mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}_c}\left[yf(\mathrm{x};\mathrm{W}^t)|y=1\right]}{\left\|\mathrm{W}^t\right\|_2}&\geq\frac{c_9}{16C_2\sqrt{d/n}\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)}\|\mu\|^2\sum_{\tau=0}^{t-1}\widehat{G}_{\text{rob}}(\mathrm{W}^\tau)\\
&=\frac{c_9\sqrt{n}}{16C_2\sqrt{d}}\|\mu\|^2.
\end{aligned}
$$

All the above proof holds for expected condition on $y=-1$.

□

B.3. Missing Proofs in Section 3.4

Theorem 3.2. We consider independent label flip with probability β. Let $p(\mathrm{x})$ be the density function of $\mathcal{D}_{\text{clust}}$. For any given classifier $f(\cdot;\mathrm{W})$, when $\alpha < \|\mu\|$, we have $L_{\text{rob}}^{0/1}(\mathrm{W}) \geq \beta + \frac{1-2\beta}{4}\int_{\mathbb{R}^d} \min\{p(\xi), p(\xi+\mathrm{v})\}d\xi$, where $\mathrm{v} = 2\left(1 - \alpha/\|\mu\|\right)\mu$. When $\alpha \geq \|\mu\|$, the robust test error satisfies $L_{\text{rob}}^{0/1}(\mathrm{W}) \geq 0.5$.

Proof of Theorem 3.2.

$$\begin{aligned}
L_{\text{rob}}^{0/1}(\mathrm{W}) &= \mathrm{P}_{(\mathrm{x},y)\sim\mathcal{D}}\left[\exists\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } yf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right] \\
&= (1-\beta)\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\exists\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y_cf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right] \\
&\quad\quad + \beta\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\exists\tilde{\mathrm{x}} \in \mathcal{B}_2(\mathrm{x},\alpha) \text{ s.t. } y_cf(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right] \\
&= (1-\beta)\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_cf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right] \\
&\quad\quad + \beta\left(1 - \mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_cf(\tilde{\mathrm{x}};\mathrm{W}) < 0\right]\right) \\
&\geq \beta + (1-2\beta)\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left[\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_cf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right].
\end{aligned}$$

Recall that $\mathrm{x} = y_c\mu + \xi$.

Case 1: Consider the case that $\alpha \geq \|\mu\|$. We have

$$\begin{aligned}
&\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) + \mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right) \\
&\geq \mathbb{1}\left(f(\xi;\mathrm{W}) \leq 0\right) + \mathbb{1}\left(f(\xi;\mathrm{W}) \geq 0\right) \\
&\geq 1,
\end{aligned}$$

where the first inequality holds because $\xi \in \mathcal{B}_2(\mu+\xi,\alpha)$ and $\xi \in \mathcal{B}_2(-\mu+\xi,\alpha)$. Therefore we have

$$\begin{aligned}
&\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_cf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) \\
&= \frac{1}{2}\mathrm{P}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) + \frac{1}{2}\mathrm{P}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right) \\
&= 0.5\mathbb{E}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) + \mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right)\right) \\
&\geq 0.5.
\end{aligned}$$

As a result,

$$L_{\text{rob}}^{0/1}(\mathrm{W}) \geq \beta + 0.5(1-2\beta) = 0.5.$$

Case 2: Consider the case that $\alpha < \|\mu\|$. Let $p(\mathrm{x})$ denote the density function of $\mathcal{D}_{\text{clust}}$. Define $\mathrm{v} = (2 - 2\frac{\alpha}{\|\mu\|})\mu$. We have

$$\begin{aligned}
&\left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) + \mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right)\right) \\
&\quad + \left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\xi+\mathrm{v}+\mu,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right) + \mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\xi+\mathrm{v}-\mu,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W}) \geq 0\right)\right) \\
&\geq \left(\mathbb{1}\left(f(\xi+\frac{\mathrm{v}}{2};\mathrm{W}) \leq 0\right) + 0\right) + \left(0 + \mathbb{1}\left(f(\xi+\frac{\mathrm{v}}{2};\mathrm{W}) \geq 0\right)\right) \\
&\geq 1,
\end{aligned}$$

where the first inequality holds because $\xi + \frac{\mathrm{v}}{2} \in \mathcal{B}_2(\mu+\xi,\alpha)$ and $\xi + \frac{\mathrm{v}}{2} \in \mathcal{B}_2(\xi+\mathrm{v}-\mu,\alpha)$. Therefore we have

$$\mathrm{P}_{(\mathrm{x},y_c)\sim\mathcal{D}_c}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mathrm{x},\alpha)} y_cf(\tilde{\mathrm{x}};\mathrm{W}) \leq 0\right)$$

$$
\begin{aligned}
&= \left(\frac{1}{2}\mathrm{P}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\le 0\right)+\frac{1}{2}\mathrm{P}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\ge 0\right)\right)\\
&= \frac{1}{2}\mathbb{E}_{\xi\sim\mathcal{D}_{\text{clust}}}\left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\le 0\right)+\mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\ge 0\right)\right)\\
&= \frac{1}{4}\int_{\mathbb{R}^d}\left\{\left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\le 0\right)+\mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(-\mu+\xi,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\ge 0\right)\right)p(\xi)\right.\\
&\qquad\left.+\left(\mathbb{1}\left(\min_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\xi+\mathrm{v}+\mu,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\le 0\right)+\mathbb{1}\left(\max_{\tilde{\mathrm{x}}\in\mathcal{B}_2(\xi+\mathrm{v}-\mu,\alpha)} f(\tilde{\mathrm{x}};\mathrm{W})\ge 0\right)\right)p(\xi+\mathrm{v})\right\}d\xi\\
&\ge \frac{1}{4}\int_{\mathbb{R}^d}\min\{p(\xi),p(\xi+\mathrm{v})\}d\xi.
\end{aligned}
$$

As a result,

$$L^{0/1}_{\text{rob}}(\mathrm{W})\ge\beta+\frac{1-2\beta}{4}\int_{\mathbb{R}^d}\min\{p(\xi),p(\xi+\mathrm{v})\}d\xi.$$

Consider a special instance where $\mathcal{D}_{\text{clust}}$ is a standard Gaussian distribution; i.e., $\mathcal{N}(0,\mathrm{I}_d)$. Then the result can be simplify as

$$L^{0/1}_{\text{rob}}(\mathrm{W})\ge\beta+\frac{1-2\beta}{2}\Phi(-(\|\mu\|-\alpha)),$$

where $\Phi(x):=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-t^2/2\right)dt$ is the normal cumulative distribution function. □

The following result shows that for certain step sizes and initialization, the neural network weights move far from the initialization after the first step of adversarial training based on gradient descent.

Proposition B.34. Consider the same setting as in Theorem 3.1. Then, for some constant $C>1$ defined in Assumption 1, with probability at least $1-2\delta$ over the random initialization and the draw of an i.i.d. sample, we have that $\frac{\|\mathrm{W}^1-\mathrm{W}^0\|_F}{\|\mathrm{W}^0\|_F}\ge\frac{\gamma\|\mu\|}{10}$.

Proof of Proposition B.34. Consider $\mathrm{V}\in\mathbb{R}^{m\times d}$ be the matrix with rows $\mathrm{v}_s=\frac{a_s\mu}{\|\mu\|\sqrt{m}}$, then we have

$$
\begin{aligned}
\frac{\left\|\mathrm{W}^1-\mathrm{W}^0\right\|_F}{\left\|\mathrm{W}^0\right\|_F}&\ge\frac{\left\langle\mathrm{W}^1-\mathrm{W}^0,\mathrm{V}\right\rangle}{\left\|\mathrm{W}^0\right\|_F}\\
&=\frac{\eta\left\langle-\nabla\widehat{L}_{\text{rob}}(\mathrm{W}^0),\mathrm{V}\right\rangle}{\left\|\mathrm{W}^0\right\|_F}\\
&\ge\frac{\gamma\|\mu\|}{4}\frac{\eta\widehat{G}_{\text{rob}}(\mathrm{W}^0)}{\left\|\mathrm{W}^0\right\|_F} && \text{(Equation (12))}\\
&\ge\frac{\gamma\|\mu\|}{4}\frac{\eta\widehat{G}_{\text{rob}}(\mathrm{W}^0)}{\sqrt{3/2md}\omega_{\text{init}}} && \text{(Lemma B.3)}\\
&\ge\frac{\gamma\|\mu\|}{5}\widehat{G}_{\text{rob}}(\mathrm{W}^0) && \text{(Assumption (3))}\\
&=\frac{\gamma\left\|\mu\right\|}{10}. && \text{(Equation (9))}
\end{aligned}
$$

□