Stability and Generalization of Adversarial Training for Shallow Neural Networks with Smooth Activation

Kaibo Zhang
Johns Hopkins University
Baltimore, MD 21218
kzhang90@jhu.edu

Yunjuan Wang
Johns Hopkins University
Baltimore, MD 21218
ywang509@jhu.edu

Raman Arora
Johns Hopkins University
Baltimore, MD 21218
arora@cs.jhu.edu

Abstract

Adversarial training has emerged as a popular approach for training models that are robust to inference-time adversarial attacks. However, our theoretical understanding of why and when it works remains limited. Prior work has offered generalization analysis of adversarial training, but they are either restricted to the Neural Tangent Kernel (NTK) regime or they make restrictive assumptions about data such as (noisy) linear separability or robust realizability. In this work, we study the stability and generalization of adversarial training for two-layer networks **without any data distribution assumptions** and **beyond the NTK regime**. Our findings suggest that for networks with *any given initialization* and *sufficiently large width*, the generalization bound can be effectively controlled via early stopping. We further improve the generalization bound by leveraging smoothing using Moreau's envelope.

1 Introduction

Despite the remarkable performance of over-parameterized deep networks in real-world applications, recent studies have revealed that they are highly vulnerable to adversarial attacks. These attacks use maliciously crafted imperceptible perturbations designed to deceive trained neural networks during inference [Szegedy et al., 2013, Biggio et al., 2013]. The lack of adversarial robustness has raised significant concerns for deploying neural network-based models in safety-critical applications. Therefore, it is crucial to design algorithms to learn robust models that can make reliable predictions on test data even in the presence of adversarial perturbations.

One principal approach to robust learning, adversarial training [Madry et al., 2018] (along with its variants [Zhang et al., 2019, Wang et al., 2020]), has proven to be an effective empirical defense mechanism against adversarial attacks. Naturally, this puts an emphasis on also developing a theoretical understanding of robust learning. To study the generalization performance of robust learning, one traditional approach is via uniform convergence [Khim and Loh, 2018, Yin et al., 2019, Awasthi et al., 2020, Mustafa et al., 2022], which provides the worst-case type uniform bounds for a given hypothesis class and are algorithm independent. Another line of work focuses on analyzing the convergence and generalization guarantees of adversarial training, yet they either focus on linear classifiers [Charles et al., 2019, Li et al., 2020, Zou et al., 2021, Chen et al., 2023], or introduce restrictive distribution assumptions such as (noisy) linear separability [Wang et al., 2024b] or robust realizability [Mianjy and Arora, 2024]. Therefore, it remains unclear whether we can derive theoretical results for adversarial training that extend beyond these simplifying assumptions.

In this work, we leverage a different machinery by analyzing adversarial training algorithm through the lens of uniform stability. Stability is a classical tool in learning theory that has been extensively studied in the literature [Bousquet and Elisseeff, 2002, Hardt et al., 2016]. Uniform argument stability measures the difference in output parameters when an algorithm is run on two training sets that

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

differ by only one sample. In the standard (non-robust) setting, Hardt et al. [2016] show a uniform stability bound of $\mathcal{O}(\frac{\eta T}{n})$ after T iterations of gradient descent with step size η on convex and smooth losses using a training dataset of size n. They further provide a uniform stability bound of $\mathcal{O}(\frac{T^q}{n})$ for smooth and non-convex losses with decaying step size $\eta = \mathcal{O}(\frac{1}{t})$, where $q \in (0, 1)$ is a constant. The choice of decaying step size is common in the non-convex setting, as maintaining a constant step size leads to an exponentially increasing bound on uniform stability.

When it comes to the robust setting, the primary challenge lies in the non-smoothness of the robust (adversarial) loss. The robust loss is generally non-smooth even if the standard counterpart is smooth [Xing et al., 2021a, Xiao et al., 2022a]. Previous work by Xing et al. [2021a] studied the convex non-smooth adversarial losses and provide an additional term of $\mathcal{O}(\eta\sqrt{T})$ compared to the convex and smooth losses. Later Xiao et al. [2022a] studied the general non-smooth adversarial losses by leveraging the approximate co-coercivity of the gradient and provide the bound with an additional term of $\mathcal{O}(\eta T\alpha)$ that grows linearly in T, where α is the size of adversarial perturbation in ℓ_p threat models. These works, while partially addressing the issue, only focus on general convex / non-convex functions. However, neural networks, which are a specific instance of non-convex functions and are widely used in practice, require further investigation.

In this work, we study the stability and generalization guarantees of variants of adversarial training algorithms. We focus on solving the binary classification problem using two-layer over-parameterized neural networks with smooth activation functions and logistic loss. Our key contributions are as follows:

1. We present a bound of $\mathcal{O}(\sqrt{\eta}T + \frac{\eta T}{n} + \sqrt{\beta\eta T})$ on the uniform argument stability of the gradient descent-based adversarial training of over-parameterized network after T iterations with step size η, where β represents the precision of generating adversarial examples at each iteration.
2. We provide robust generalization guarantees that depend on the Adversarial Regularized Empirical Risk Minimization (ARERM) Oracle. Our results hold for any given initialization and any data distribution. Specifically, if the learner is provided with a good initialization such that there exist robust networks around this initialization, then a small robust test loss is achieved via early stopping. Furthermore, our results can be extended to stochastic gradient descent-based adversarial training.
3. We leverage Moreau's envelope to construct a smooth loss that approximates robust empirical loss. We present bounds on the stability and generalization error of gradient descent with Moreau's smoothing, and demonstrate its superiority compared with gradient descent-based adversarial training algorithm.

1.1 Related Work

Stability Analysis. The notion of stability was initially introduced in Bousquet and Elisseeff [2002] to study the generalization of statistical learning problems. More recently, a fine-grained analysis has been presented by Feldman and Vondrak [2019] and Bousquet et al. [2020]. For smooth loss functions, Hardt et al. [2016] explored the stability of SGD in both convex and non-convex settings, which was later extended to convex non-smooth loss functions by Bassily et al. [2020] and the bound incorporated an additional term of $\mathcal{O}(\eta\sqrt{T})$ due to non-smoothness. Lei and Ying [2020] tackled the non-smoothness differently by assuming the gradient of the loss to be Hölder continuous. For non-convex and non-smooth loss, Lei [2023] introduced the stability of sub-gradient, as convergence to local minimizers is observed in this setting.

Robust Generalization Guarantee. The standard method of giving a generalization guarantee is through uniform convergence. These theories typically yield an upper bound of $\mathcal{O}(\frac{1}{\sqrt{n}})$ and require a large number of training samples in order to get a small generalization gap. Techniques in this category include analyzing the Rademacher complexity [Yin et al., 2019, Khim and Loh, 2018, Awasthi et al., 2020], VC dimension [Cullina et al., 2018, Montasser et al., 2020], covering number [Balda et al., 2019, Mustafa et al., 2022, Li and Telgarsky, 2023], PAC Bayesian analysis [Farnia et al., 2018, Viallard et al., 2021, Xiao et al., 2023] and margin-based analysis [Farnia et al., 2018].

Generalization Guarantee of Adversarial Training. Providing generalization guarantees for adversarial training of neural networks is challenging due to its non-convex nature. A series of

works [Charles et al., 2019, Li et al., 2020, Zou et al., 2021, Chen et al., 2023] have focused on a simpler problem – adversarial training of linear models with a convex loss wherein generating adversarial examples admits a closed-form solution. Several works bypass this challenge by considering a lazy training regime [Gao et al., 2019, Zhang et al., 2020, Li and Telgarsky, 2023] in which the landscape of the neural network can be studied near certain random initialization, and the generalization guarantee is usually obtained via uniform convergence. Unfortunately, Wang et al. [2022] proved that adversarial robustness is at odds with lazy regime. Recently, Mianjy and Arora [2024], Wang et al. [2024b] provide convergence and generalization guarantees for adversarial training of neural networks, yet they make restrictive assumptions on the data distribution such as (noisy) linear separability and robust realizability.

Another line of research investigates the generalization of adversarial training through algorithmic stability analysis. Despite the smoothness of the standard loss, the adversarial loss remains non-smooth [Liu et al., 2020, Xing et al., 2021a, Xiao et al., 2022b]. To resolve this issue, Farnia and Ozdaglar [2021] make a strong assumption that the loss is concave in input x. Xing et al. [2021a] provide adversarial training of convex and non-smooth losses, yielding an additional term of $\mathcal{O}(\sqrt{\eta^2 T})$ compared to the standard non-robust counterpart. Xiao et al. [2022a] and Wang et al. [2024a] leverage the idea of approximate smoothness and provide bounds that scale linearly with ηT and the perturbation size. Cheng et al. [2024] consider generating adversarial examples via a single step of gradient descent and demonstrate that such variant of adversarial training algorithm achieves better stability. Farnia et al. [2018] also consider specific attack algorithms – these attacks while being more practical and designed with continuity and Lipschitzness property, may differ significantly from the worst-case attack, and do not yield a good bound on the robust generalization gap.

2 Problem Setup

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 the standard O-notation ($\mathcal{O}$, Θ and Ω).

We consider a binary classification problem with a bounded input space $\mathcal{X}$ inside a Euclidean ball of radius C_x, and label space $\mathcal{Y} = \{\pm 1\}$. We assume that data are drawn according to an unknown probability distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. The learner has access to n training data drawn i.i.d. from $\mathcal{D}$; i.e., $S = \{\mathrm{z}_i = (\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^n \sim \mathcal{D}^n$. We do not make any restrictive distributional assumptions such as realizability [Mianjy and Arora, 2024] or (noisy) linearly separability [Wang et al., 2024b].

We focus on learning two-layer neural networks, parameterized by a pair of weight matrices (a, W):

$$f_{\mathrm{W}}(\mathrm{x}) = f(\mathrm{x}; \mathrm{W}) := \textstyle\sum_{s=1}^m a_s \phi(\langle \mathrm{w}_s, \mathrm{x}\rangle).$$

Here, m is a positive integer representing the number of hidden units, i.e., the width of the networks. $\phi : \mathbb{R} \to \mathbb{R}$ is a 1-Lipschitz, H-smooth activation function. Formally, $\forall z, z' \in \mathbb{R}, |\phi'(z)| \leq 1, |\phi'(z) - \phi'(z')| \leq H\,|z - z'|$. The smoothness property of activation functions is commonly assumed in algorithmic stability literature and in theory of deep learning and covers a wide range of activation functions such as smoothed ReLU and smoothed leaky ReLU [Frei et al., 2022]. The weight matrices at the top and bottom layer are denoted as $\mathrm{a} = [a_1, \ldots, a_m] \in \mathbb{R}^m$ and $\mathrm{W} = [\mathrm{w}_1, \ldots, \mathrm{w}_m] \in \mathbb{R}^{d\times m}$, respectively. The top layer weights are initialized such that $|a_i| = \frac{1}{\sqrt{m}}, \forall i \in [m]$, and are kept fixed throughout the training process. Prior works [Du et al., 2018, Arora et al., 2019, Ji and Telgarsky, 2019] often initialize a_i to be uniformly sampled from $\{\pm\frac{1}{\sqrt{m}}\}$, which can be seen as a special instance of ours. We do not make any assumption on the initialization of the bottom layer matrix, i.e., W_0 can be either a standard Gaussian [Du et al., 2018, Ji and Telgarsky, 2019], or a vanishing initialization [Ba et al., 2019, Xing et al., 2021b], or a pre-trained model.

Adversarial Attacks. We consider a general threat model where the adversary's perturbation set is defined as $\mathcal{B} : \mathcal{X} \to 2^{\mathcal{X}}$. Given an input x, $\mathcal{B}(\mathrm{x})$ represents the set of all possible perturbations of x that an adversary can choose from. This broader definition of attack includes both the standard ℓ_p threat models with perturbation size of α, i.e., $\mathcal{B}(\mathrm{x}) = \{\tilde{\mathrm{x}} : \|\tilde{\mathrm{x}} - \mathrm{x}\|_p \leq \alpha\}$, as well as a discrete set of large-norm transformations. Unlike prior works [Mianjy and Arora, 2024, Wang et al., 2024b], we do not make any assumptions on the perturbation size.

In this work, we focus on logistic loss, $\ell(z) = \ln(1 + e^{-z})$, which serves as a smooth and convex surrogate loss for the 0-1 loss. With a slight abuse of notation, for a fixed sample $\mathrm{z} = (\mathrm{x}, y)$, we define $\ell(\mathrm{z}, \mathrm{W}) := \ell(yf(\mathrm{x}; \mathrm{W}))$. The population and empirical loss w.r.t. $\ell(\cdot)$ are denoted, respectively, as

$$L(\mathrm{W}) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}\ell(yf(\mathrm{x}; \mathrm{W})), \quad \widehat{L}(\mathrm{W}; S) := \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(\mathrm{x}_i; \mathrm{W})).$$

Given $\mathcal{B}$, for a fixed sample $\mathrm{z} = (\mathrm{x}, y)$, we define the robust loss as $\ell_{rob}(\mathrm{z}, \mathrm{W}) := \max_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}; \mathrm{W}))$. The robust population and empirical loss w.r.t. $\ell(\cdot)$ are defined as

$$L_{rob}(\mathrm{W}) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}} \max_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}; \mathrm{W})) \quad \widehat{L}_{rob}(\mathrm{W}; S) := \frac{1}{n}\sum_{i=1}^{n} \max_{\tilde{\mathrm{x}}_i\in\mathcal{B}(\mathrm{x}_i)} \ell(y_i f(\tilde{\mathrm{x}}_i; \mathrm{W})).$$

Adversarial Training. During training, the network bottom layer weight W are updated using gradient descent-based adversarial training (or its stochastic version). We denote the weight matrix at the t-th iterate of adversarial training as W_t. For each training example (x_i, y_i), at iteration t, we generate a *β_1-optimal adversarial example* $(\tilde{\mathrm{x}}_i(\mathrm{W}_t), y_i)$, which satisfies the following condition:

$$\ell(y_i f(\tilde{\mathrm{x}}_i(\mathrm{W}_t); \mathrm{W}_t)) \geq \max_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x}_i)} \ell(y_i f(\tilde{\mathrm{x}}; \mathrm{W}_t)) - \beta_1. \tag{1}$$

Setting $\beta_1 = 0$ recovers the scenario where we have access to the worst-case adversarial attack. As this may not be feasible in practice due to computational reason, the parameter β_1 allows us to capture the precision of the attack algorithm, which includes common attacks such as projected gradient descent (PGD) [Madry et al., 2018]. We should regard β_1 as a parameter we can choose. Our results in Section 3 suggest that we can achieve better generalization by adding more computation and making β_1 smaller.

Algorithm 1 Variants of Adversarial Training Algorithms

Input: Step size η. Number of iterations T. Initial weight W_0. $\beta \geq 0$. $\mu > 0$.
for $t = 0, \ldots, T-1$ **do**
GD: $\forall i \in [n]$, compute a β_1-optimal adversarial example $\tilde{\mathrm{x}}_i(\mathrm{W}_t)$ that satisfies Equation (1).
Update $\mathrm{W}_{t+1} = \mathrm{W}_t - \frac{\eta}{n}\sum_{i=1}^{n} \nabla_{\mathrm{W}}\ell(y_i f(\tilde{\mathrm{x}}_i(\mathrm{W}_t); \mathrm{W}_t))$.
SGD: Compute a β_1-optimal adversarial example $\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t)$ that satisfies Equation (1).
Update $\mathrm{W}_{t+1} = \mathrm{W}_t - \eta\nabla_{\mathrm{W}}\ell(y_{t+1} f(\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t); \mathrm{W}_t))$.
Moreau Envelope: Compute a β_2-optimal minimizer $\widetilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S)$ that satisfies Equation (2).
Update $\mathrm{W}_{t+1} = \mathrm{W}_t - \frac{\eta}{\mu}(\mathrm{W}_t - \widetilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S))$.
end for
return: $\{\mathrm{W}_t\}_{t=0}^{T}$.

Optimizing the Moreau Envelope. Since the robust loss is non-smooth [Xiao et al., 2022a], we utilize Moreau's envelope to construct a smooth function that approximates the empirical robust loss. Such an idea has previously been explored in Xiao et al. [2024]. Given training data S and $\mu > 0$, we redefine the robust surrogate loss as follows:

$$M^{\mu}(\mathrm{W}; S) = \min_{\mathrm{U}} \left(\widehat{L}_{rob}(\mathrm{U}; S) + \frac{1}{2\mu}\|\mathrm{U} - \mathrm{W}\|_F^2 \right).$$

Selecting μ appropriately ensures that $\widehat{L}_{rob}(\mathrm{U}; S) + \frac{1}{2\mu}\|\mathrm{U} - \mathrm{W}\|_F^2$ is a strongly convex function w.r.t. U. Given W and S, we define $\mathrm{U}^{\mu}(\mathrm{W}; S) = \mathrm{argmin}_{\mathrm{U}\in\mathbb{R}^{d\times m}} \widehat{L}_{rob}(\mathrm{U}; S) + \frac{1}{2\mu}\|\mathrm{U} - \mathrm{W}\|_F^2$, which can be obtained via subgradient-based method (solve a min-max optimization). The gradient of the Moreau envelope can be simply calculated as $\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}; S) = \frac{1}{\mu}(\mathrm{W} - \mathrm{U}^{\mu}(\mathrm{W}; S))$. Given training data S, at each iteration t, we generate a *β_2-optimal minimizer* $\widetilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S)$ that satisfies

$$\widehat{L}_{rob}(\widetilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S); S) + \frac{1}{2\mu}\|\widetilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S) - \mathrm{W}_t\|_F^2 \leq \beta_2 + M^{\mu}(\mathrm{W}_t; S). \tag{2}$$

We remark that β_2-optimal minimizer defined in Equation (2) and β_1-optimal adversarial example defined in Equation (1) are approximating different quantities, which are not comparable. All the algorithms described above are summarized in Algorithm 1.

Uniform Argument Stability. Given a training set $S = \{\mathrm{z}_i\}_{i=1}^n$ drawn i.i.d. from $\mathcal{D}$, let S' denote the training set obtained by replacing one example in S with an independently drawn example $\mathrm{z}' \sim \mathcal{D}$. We refer to S, S' as neighboring samples and write $S \simeq S'$. Given an algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$, where the hypothesis class $\mathcal{H}$ is parameterized using a parameter matrix $\mathrm{W} \in \mathbb{R}^{d \times m}$, we define the uniform argument stability as

$$\delta_{\mathcal{A}}(S, S') := \|\mathcal{A}(S) - \mathcal{A}(S')\|_F.$$

For any L-Lipschitz loss function g, $|g(\mathcal{A}(S), \mathrm{z}) - g(\mathcal{A}(S'), \mathrm{z})| \leq L\delta_{\mathcal{A}}(S, S')$. The standard stability argument [Mohri et al., 2018] relates the expected generalization gap to the uniform argument stability.

$$\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathcal{A}(S)) := \mathbb{E}_{S\sim\mathcal{D}^n}\left(\mathbb{E}_{\mathrm{z}\sim\mathcal{D}}g(\mathcal{A}(S), \mathrm{z}) - \frac{1}{n}\sum_{i=1}^n g(\mathcal{A}(S), \mathrm{z}_i)\right) \leq L \sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S'). \quad (3)$$

In this paper, we consider robust generalization using logistic loss, so function $g(\mathrm{W}, \mathrm{z}) = \ell_{rob}(\mathrm{z}, \mathrm{W})$, and $\varepsilon_{gen}(\mathrm{W}) = \mathbb{E}_{\mathrm{z}\sim\mathcal{D}}[\ell_{rob}(\mathrm{z}, \mathrm{W})] - \frac{1}{n}\sum_{i=1}^n \ell_{rob}(\mathrm{z}_i, \mathrm{W})$. We also remark that a high probability bound for stabile algorithms can be given based on Feldman and Vondrak [2019]. For simplicity, our generalization bounds in this paper are only in expectation.

3 Main Result

In this section, we present our main results, providing theoretical guarantees for adversarial training of two-layer neural networks with smooth activation functions. We discuss (stochastic) adversarial training in Section 3.1 and gradient descent-based Moreau's smoothing in Section 3.2. Our generalization bounds rely on a key quantity, the *Adversarial Regularized Empirical Risk Minimization (ARERM) Oracle* defined as

$$\Delta_S^{\text{oracle}} := \min_{\mathrm{W}\in\mathbb{R}^{d\times m}} \left(\widehat{L}_{rob}(\mathrm{W}; S) + \frac{2\|\mathrm{W} - \mathrm{W}_0\|_F^2}{\eta T}\right).$$

Given a sample, Δ_S^{oracle} returns the minimal empirical risk in the vicinity of an initialization W_0.

3.1 Generalization Guarantees for Adversarial Training

We begin by presenting a bound on the uniform argument stability (UAS) of Algorithm 1 with GD.

Theorem 3.1. Assume that the network width satisfies $m \geq H^2 C_x^4 \eta^2 (T+1)^2$. Then, after T iterations of Algorithm 1 with GD, for any neighboring datasets S, S', we have

$$\sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \leq \mathcal{O}(C_x \eta\sqrt{T} + C_x \frac{\eta T}{n} + \sqrt{\beta_1 \eta T}).$$

Remarkably, setting $\beta_1 = 0$ yields a bound of $\mathcal{O}(\eta\sqrt{T} + \frac{\eta T}{n})$ on the UAS of Algorithm 1, thereby recovering the result in prior work of Xing et al. [2021a]. However, note that Xing et al. [2021a] show the result only for convex learning problems, whereas we consider training two-layer neural networks using logistic loss, which is non-convex and non-smooth. Further note that we assume that the networks are sufficiently over-parameterized, i.e., $m \geq \Omega(\eta^2 T^2)$, a condition that is commonly assumed in deep learning theory. We can also regard this condition as early stopping, wherein $T \leq \mathcal{O}\left(\frac{\sqrt{m}}{HC_x^2\eta}\right)$. This view is also consistent with several empirical studies [Caruana et al., 2000, Rice et al., 2020, Pang et al., 2021].

Next, we show that stable robust learning rules do not overfit.

Theorem 3.2. Define $\alpha_1(\eta, T) := \mathcal{O}(C_x^2\eta\sqrt{T} + C_x^2\frac{\eta T}{n} + C_x\sqrt{\beta_1\eta T})$. Assume that the width of the networks satisfies $m \geq H^2 C_x^4 \eta^2 (T+1)^2$, and $\alpha_1(\eta, T) < 1$. Then, after T iterations of Algorithm 1 with GD, we have

$$\min_{\lceil\frac{9T}{10}\rceil\leq t\leq T} \mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathrm{W}_t) \leq \frac{17\alpha_1(\eta, T)}{1 - \alpha_1(\eta, T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2} + \beta_1\right],$$

and

$$\min_{0\leq t\leq T} \mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \frac{1}{1 - \alpha_1(\eta, T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2} + \beta_1\right].$$

The result above bounds the robust generalization gap and the robust loss in terms of the ARERM oracle, a step size-dependent term $\mathcal{O}(\eta)$, and the precision of the adversarial examples β_1. Note though that the bound holds for the minimum over the last few iterates (past iterates), rather than for the last iteration. This distinction arises because, unlike standard gradient descent for neural networks, we cannot guarantee a decreasing robust training loss without additional assumptions on the data distributions owing to the non-smooth nature of the robust loss. The step size-dependent term arises for the same reason. A direct corollary gives us a bound on the expected robust loss.

Corollary 3.3. After $T \leq \mathcal{O}(\min\{n^2, \frac{1}{\beta_1^2}\})$ iterations of Algorithm 1 with GD using a step size of $\eta = \Theta(\frac{1}{C_x^2\sqrt{T}})$ on a network with width $m \geq \Omega(T)$, for any weight matrix W

$$\min_{0\leq t\leq T} \mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq 1.1 L_{rob}(\mathrm{W}) + \mathcal{O}\left(\frac{C_x^2\|\mathrm{W}-\mathrm{W}_0\|_F^2}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Since corollary 3.3 holds for any W_0, it underscores the importance of initialization for robust learning. Given a good initialization, such as a pre-trained model, and assuming that there exists a robust network W_* in the vicinity of the initialization (i.e., $\|\mathrm{W}_* - \mathrm{W}_0\|_F = \mathcal{O}(1)$) that achieves a small robust loss $L_{\text{rob}}(\mathrm{W}_*) \approx 0$, we have that the minimum expected robust loss over all iterates approaches $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. Further, if β_1 is small enough and $m \gtrsim n^2$, then T can be of the order $\Theta(n^2)$, leading to a $\mathcal{O}(1/n)$ upper bound on the robust test loss.

We remark that by a similar analysis, our result can be reduced to the standard (non-robust) setting for gradient descent training of two-layer networks by setting the perturbation set $\mathcal{B}(\mathrm{x}) = \{\mathrm{x}\}, \forall \mathrm{x} \in \mathcal{X}$, $\beta_1 = 0$, and redefining $\alpha_1(\eta, T) = \mathcal{O}(C_x^2\frac{\eta T}{n})$. In this context, we can show that gradient descent for the binary classification problem can achieve excess risk bound of $\mathcal{O}(1/\sqrt{n})$ by taking $\eta T = \Theta(\sqrt{n})$ if $m \gtrsim n$ and assuming $\|\mathrm{W}_* - \mathrm{W}_0\|_F = \mathcal{O}(1)$, where $\mathrm{W}_* \in \operatorname{argmin}_{\mathrm{W}} L_{rob}(\mathrm{W})$.

Next, we extend our result to the stochastic adversarial training.

Theorem 3.4. After T iterations of Algorithm 1 with SGD on a network of width $m \geq H^2C_x^4\eta^2(T+1)^2$ we have that for any weight matrix W,

$$\min_{0\leq t\leq T} \mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} L_{rob}(\mathrm{W}_t) \leq L_{rob}(\mathrm{W}) + \frac{\|\mathrm{W}-\mathrm{W}_0\|_F^2}{\eta(T+1)} + \frac{C_x^2\eta}{2} + \beta_1.$$

Similar to the discussion following Corollary 3.3, we assert that if we assume that there exists an over-parameterized robust network with small robust loss, then using a step size of $\eta = 1/\sqrt{T}$, stochastic adversarial training yields an excess risk bound of $\mathcal{O}(1/\sqrt{T})$.

3.2 Generalization Guarantees for Gradient Descent on Moreau's Envelope

We now present a bound on the uniform argument stability of gradient descent with smoothing based on Moreau's envelope.

Theorem 3.5. After T iterations of Algorithm 1 with Moreau Envelope with step-size $\eta \leq \min\{\mu, \frac{\sqrt{m}}{8HC_x^2}\} \leq \frac{\sqrt{m}}{2HC_x^2}$, on a network of width $m \geq H^2C_x^4\eta^2T^2$, for any neighboring datasets S, S', we have

$$\sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \leq \mathcal{O}\left(C_x\frac{\eta T}{n} + \eta T\sqrt{\frac{\beta_2}{\mu}}\right).$$

Setting $\beta_2 = 0$ yields a bound of $\mathcal{O}(\frac{\eta T}{n})$ on the UAS of Algorithm 1, thereby recovering the result in prior work of [Hardt et al., 2016, Xiao et al., 2024] for convex and smooth functions. Note that by using Moreau's envolope, we are able to shave off the $\mathcal{O}(\eta\sqrt{T})$ term that appears in Theorem 3.1.

Although inspired by Xiao et al. [2024], Theorem 3.5 differs from the non-convex setting of Xiao et al. [2024]. Our result utilizes the specific structure of over-parameterized neural networks that exhibit weakly convex properties, a special instance of non-convex functions, and allows for a constant step size. In contrast, [Xiao et al., 2024, Theorem 4.7] follows the traditional stability argument for non-convex and smooth functions in Hardt et al. [2016], considering a decaying step size $\eta_t \leq \frac{\mu}{t}$.

Such a condition might be impractical if μ is chosen to be sufficiently small. In fact, our results indicate that it is necessary to select a sufficiently small μ so that the robust training loss is well approximated by the Moreau envelope (see Lemma C.1 in the Appendix).

Even though the gradient descent-based algorithm with Moreau's smoothing achieves better stability guarantees compared to gradient descent-based adversarial training when $\beta_1 = \beta_2 = 0$, it requires more computational resources. Specifically, for the calculation of the gradient at each step, we need to solve a min-max optimization problem with a strongly convex and non-smooth objective to obtain a β-optimal minimizer. Additionally, for every step of this min-max optimization, we need to generate adversarial examples and apply sub-gradient descent.

Theorem 3.6. Define $\alpha_2(\eta, T) := \mathcal{O}(C_x^2 \frac{\eta T}{n} + C_x \eta T \sqrt{\frac{\beta_2}{\mu}})$. Assume $\alpha_2(\eta, T) < 1$. Then, after $T \geq 8$ iterations of Algorithm Moreau Envelope with step-size $\eta \leq \mu$ on a network of width $m \geq H^2 C_x^4 \eta^2 T^2$, we have

$$\min_{[\frac{9T}{10}] \leq t \leq T} \mathbb{E}_{S \sim \mathcal{D}^n} \varepsilon_{gen}(\mathrm{W}_t) \leq \frac{55\alpha_2(\eta, T)}{1 - \alpha_2(\eta, T)} \left[\mathbb{E}_{S \sim \mathcal{D}^n} \Delta_S^{\text{oracle}} + C_x^2 \mu + 2\eta(T+1)\frac{\beta_2}{\mu} \right],$$

and

$$\min_{1 \leq t \leq T} \mathbb{E}_{S \sim \mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \frac{1}{1 - \alpha_2(\eta, T)} \left[\mathbb{E}_{S \sim \mathcal{D}^n} \Delta_S^{\text{oracle}} + C_x^2 \mu + 2\eta(T+1)\frac{\beta_2}{\mu} \right].$$

Similar to Theorem 3.2, the result above shows that both the robust generalization gap as well as the robust loss can be bounded in terms of the ARERM oracle, parameter μ in Moreau's envelope, and a term of $\mathcal{O}(\eta T \beta_2 / \mu)$ dependent on the precision of generating the minimizer of Moreau envelope. While the bound above is on the minimum expected generalization gap (and expected robust test loss) over the last few iterates (past iterates), we can give a bound for the the last iterate for the case when $\beta_2 = 0$. We conclude the section by presenting the following direct corollary.

Corollary 3.7. After $T \leq \mathcal{O}(\min\{n^2, \frac{1}{\beta_2^{2/3}}\})$ iterations of Algorithm 1 with Moreau Envelope with step-size $\eta = \mu = \Theta(\frac{1}{C_x^2 \sqrt{T}})$ on a network of width $m \geq \Omega(T)$, we have for any weight matrix W,

$$\min_{1 \leq t \leq T} \mathbb{E}_{S \sim \mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq 1.1 L_{rob}(\mathrm{W}) + \mathcal{O}\left(\frac{C_x^2 \|\mathrm{W} - \mathrm{W}_0\|_F^2}{\sqrt{T}} \right) + \mathcal{O}\left(\frac{1}{\sqrt{T}} \right).$$

4 Proof Sketch

We begin by providing a high level intuition behind our analysis technique, and then we highlight the key ideas in the proofs of the main theorems. For simplicity, we assume that the learner can generate optimal attacks during adversarial training, i.e., we consider $\beta_1 = 0$, $\beta_2 = 0$ in this section. We refer the reader to the Appendix for proofs of the more general case.

Our analysis relies on a key lemma demonstrating that the objective function (i.e., the robust empirical risk) being minimized in adversarial training of two-layer neural networks with smooth activation functions using the logistic loss function is "almost" convex.

Definition 4.1. Let $l > 0$. A function $f(x)$ is said to be $-l$-weakly convex if $f(x) + \frac{l}{2}\|x\|_2^2$ is convex in x.

Lemma 4.2. (Restatement of Lemma A.4) For any weight matrices W^1 and W^2,

$$\widehat{L}_{rob}(\mathrm{W}^2; S) \geq \widehat{L}_{rob}(\mathrm{W}^1; S) + \left\langle \nabla_{\mathrm{W}} \widehat{L}_{rob}(\mathrm{W}^1; S), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle - \frac{H C_x^2}{2\sqrt{m}} \|\mathrm{W}^2 - \mathrm{W}^1\|_F^2.$$

Equivalently, $\widehat{L}_{rob}(\mathrm{W}; S)$ is $-\frac{H C_x^2}{\sqrt{m}}$-weakly convex.

We borrow many ideas from Xiao et al. [2024] and Xing et al. [2021a] in our proofs. These papers primarily focus on the convex setting, while only giving a general result for non-convex functions. We extend their results to a special case of learning neural networks. We argue that by specializing our analysis to neural networks would lead to sharper results than a general non-convex function class, as we will be able to leverage the "almost" convexity of neural network training [Richards

and Rabbat, 2021, Richards and Kuzborskij, 2021]. This allows us to get stability and optimization guarantees that are similar to the convex setting when we consider an over-parameterized network $m \geq \text{poly}(\eta T)$. An additional challenge we face is that the robust loss is non-smooth even if its standard counterpart (logistic loss) is smooth, making the analysis more complicated than the standard (non-robust) scenario. Nevertheless, we can still leverage the "almost" convex nature of the loss to establish the stability of adversarial training.

The following lemma gives a relationship between stability and generalization which is useful in both standard adversarial training as well as gradient descent with Moreau's envelope. When the robust training loss $\widehat{L}_{rob}(\mathrm{W}_T; S)$ is small, Lemma 4.3 provides a tighter bound than directly applying Equation (3). See Proposition A.3 for both results.

Lemma 4.3. (Restatement of Proposition A.3) The robust test loss satisfies the following:

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_T) \leq \mathbb{E}_{S\sim\mathcal{D}^n} \frac{1}{1 - C_x \cdot \sup\limits_{S\simeq S'} \delta_{\mathcal{A}}(S, S')} \widehat{L}_{rob}(\mathrm{W}_T; S).$$

This result gives a way to bound the expected robust loss. Say you want to bound the expected robust test loss by $(1+\epsilon)$ times the expected training loss. Then, to ensure $\frac{1}{1-\alpha_1(\eta,T)} \leq 1+\epsilon$, we need $\alpha_1(\eta, T) \leq \frac{\epsilon}{1+\epsilon} = O(\epsilon)$.Since $\alpha_1(\eta, T) = O(\eta\sqrt{T} + \frac{\eta T}{n} + \sqrt{\beta_1 \eta T})$, we can set different parameters in more than one way to ensure that $\alpha_1(\eta, T) = O(\epsilon)$. We can set $\beta_1 = O(\epsilon^2)$, $n = \Theta(1/\epsilon)$, $T = \Theta(1/\epsilon^2)$, $\eta = O(\frac{1}{T})$; or set $\beta_1 = O(\epsilon^3)$, $n = \Theta(1/\epsilon^2)$, $T = \Theta(1/\epsilon^4)$, $\eta = O(\frac{\epsilon}{\sqrt{T}})$.

4.1 Generalization Guarantees for Gradient-Based Adversarial Training

The stability guarantee we give in the following Theorem 4.4 is similar to the result in the convex case [Xing et al., 2021a]. While [Xing et al., 2021a] use the monotone subgradient condition of the convex functions, we show that the subgradients of an "almost" convex loss function are "almost" monotone. We do incur an additional term of $\exp\left(2HC_x^2\eta T/\sqrt{m}\right)$, which is small for over-parameterized neural networks ($m \geq \text{ploy}(\eta T)$).

Theorem 4.4. (Restatement of Theorem 3.1) Let S and S' be any two neighboring data sets, i.e., they differ only in one example. Let W_T and W'_T denote the weight matrices returned after T iterations of Algorithm 1 with GD on S and S', respectively. Then, we have

$$\|\mathrm{W}_T - \mathrm{W}'_T\|_F^2 \leq \exp\left(1 + \frac{2HC_x^2\eta T}{\sqrt{m}}\right) \cdot \left(4C_x^2\eta^2(T+1) + \frac{4C_x^2\eta^2(T+1)^2}{n^2}\right).$$

We next provide an intermediate lemma that lead us to Theorem 3.2.

Lemma 4.5. (Restatement of Theorem B.2) Set $k = \left(1 + \frac{HC_x^2\eta}{\sqrt{m}}\right)^{-1}$. Then after $T \leq \frac{\sqrt{m}}{HC_x^2\eta} - 1$ iterations of Algorithm 1 with GD,

$$\frac{1}{\sum\limits_{t=0}^{T} k^t} \sum_{t=0}^{T} k^t \widehat{L}_{rob}(\mathrm{W}_t; S) \leq \Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2}.$$

Richards and Kuzborskij [2021] (see Lemma 2 in their paper) give an optimization guarantee by providing an upper bound on the averaged training loss $\frac{1}{T}\sum_{t=1}^{T} \widehat{L}(\mathrm{W}_t; S)$ of all iterates. In Lemma 4.5 we use a more refined analysis by considering the weighted average of the training loss. Specifically, for any weight matrix W, we follow the standard technique in the convex case and upper bound the following:

$$\|\mathrm{W} - \mathrm{W}_{t+1}\|_F^2 = \|\mathrm{W} - \mathrm{W}_t\|_F^2 + \eta^2\|\nabla_{\mathrm{W}}\widehat{L}_{rob}(\mathrm{W}_t; S)\|_F^2 + 2\eta\left\langle \nabla_{\mathrm{W}}\widehat{L}_{rob}(\mathrm{W}_t; S), \mathrm{W} - \mathrm{W}_t \right\rangle.$$

The second term on the right hand side is bounded by the Lipschitzness of the logistic loss. The inner product in the third term is bounded by $\widehat{L}_{rob}(\mathrm{W}; S) - \widehat{L}_{rob}(\mathrm{W}_t; S) + \frac{HC_x^2}{2\sqrt{m}}\|\mathrm{W} - \mathrm{W}_t\|_F^2$ using Lemma A.4. We finish the proof by telescoping. The weighted telescoping technique removes all of the $\|\mathrm{W} - \mathrm{W}_t\|_F^2$ terms ($t > 0$) in the upper bound, thereby giving a simpler result. The term $C_x^2\eta/2$ in the upper bound stems from the non-smoothness of the robust loss, and is unavoidable even if the robust loss is convex. Finally, Theorem 3.2 follows from Theorem 4.4 and Lemmas 4.3 and 4.5.

4.2 Generalization Guarantees for Gradient-Descent on Moreau's Envelope

Below we give the key lemmas for bounding the generalization error of GD with Moreau's envelope. The proof technique here is similar to that for standard adversarial training (in the previous section), except that we get to utilize the smoothness of Moreau's envelope. Specifically, Lemma 4.6 leverages the fact that the gradient is "almost" co-coercive to control the uniform argument stability.

Theorem 4.6. (Restatement of Theorem C.4) Let $S \simeq S'$ be any two neighboring data sets, i.e., S and S' differ only in one example. For any $\eta \leq \min\{\mu, \frac{\sqrt{m}}{8HC_x^2}\} \leq \frac{\sqrt{m}}{2HC_x^2}$, let W_T and W'_T be the weight matrices obtained by T iterations of gradient descent with Moreau's envelopes on datasets S and S', respectively. Then, we have that

$$\|\mathrm{W}_T - \mathrm{W}'_T\|_F^2 \leq \exp\left(1 + \frac{8HC_x^2\eta T}{\sqrt{m}}\right) \cdot \frac{16C_x^2\eta^2(T+1)^2}{n^2}.$$

Lemma 4.7 also leverages smoothness due to Moreau's envelope and yields a bound that does not involve the additional term $C_x^2\eta/2$ compared with Lemma 4.5.

Lemma 4.7. (Restatement of Theorem C.6) Set $k = \left(1 + \frac{2HC_x^2\eta}{\sqrt{m}}\right)^{-1}$. After T iterations of Algorithm 1 with Moreau Envelope with $\eta \leq \mu \leq \frac{\sqrt{m}}{2HC_x^2}$ and $T \leq \frac{\sqrt{m}}{HC_x^2\eta}$, we have

$$\frac{1}{\sum_{t=1}^{T} k^t} \sum_{t=1}^{T} k^t M^\mu(\mathrm{W}_t; S) \leq \Delta_S^{\mathrm{oracle}}.$$

Theorem 3.6 is naturally derived via Theorem 4.6, Lemma 4.3 and 4.7.

5 Conclusion

In this work, we establish the generalization guarantees for variants of adversarial training applied to two-layer networks with smooth activation functions. For over-parameterized neural networks, we present robust generalization bound that are controlled by the Adversarial Regularized Empirical Risk Minimization (ARERM) oracle, applicable to any given initialization and any data distributions. One future direction is to extend our analysis to deep neural networks and beyond neural networks with smooth activation functions.

Acknowledgments and Disclosure of Funding

This research was supported, in part, by the DARPA GARD award HR00112020004, NSF CAREER award IIS-1943251, funding from the Institute for Assured Autonomy (IAA) at JHU, and the Spring'22 workshop on "Learning and Games" at the Simons Institute for the Theory of Computing.

References

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

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2019.

Emilio Rafael Balda, Arash Behboodi, Niklas Koep, and Rudolf Mathar. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim vSrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Rich Caruana, Steve Lawrence, and C. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Zachary Charles, Shashank Rajput, Stephen Wright, and Dimitris Papailiopoulos. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*, 2019.

Jinghui Chen, Yuan Cao, and Quanquan Gu. Benign overfitting in adversarially robust linear classification. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Xiwei Cheng, Kexin Fu, and Farzan Farnia. Stability and generalization in free adversarial training. *arXiv preprint arXiv:2404.08980*, 2024.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185. PMLR, 2021.

Farzan Farnia, Jesse M Zhang, and David Tse. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.

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

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

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

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

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 191–227. PMLR, 2023.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

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

Yan Li, Ethan Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Poorya Mianjy and Raman Arora. Robustness guarantees for adversarially trained neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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

Waleed Mustafa, Yunwen Lei, and Marius Kloft. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pages 16174–16196. PMLR, 2022.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in neural information processing systems*, 34:8609–8621, 2021.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2021.

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

Paul Viallard, Eric Guillaume VIDOT, Amaury Habrard, and Emilie Morvant. A PAC-Bayes analysis of adversarial robustness. *Advances in Neural Information Processing Systems*, 34:14421–14433, 2021.

Yihan Wang, Shuang Liu, and Xiao-Shan Gao. Data-dependent stability analysis of adversarial training. *arXiv preprint arXiv:2401.03156*, 2024a.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Yunjuan Wang, Enayat Ullah, Poorya Mianjy, and Raman Arora. Adversarial robustness is at odds with lazy training. *Advances in Neural Information Processing Systems*, 35:6505–6516, 2022.

Yunjuan Wang, Kaibo Zhang, and Raman Arora. Benign overfitting in adversarially trained neural networks. In *International Conference on Machine Learning*, 2024b.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, and Zhi-Quan Luo. Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*, 2022a.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022b.

Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. PAC-Bayesian adversarially robust generalization bounds for deep neural networks. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Jiancong Xiao, Jiawei Zhang, Zhi-Quan Luo, and Asuman Ozdaglar. Uniformly stable algorithms for adversarial training and beyond. *arXiv preprint arXiv:2405.01817*, 2024.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021a.

Yue Xing, Qifan Song, and Guang Cheng. On the generalization properties of adversarial training. In *International Conference on Artificial Intelligence and Statistics*, pages 505–513. PMLR, 2021b.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Yi Zhang, Orestis Plevrakis, Simon S Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *Advances in Neural Information Processing Systems*, 33:679–688, 2020.

Difan Zou, Spencer Frei, and Quanquan Gu. Provable robustness of adversarial training for learning halfspaces with noise. In *International Conference on Machine Learning*, pages 13002–13011. PMLR, 2021.

Supplementary Material

A Technical Theorems and Lemmas

Lemma A.1. Let $\ell(z) = \ln(1 + e^{-z})$ be the logistic loss function. We have $|\ell'(z)| \leq \min\{1, \ell(z)\}$.

Proof of Lemma A.1.

$$|\ell'(z)| = -\ell'(z) = \frac{1}{1+e^z} \leq \begin{cases} 1; & (e^z > 0) \\ \ln(1+e^{-z}). & \left(\frac{x}{1+x} \leq \ln(1+x)\right) \end{cases}$$

□

Lemma A.2. For any sample $\mathrm{z} = (\mathrm{x}, y)$, and any weight matrices W and W', we have

$$\ell_{rob}(\mathrm{z}, \mathrm{W}) - \ell_{rob}(\mathrm{z}, \mathrm{W}') \leq C_x \|\mathrm{W} - \mathrm{W}'\|_2 \cdot \min\{1, \ell_{rob}(\mathrm{z}, \mathrm{W})\}.$$

Proof of Lemma A.2.

$$\begin{aligned}
&\ell_{rob}(\mathrm{z}, \mathrm{W}) - \ell_{rob}(\mathrm{z}, \mathrm{W}') \\
=&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}; \mathrm{W})) - \max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}; \mathrm{W}')) \\
\leq&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \left(\ell(yf(\tilde{\mathrm{x}}; \mathrm{W})) - \ell(yf(\tilde{\mathrm{x}}; \mathrm{W}'))\right) \\
\leq&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} |\ell'(yf(\tilde{\mathrm{x}}; \mathrm{W})) \cdot (yf(\tilde{\mathrm{x}}; \mathrm{W}) - yf(\tilde{\mathrm{x}}; \mathrm{W}'))| && (\ell \text{ is convex}) \\
=&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \left| \ell'(yf(\tilde{\mathrm{x}}; \mathrm{W})) \cdot \left(\sum_{s=1}^{m} a_s \phi(\langle \mathrm{w}_s, \tilde{\mathrm{x}} \rangle) - \sum_{s=1}^{m} a_s \phi(\langle \mathrm{w}'_s, \tilde{\mathrm{x}} \rangle) \right) \right| \\
\leq&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \left| \ell'(yf(\tilde{\mathrm{x}}; \mathrm{W})) \cdot \frac{1}{\sqrt{m}} \left(\sum_{s=1}^{m} |\langle \mathrm{w}_s - \mathrm{w}'_s, \tilde{\mathrm{x}} \rangle| \right) \right| && (\phi \text{ is 1-Lip}) \\
\leq&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \left| \ell'(yf(\tilde{\mathrm{x}}; \mathrm{W})) \cdot \sqrt{\sum_{s=1}^{m} \langle \mathrm{w}_s - \mathrm{w}'_s, \tilde{\mathrm{x}} \rangle^2} \right| && (\text{Cauchy's inequality}) \\
\leq&\max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} |\ell'(yf(\tilde{\mathrm{x}}; \mathrm{W})) \cdot \|\mathrm{W} - \mathrm{W}'\|_2 \cdot \|\tilde{\mathrm{x}}\|_2| \\
\leq&C_x \|\mathrm{W} - \mathrm{W}'\|_2 \cdot \max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} |\ell'(yf(\tilde{\mathrm{x}}; \mathrm{W}))| \\
\leq&C_x \|\mathrm{W} - \mathrm{W}'\|_2 \cdot \min\{1, \max_{\tilde{\mathrm{x}} \in \mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}; \mathrm{W}))\} && (\text{Lemma A.1}) \\
=&C_x \|\mathrm{W} - \mathrm{W}'\|_2 \cdot \min\{1, \ell_{rob}(\mathrm{z}, \mathrm{W})\}.
\end{aligned}$$

□

In the following proposition, we build the relationship between the generalization gap and uniform stability.

Proposition A.3. Let S and S' be any two neighboring data sets that differ only in one example. Let $\mathrm{W}_t = \mathcal{A}(S), \mathrm{W}'_t = \mathcal{A}(S')$ be the weight returned after running algorithm $\mathcal{A}$ for t steps using S and S', respectively. $\delta_{\mathcal{A}}(S, S') = \|\mathcal{A}(S) - \mathcal{A}(S')\|_F$. Then if $\sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S') < \frac{1}{C_x}$, we have

$$\mathbb{E}_{S \sim \mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \mathbb{E}_{S \sim \mathcal{D}^n} \frac{1}{1 - C_x \cdot \sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S')} \widehat{L}_{rob}(\mathrm{W}_t; S) \tag{4}$$

and

$$\mathbb{E}_{S \sim \mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \mathbb{E}_{S \sim \mathcal{D}^n} \widehat{L}_{rob}(\mathrm{W}_t; S) + C_x \cdot \sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S'). \tag{5}$$

Proof of Proposition A.3. Let S and S' differ in one example, and $\mathrm{z}' = S' \backslash S$.

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left(L_{rob}(\mathrm{W}_t) - \widehat{L}_{rob}(\mathrm{W}_t; S)\right) = \mathbb{E}_{S\cup\{\mathrm{z}'\}\sim\mathcal{D}^{n+1}}\left[(\ell_{rob}(\mathrm{z}', \mathrm{W}_t) - \ell_{rob}(\mathrm{z}', \mathrm{W}'_t))\right]. \tag{6}$$

Combining Lemma A.2 and Equation (6) we get

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left(L_{rob}(\mathrm{W}_t) - \widehat{L}_{rob}(\mathrm{W}_t; S)\right) \leq \mathbb{E}_{S\cup\{\mathrm{z}'\}\sim\mathcal{D}^{n+1}}[C_x\|\mathrm{W}_t - \mathrm{W}'_t\|_2 \cdot \min\{1, \ell_{rob}(\mathrm{z}', \mathrm{W}_t)\}].$$

Based on the definition of $\delta_{\mathcal{A}}$,

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\left(L_{rob}(\mathrm{W}_t) - \widehat{L}_{rob}(\mathrm{W}_t; S)\right) \\
&\leq C_x \cdot \sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \cdot \mathbb{E}_{S\cup\{\mathrm{z}'\}\sim\mathcal{D}^{n+1}} \min\{1, \ell_{rob}(\mathrm{z}', \mathrm{W}_t)\} \\
&\leq C_x \cdot \sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \cdot \min\{1, \mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_t)\}.
\end{aligned}$$

Simplifying this inequality, we get

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \mathbb{E}_{S\sim\mathcal{D}^n} \frac{1}{1 - C_x \cdot \sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S')} \widehat{L}_{rob}(\mathrm{W}_t; S)$$

and

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathrm{W}_t) \leq \mathbb{E}_{S\sim\mathcal{D}^n} \widehat{L}_{rob}(\mathrm{W}_t; S) + C_x \cdot \sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S').$$

□

The following lemma gives the weakly convex property of the robust loss (by considering the special case of $\beta_1 = 0$).

Lemma A.4. Given any data (x, y), for model with weight W, let $\tilde{\mathrm{x}}(\mathrm{W}) \in \mathcal{B}(\mathrm{x})$ be an β_1-optimal adversarial examples such that $\ell(yf(\tilde{\mathrm{x}}(\mathrm{W}), \mathrm{W})) \geq \max_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x})} \ell(yf(\tilde{\mathrm{x}}, \mathrm{W})) - \beta_1$. Then for any two weight matrices $\mathrm{W}^1, \mathrm{W}^2 \in \mathbb{R}^{d\times m}$, we have

$$\ell((\tilde{\mathrm{x}}(\mathrm{W}^2), y), \mathrm{W}^2) \geq \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1) + \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle - \beta_1 - \frac{HC_x^2}{2\sqrt{m}} \|\mathrm{W}^2 - \mathrm{W}^1\|_2^2.$$

Proof of Lemma A.4.

$$\begin{aligned}
&\ell((\tilde{\mathrm{x}}(\mathrm{W}^2), y), \mathrm{W}^2) - \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1) - \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle + \beta_1 \\
=&\ell(yf_{\mathrm{W}^2}(\tilde{\mathrm{x}}(\mathrm{W}^2))) - \ell(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1))) - \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle + \beta_1 \\
\geq& \max_{\tilde{\mathrm{x}}\in\mathcal{B}(\mathrm{x})} \ell(yf_{\mathrm{W}^2}(\tilde{\mathrm{x}})) - \ell(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1))) - \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle \\
&\qquad \text{(By definition of } \beta_1\text{-optimal adversarial examples)} \\
\geq& \ell(yf_{\mathrm{W}^2}(\tilde{\mathrm{x}}(\mathrm{W}^1))) - \ell(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1))) - \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle \\
\geq& \ell'(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1))) \cdot \left(yf_{\mathrm{W}^2}(\tilde{\mathrm{x}}(\mathrm{W}^1)) - yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1))\right) - \left\langle \nabla_{\mathrm{W}} \ell((\tilde{\mathrm{x}}(\mathrm{W}^1), y), \mathrm{W}^1), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle \\
&\qquad (\ell \text{ is convex}) \\
=& \ell'(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1)))y \sum_{s=1}^m a_s \left(\phi(\langle \mathrm{w}_s^2, \tilde{\mathrm{x}}(\mathrm{W}^1)\rangle) - \phi(\langle \mathrm{w}_s^1, \tilde{\mathrm{x}}(\mathrm{W}^1)\rangle) - \phi'(\langle \mathrm{w}_s^1, \tilde{\mathrm{x}}(\mathrm{W}^1)\rangle) \langle \mathrm{w}_s^2 - \mathrm{w}_s^1, \tilde{\mathrm{x}}(\mathrm{W}^1)\rangle\right) \\
\geq& -|\ell'(yf_{\mathrm{W}^1}(\tilde{\mathrm{x}}(\mathrm{W}^1)))| \sum_{s=1}^m \frac{1}{\sqrt{m}} \cdot \frac{H}{2} \langle \mathrm{w}_s^2 - \mathrm{w}_s^1, \tilde{\mathrm{x}}(\mathrm{W}^1)\rangle^2 \qquad (\phi \text{ is } H\text{-smooth}) \\
\geq& -1 \cdot \frac{H}{2\sqrt{m}} \|(\mathrm{W}^2 - \mathrm{W}^1)^{\mathsf{T}} \tilde{\mathrm{x}}(\mathrm{W}^1)\|_2^2 \qquad \text{(Lemma A.1)} \\
\geq& -\frac{HC_x^2}{2\sqrt{m}} \|\mathrm{W}^2 - \mathrm{W}^1\|_2^2. \qquad (\|\tilde{\mathrm{x}}(\mathrm{W}^1)\|_2 \leq C_x)
\end{aligned}$$

□

The following lemma tells us the gradient has a universal upper bound.

Lemma A.5. For any data (x, y) and any weight matrix W,

$$\|\nabla_{\mathrm{W}}\ell(yf(\mathrm{x};\mathrm{W}))\|_F \le C_x.$$

Proof of Lemma A.5. Since $\nabla_{\mathrm{W}}\ell(yf(\mathrm{x};\mathrm{W})) = [\ell'(yf(\mathrm{x};\mathrm{W}))ya_s\phi'(\langle \mathrm{w}_s, \mathrm{x}\rangle)\mathrm{x}]_{s=1}^m$,

$$\begin{aligned}
\|\nabla_{\mathrm{W}}\ell(yf(\mathrm{x};\mathrm{W}))\|_F &= \sqrt{\sum_{s=1}^m \|\ell'(yf(\mathrm{x};\mathrm{W}))ya_s\phi'(\langle \mathrm{W}_s, \mathrm{x}\rangle)\mathrm{x}\|_2^2} \\
&\le -C_x \cdot \ell'(yf(\mathrm{x};\mathrm{W})) && (|a_s| = \tfrac{1}{\sqrt{m}}, \phi' \le 1, \|\mathrm{x}\|_2 \le C_x) \\
&\le C_x. && \text{(Lemma A.1)}
\end{aligned}$$

□

B Missing Proofs in Section 3.1

Now we give the uniform argument stability upper bound.

Theorem B.1. (Restatement of Theorem 3.1) Let S and $S^{(i)}$ only differ in the i-th data. W_T and $\mathrm{W}_T^{(i)}$ denote the weight matrices returned after after running Algorithm GD for T iterations on S and $S^{(i)}$, respectively. Then we have

$$\|\mathrm{W}_T - \mathrm{W}_T^{(i)}\|_F^2 \le e^{1+\frac{2HC_x^2\eta T}{\sqrt{m}}}\left(4C_x^2\eta^2(T+1) + \frac{4C_x^2\eta^2(T+1)^2}{n^2} + 4\beta_1\eta(T+1)\right).$$

Proof of Theorem B.1. For any weight matrix W and any perturbation of the j-th data $\tilde{\mathrm{x}}_j \in \mathcal{B}(\mathrm{x}_j)$ $(j \ne i)$, define $L_{S^{\setminus i}}(\mathrm{W}; \{\tilde{\mathrm{x}}_j\}_{j\ne i}) = \frac{1}{n}\sum_{j\ne i}\ell(y_j f(\tilde{\mathrm{x}}_j;\mathrm{W}))$ to be the loss of the (perturbed) data set without including the i-th data. Let $\tilde{\mathrm{x}}_j(\mathrm{W}_t)$ denote the β_1-optimal adversarial example of x_j given W_t, and $\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})$ denote the β_1-optimal adversarial example of x_j given $\mathrm{W}_t^{(i)}$. We first show that the gradient is an "almost" monotone operator, which is derived from the weakly convex property of the robust loss (see Lemma A.4).

$$\begin{aligned}
&\beta_1 + \frac{HC_x^2}{2\sqrt{m}}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\ne i}) \\
\ge&\frac{1}{n}\sum_{j\ne i}\left[\beta_1 + \frac{HC_x^2}{2\sqrt{m}}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + \ell(y_j f(\tilde{\mathrm{x}}_j(\mathrm{W}_t);\mathrm{W}_t))\right] \\
\ge&\frac{1}{n}\sum_{j\ne i}\left[\ell(y_j f(\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)});\mathrm{W}_t^{(i)})) + \left\langle \nabla_{\mathrm{W}}\ell(y_j f(\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)});\mathrm{W}_t^{(i)})), \mathrm{W}_t - \mathrm{W}_t^{(i)}\right\rangle\right] \quad \text{(Lemma A.4)} \\
=&L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\ne i}) + \left\langle \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\ne i}), \mathrm{W}_t - \mathrm{W}_t^{(i)}\right\rangle.
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
&\beta_1 + \frac{HC_x^2}{2\sqrt{m}}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\ne i}) \\
\ge&L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\ne i}) + \left\langle \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\ne i}), \mathrm{W}_t^{(i)} - \mathrm{W}_t\right\rangle.
\end{aligned}$$

Adding these two inequalities together, we get

$$\begin{aligned}
&\left\langle \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\ne i}) - \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\ne i}), \mathrm{W}_t^{(i)} - \mathrm{W}_t\right\rangle \\
\ge& -2\beta_1 - \frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2. \qquad (7)
\end{aligned}$$

Now we start upper bounding $\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F$.

$$
\begin{aligned}
&\|\mathrm{W}_{t+1} - \mathrm{W}_{t+1}^{(i)}\|_F^2 \\
=&\|\mathrm{W}_t - \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\neq i}) - \frac{\eta}{n}\nabla_{\mathrm{W}}\ell(y_i f(\tilde{\mathrm{x}}_i(\mathrm{W}_t); \mathrm{W}_t)) \\
&\quad - \mathrm{W}_t^{(i)} + \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\neq i}) + \frac{\eta}{n}\nabla_{\mathrm{W}}\ell(y_i' f(\tilde{\mathrm{x}}_i'(\mathrm{W}_t^{(i)}); \mathrm{W}_t^{(i)}))\|_F^2 \\
\leq& \left(\|\mathrm{W}_t - \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\neq i}) - \mathrm{W}_t^{(i)} + \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\neq i})\|_F + \frac{2\eta C_x}{n} \right)^2 \\
&\hfill \text{(Lemma A.5)} \\
\leq& \frac{T+2}{T+1}\|\mathrm{W}_t - \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\neq i}) - \mathrm{W}_t^{(i)} + \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\neq i})\|_F^2 \\
&\quad + (T+2)\frac{4\eta^2 C_x^2}{n^2} \qquad ((a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2 \text{ for } p > 0.) \\
=& \frac{T+2}{T+1}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + \frac{T+2}{T+1}\|\eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\neq i}) - \eta\nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\neq i})\|_F^2 \\
&\quad - 2\eta\frac{T+2}{T+1}\left\langle \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t^{(i)}; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t^{(i)})\}_{j\neq i}) - \nabla_{\mathrm{W}} L_{S^{\setminus i}}(\mathrm{W}_t; \{\tilde{\mathrm{x}}_j(\mathrm{W}_t)\}_{j\neq i}), \mathrm{W}_t^{(i)} - \mathrm{W}_t \right\rangle \\
&\quad + (T+2)\frac{4\eta^2 C_x^2}{n^2} \\
\leq& \frac{T+2}{T+1}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + \frac{T+2}{T+1}4\eta^2 C_x^2 + \frac{T+2}{T+1}(4\beta_1\eta + \frac{2HC_x^2}{\sqrt{m}}\eta\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2) + (T+2)\frac{4\eta^2 C_x^2}{n^2} \\
&\hfill \text{(Lemma A.5 and Equation (7))} \\
=& \frac{T+2}{T+1}\left((1 + \frac{2HC_x^2}{\sqrt{m}}\eta)\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + 4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right).
\end{aligned}
$$

Define $\gamma = \frac{T+2}{T+1}(1 + \frac{2HC_x^2}{\sqrt{m}}\eta)$, then the inequality above can be written as

$$
\|\mathrm{W}_{t+1} - \mathrm{W}_{t+1}^{(i)}\|_F^2 \leq \gamma\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + \frac{T+2}{T+1}\left(4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right).
$$

Dividing both sides by γ^{t+1},

$$
\frac{\|\mathrm{W}_{t+1} - \mathrm{W}_{t+1}^{(i)}\|_F^2}{\gamma^{t+1}} \leq \frac{\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2}{\gamma^t} + \frac{T+2}{T+1}\frac{4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2}}{\gamma^{t+1}}.
$$

Summing up this inequality for $t = 0, 1, \ldots, T-1$, we obtain

$$
\begin{aligned}
\|\mathrm{W}_T - \mathrm{W}_T^{(i)}\|_F^2 &\leq \gamma^T \sum_{t=0}^{T-1} \frac{T+2}{T+1}\frac{4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2}}{\gamma^{t+1}} \\
&\leq \frac{\gamma^T}{\gamma - 1}\frac{T+2}{T+1}\left(4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right) \\
&\leq (T+1)\gamma^T\frac{T+2}{T+1}\left(4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right) \\
&= (T+1)(\frac{T+2}{T+1})^{T+1}(1 + \frac{2HC_x^2}{\sqrt{m}}\eta)^T\left(4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right) \\
&\leq (T+1)e \cdot e^{\frac{2HC_x^2\eta T}{\sqrt{m}}}\left(4\eta^2 C_x^2 + 4\beta_1\eta + (T+1)\frac{4\eta^2 C_x^2}{n^2} \right) \qquad (1 + x \leq e^x) \\
&= e^{1 + \frac{2HC_x^2\eta T}{\sqrt{m}}}\left(4C_x^2\eta^2(T+1) + \frac{4C_x^2\eta^2(T+1)^2}{n^2} + 4\beta_1\eta(T+1) \right).
\end{aligned}
$$

□

The proof of Theorem 3.1 is immediately obtained from Theorem B.1 by observing $e^{1+\frac{2HC_x^2\eta T}{\sqrt{m}}} \leq e^3$.

Next we give an optimization guarantee. We show that when T is sufficiently large, the robust training loss can approach the adversarial regularized empirical risk minimization oracle. We do not need early stopping in the Theorem below.

Theorem B.2. After running Algorithm GD for T iterations, we have

$$\min_{0\leq t\leq T}\widehat{L}_{rob}(\mathrm{W}_t;S)\leq \min_{\mathrm{W}\in\mathbb{R}^{d\times m}}\left(\widehat{L}_{rob}(\mathrm{W};S)+\frac{HC_x^2}{2\sqrt{m}(1-\frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{T+1}})}\|\mathrm{W}-\mathrm{W}_0\|_F^2+\frac{C_x^2\eta}{2}+\beta_1\right).$$

Proof of Theorem B.2. For any given W, we have

$$\begin{aligned}
&\|\mathrm{W}-\mathrm{W}_{t+1}\|_F^2\\
=&\|\mathrm{W}-\mathrm{W}_t+\eta\nabla_{\mathrm{W}}L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n)\|_F^2\\
=&\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2\|\nabla_{\mathrm{W}}L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n)\|_F^2+2\eta\left\langle\nabla_{\mathrm{W}}L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n),\mathrm{W}-\mathrm{W}_t\right\rangle\\
\leq&\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2C_x^2+\frac{HC_x^2\eta}{\sqrt{m}}\|\mathrm{W}-\mathrm{W}_t\|_F^2+2\eta\widehat{L}_{rob}(\mathrm{W};S)-2\eta L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n)\\
&\qquad\qquad\text{(Lemma A.5 and Lemma A.4)}\\
=&(1+\frac{HC_x^2\eta}{\sqrt{m}})\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2C_x^2+2\eta\widehat{L}_{rob}(\mathrm{W};S)-2\eta L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n).
\end{aligned}$$

Dividing both sides by $(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}$ we get

$$\frac{\|\mathrm{W}-\mathrm{W}_{t+1}\|_F^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}\leq\frac{\|\mathrm{W}-\mathrm{W}_t\|_F^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t}}+\frac{\eta^2C_x^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}+\frac{2\eta(\widehat{L}_{rob}(\mathrm{W};S)-L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n))}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}. \tag{8}$$

Taking the sum for $t=0,1,\ldots,T$ we get

$$\begin{aligned}
&2\eta\sum_{t=0}^{T}\frac{\widehat{L}_{rob}(\mathrm{W};S)}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}+\|\mathrm{W}-\mathrm{W}_0\|_F^2+\sum_{t=0}^{T}\frac{\eta^2C_x^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}\\
\geq&2\eta\sum_{t=0}^{T}\frac{L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n)}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \qquad (9)\\
\geq&2\eta\sum_{t=0}^{T}\frac{\widehat{L}_{rob}(\mathrm{W}_t;S)-\beta_1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}\\
\geq&2\eta\min_{0\leq t\leq T}\widehat{L}_{rob}(\mathrm{W}_t;S)\cdot\sum_{t=0}^{T}\frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}-2\eta\sum_{t=0}^{T}\frac{\beta_1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}.
\end{aligned}$$

Simplifying the above inequality, we have

$$\min_{0\leq t\leq T}\widehat{L}_{rob}(\mathrm{W}_t;S)\leq\widehat{L}_{rob}(\mathrm{W};S)+\frac{HC_x^2}{2\sqrt{m}(1-\frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{T+1}})}\|\mathrm{W}-\mathrm{W}_0\|_F^2+\frac{C_x^2\eta}{2}+\beta_1.$$

□

Theorem 3.2. Define $\alpha_1(\eta,T):=\mathcal{O}(C_x^2\eta\sqrt{T}+C_x^2\frac{\eta T}{n}+C_x\sqrt{\beta_1\eta T})$. Assume that the width of the networks satisfies $m\geq H^2C_x^4\eta^2(T+1)^2$, and $\alpha_1(\eta,T)<1$. Then, after T iterations of Algorithm 1 with GD, we have

$$\min_{[\frac{9T}{10}]\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathrm{W}_t)\leq\frac{17\alpha_1(\eta,T)}{1-\alpha_1(\eta,T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}+\beta_1\right],$$

and

$$\min_{0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathrm{W}_t)\leq\frac{1}{1-\alpha_1(\eta,T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}+\beta_1\right].$$

Proof of Theorem 3.2. Define $t_0 := [\frac{9T}{10}]$ and $k := \frac{1}{1+\frac{HC_x^2\eta}{\sqrt{m}}}$. From equation (9), we have

$$\begin{aligned}
&2\eta\sum_{t=0}^{T}\frac{\widehat{L}_{rob}(\mathrm{W};S)}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}+\|\mathrm{W}-\mathrm{W}_0\|_F^2+\sum_{t=0}^{T}\frac{\eta^2C_x^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}\\
\geq&2\eta\sum_{t=t_0}^{T}\frac{L(\mathrm{W}_t;\{\tilde{\mathrm{x}}_i(\mathrm{W}_t)\}_{i=1}^n)}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}\\
\geq&2\eta\sum_{t=t_0}^{T}\frac{\widehat{L}_{rob}(\mathrm{W}_t;S)-\beta_1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}.
\end{aligned}$$

Taking minimum over all weight matrices W,

$$\begin{aligned}
&\sum_{t=t_0}^{T}k^{t+1}\left(\widehat{L}_{rob}(\mathrm{W}_t;S)-\beta_1\right)\\
\leq&\min_{\mathrm{W}}\left(\sum_{t=0}^{T}k^{t+1}\widehat{L}_{rob}(\mathrm{W};S)+\frac{\|\mathrm{W}-\mathrm{W}_0\|_F^2}{2\eta}+\sum_{t=0}^{T}\frac{\eta C_x^2}{2}k^{t+1}\right)\\
=&\sum_{t=0}^{T}k^{t+1}\cdot\min_{\mathrm{W}}\left(\widehat{L}_{rob}(\mathrm{W};S)+\frac{\|\mathrm{W}-\mathrm{W}_0\|_F^2}{2\eta\sum_{t=0}^{T}k^{t+1}}+\frac{C_x^2\eta}{2}\right)\\
\leq&\sum_{t=0}^{T}k^{t+1}\cdot\min_{\mathrm{W}}\left(\widehat{L}_{rob}(\mathrm{W};S)+\frac{\|\mathrm{W}-\mathrm{W}_0\|_F^2}{\eta(T+1)}+\frac{C_x^2\eta}{2}\right)\qquad(m\geq H^2C_x^4\eta^2(T+1)^2)\\
\leq&\sum_{t=0}^{T}k^{t+1}\cdot\left(\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right).
\end{aligned}$$

Taking the expectation on both sides, we get

$$\begin{aligned}
\sum_{t=t_0}^{T}k^{t+1}\left(\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)-\beta_1\right)&\leq\sum_{t=t_0}^{T}k^{t+1}\left(\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)-\beta_1\right)\\
&\leq\sum_{t=0}^{T}k^{t+1}\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right).
\end{aligned}$$

Simplifying the equation above, we get

$$\begin{aligned}
\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)&\leq\beta_1+\frac{\sum_{t=0}^{T}k^{t+1}}{\sum_{t=t_0}^{T}k^{t+1}}\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right)\qquad(10)\\
&\leq\beta_1+\left(\sum_{r=0}^{9}\left(\frac{1}{k}\right)^{r\frac{T+1}{10}}\right)\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right)\\
&\leq\beta_1+\left(\sum_{r=0}^{9}e^{\frac{r}{10}}\right)\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right)\\
&\qquad\qquad(m\geq H^2C_x^4\eta^2(T+1)^2)\\
&\leq\beta_1+17\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}}+\frac{C_x^2\eta}{2}\right).
\end{aligned}$$

Equation (4) gives us for any $t \le T$, $\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathrm{W}_t) \le \frac{\alpha_1(\eta,T)}{1-\alpha_1(\eta,T)}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)$. Therefore,

$$\begin{aligned}\min_{t_0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathrm{W}_t) &\le \frac{\alpha_1(\eta,T)}{1-\alpha_1(\eta,T)}\min_{t_0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)\\ &\le \frac{\alpha_1(\eta,T)}{1-\alpha_1(\eta,T)}\left(\beta_1 + 17\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2}\right)\right).\end{aligned}$$

The proof of the second statement takes a similar approach. Following the same procedure, we can replace t_0 by 0 in equation (10), and get

$$\min_{0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S) \le \mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2} + \beta_1.$$

Combining with equation (4),

$$\begin{aligned}\min_{0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathrm{W}_t) &\le \frac{1}{1-\alpha_1(\eta,T)}\min_{0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S)\\ &\le \frac{1}{1-\alpha_1(\eta,T)}\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + \frac{C_x^2\eta}{2} + \beta_1\right).\end{aligned}$$

□

Corollary 3.3. After $T \le \mathcal{O}(\min\{n^2, \frac{1}{\beta_1^2}\})$ iterations of Algorithm 1 with GD using a step size of $\eta = \Theta(\frac{1}{C_x^2\sqrt{T}})$ on a network with width $m \ge \Omega(T)$, for any weight matrix W

$$\min_{0\le t\le T}\mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathrm{W}_t) \le 1.1L_{rob}(\mathrm{W}) + \mathcal{O}\left(\frac{C_x^2\|\mathrm{W}-\mathrm{W}_0\|_F^2}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Proof of Corollary 3.3. Under the conditions of the corollary, we have $m \ge H^2C_x^4\eta^2(T+1)^2$, and $\alpha_1(\eta,T) = \mathcal{O}(C_x^2\eta\sqrt{T} + C_x^2\frac{\eta T}{n} + C_x\sqrt{\beta_1\eta T})$ can be small enough so that $\frac{1}{1-\alpha_1(\eta,T)} \le 1.1$. Then it is clear that this corollary is a special case of Theorem 3.2. □

We now extend the previous ideas to stochastic adversarial training.

Lemma B.3. After T iterations of Algorithm 1 with SGD, for any weight matrix W,

$$\min_{0\le t\le T}\left(\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W}_t) - \ell_{rob}(\mathrm{z}_{t+1},\mathrm{W})\right) \le \frac{HC_x^2}{2\sqrt{m}(1-\frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{T+1}})}\|\mathrm{W}-\mathrm{W}_0\|_F^2 + \frac{C_x^2\eta}{2} + \beta_1.$$

Proof of Lemma B.3. The proof proceeds similarly as Theorem B.2.

$$\begin{aligned}&\|\mathrm{W}-\mathrm{W}_{t+1}\|_F^2\\ =&\|\mathrm{W}-\mathrm{W}_t+\eta\nabla_{\mathrm{W}}\ell((\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t),y_{t+1}),\mathrm{W}_t)\|_F^2\\ =&\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2\|\nabla_{\mathrm{W}}\ell((\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t),y_{t+1}),\mathrm{W}_t)\|_F^2+2\eta\left\langle\nabla_{\mathrm{W}}\ell((\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t),y_{t+1}),\mathrm{W}_t),\mathrm{W}-\mathrm{W}_t\right\rangle\\ \le&\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2C_x^2+\frac{HC_x^2\eta}{\sqrt{m}}\|\mathrm{W}-\mathrm{W}_t\|_F^2+2\eta\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W})-2\eta\ell((\tilde{\mathrm{x}}_{t+1}(\mathrm{W}_t),y_{t+1}),\mathrm{W}_t)\\ &\qquad\qquad\text{(Lemma A.5 and Lemma A.4)}\\ \le&(1+\frac{HC_x^2\eta}{\sqrt{m}})\|\mathrm{W}-\mathrm{W}_t\|_F^2+\eta^2C_x^2+2\eta\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W})-2\eta\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W}_t)+2\eta\beta_1.\end{aligned}$$

Dividing both sides by $(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}$, we get

$$\frac{\|\mathrm{W}-\mathrm{W}_{t+1}\|_F^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \le \frac{\|\mathrm{W}-\mathrm{W}_t\|_F^2}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t}} + \frac{\eta^2C_x^2+2\eta\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W})-2\eta\ell_{rob}(\mathrm{z}_{t+1},\mathrm{W}_t)+2\eta\beta_1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}.$$

Taking the sum of the above equation for $t = 0, 1, \ldots, T$:

$$\begin{aligned}
&\|\mathrm{W} - \mathrm{W}_0\|_F^2 + \sum_{t=0}^{T} \frac{\eta^2 C_x^2 + 2\eta\beta_1}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \\
\geq & 2\eta \sum_{t=0}^{T} \frac{\ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W}_t) - \ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W})}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \\
\geq & 2\eta \sum_{t=0}^{T} \frac{\min_{0\leq t\leq T} (\ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W}_t) - \ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W}))}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}.
\end{aligned} \tag{11}$$

Simplifying the above inequality we have

$$\min_{0\leq t\leq T} (\ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W}_t) - \ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W})) \leq \frac{HC_x^2}{2\sqrt{m}(1 - \frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{T+1}})} \|\mathrm{W} - \mathrm{W}_0\|_F^2 + \frac{C_x^2\eta}{2} + \beta_1.$$

□

Theorem 3.4. After T iterations of Algorithm 1 with SGD on a network of width $m \geq H^2 C_x^4 \eta^2 (T+1)^2$ we have that for any weight matrix W,

$$\min_{0\leq t\leq T} \mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} L_{rob}(\mathrm{W}_t) \leq L_{rob}(\mathrm{W}) + \frac{\|\mathrm{W} - \mathrm{W}_0\|_F^2}{\eta(T+1)} + \frac{C_x^2\eta}{2} + \beta_1.$$

Proof of Theorem 3.4. Taking the expectation over $S \sim \mathcal{D}^n$ on both sides of Equation (11), we obtain

$$\begin{aligned}
&\|\mathrm{W} - \mathrm{W}_0\|_F^2 + \sum_{t=0}^{T} \frac{\eta^2 C_x^2 + 2\eta\beta}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \\
\geq & 2\eta \sum_{t=0}^{T} \frac{\mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} \mathbb{E}_{\mathrm{z}_{t+1}\sim\mathcal{D}} \ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W}_t) - \mathbb{E}_{\mathrm{z}_{t+1}\sim\mathcal{D}} \ell_{rob}(\mathrm{z}_{t+1}, \mathrm{W})}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \\
= & 2\eta \sum_{t=0}^{T} \frac{\mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} L_{rob}(\mathrm{W}_t) - L_{rob}(\mathrm{W})}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}} \\
\geq & 2\eta \sum_{t=0}^{T} \frac{\min_{0\leq t\leq T} \mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} L_{rob}(\mathrm{W}_t) - L_{rob}(\mathrm{W})}{(1 + \frac{HC_x^2\eta}{\sqrt{m}})^{t+1}}.
\end{aligned}$$

Simplifying the above inequality, we get

$$\begin{aligned}
\min_{0\leq t\leq T} \mathbb{E}_{\{\mathrm{z}_1,\ldots,\mathrm{z}_t\}\sim\mathcal{D}^t} L_{rob}(\mathrm{W}_t) &\leq L_{rob}(\mathrm{W}) + \frac{HC_x^2}{2\sqrt{m}(1 - \frac{1}{(1+\frac{HC_x^2\eta}{\sqrt{m}})^{T+1}})} \|\mathrm{W} - \mathrm{W}_0\|_F^2 + \frac{C_x^2\eta}{2} + \beta_1 \\
&\leq L_{rob}(\mathrm{W}) + \frac{\|\mathrm{W} - \mathrm{W}_0\|_F^2}{\eta(T+1)} + \frac{C_x^2\eta}{2} + \beta_1. \\
&\qquad (m \geq H^2 C_x^4 \eta^2 (T+1)^2)
\end{aligned}$$

□

C Missing Proofs in Section 3.2

From Lemma A.4, for any $\mu < \frac{\sqrt{m}}{HC_x^2}$, $L_{rob}(\mathrm{U}) + \frac{1}{2\mu}\|\mathrm{U} - \mathrm{W}\|_F^2$ is strongly convex in U. Recall that the Moreau envelope is defined as

$$M^{\mu}(\mathrm{W}; S) = \min_{\mathrm{U}} \left(\widehat{L}_{rob}(\mathrm{U}; S) + \frac{1}{2\mu}\|\mathrm{U} - \mathrm{W}\|_F^2 \right).$$

The minimizer of the optimization problem above is denoted as

$$\mathrm{U}^{\mu}(\mathrm{W};S)=\operatorname*{argmin}_{\mathrm{U}}\left(\widehat{L}_{rob}(\mathrm{U};S)+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right).$$

We borrow a few properties of the Moreau envelope from Xiao et al. [2024].

Lemma C.1. For any $\mu<\frac{\sqrt{m}}{HC_x^2}$,

1. $\min_{\mathrm{W}} M^{\mu}(\mathrm{W};S)$ has the same global solution set as $\min_{\mathrm{W}} \widehat{L}_{rob}(\mathrm{W};S)$.
2. The gradient of $M^{\mu}(\mathrm{W};S)$ is $\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W};S)=\frac{1}{\mu}\left(\mathrm{W}-\mathrm{U}^{\mu}(\mathrm{W};S)\right)$.
3. $M^{\mu}(\mathrm{W};S)+\frac{\|\mathrm{W}\|_F^2}{2\left(\frac{\sqrt{m}}{HC_x^2}-\mu\right)}$ is convex.
4. $\mathrm{U}^{\mu}(\mathrm{W};S)$ is $\frac{\frac{\sqrt{m}}{HC_x^2}}{\frac{\sqrt{m}}{HC_x^2}-\mu}$-Lipschitz in W w.r.t the Frobenius norm.
5. $M^{\mu}(\mathrm{W};S)$ is $\max\left\{\frac{1}{\mu},\frac{1}{\frac{\sqrt{m}}{HC_x^2}-\mu}\right\}$-smooth.
6. $\widehat{L}_{rob}(\mathrm{W};S)-\frac{C_x^2}{2\left(\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}}\right)}\le M^{\mu}(\mathrm{W};S)\le \widehat{L}_{rob}(\mathrm{W};S)$.
7. $\|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W};S)\|_F\le C_x$.

Proof of Lemma C.1. The first 5 statements are covered in the proof of [Xiao et al., 2024, Lemma A.1]. For the statement 6,

$$\begin{aligned}
\widehat{L}_{rob}(\mathrm{W};S)&=\widehat{L}_{rob}(\mathrm{W};S)+\frac{1}{2\mu}\|\mathrm{W}-\mathrm{W}\|_F^2\\
&\ge\min_{\mathrm{U}}\left(\widehat{L}_{rob}(\mathrm{U};S)+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right)\\
&=M^{\mu}(\mathrm{W};S)\\
&=\widehat{L}_{rob}(\mathrm{W};S)+\min_{\mathrm{U}}\left(\widehat{L}_{rob}(\mathrm{U};S)-\widehat{L}_{rob}(\mathrm{W};S)+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right)\\
&\ge\widehat{L}_{rob}(\mathrm{W};S)+\min_{\mathrm{U}}\left(\left\langle\nabla_{\mathrm{W}}\widehat{L}_{rob}(\mathrm{W};S),\mathrm{U}-\mathrm{W}\right\rangle-\frac{HC_x^2}{2\sqrt{m}}\|\mathrm{U}-\mathrm{W}\|_F^2+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right)\\
&\qquad\text{(“almost” convex robust loss from Lemma A.4)}\\
&\ge\widehat{L}_{rob}(\mathrm{W};S)+\min_{\mathrm{U}}\left(-C_x\|\mathrm{U}-\mathrm{W}\|_F-\frac{HC_x^2}{2\sqrt{m}}\|\mathrm{U}-\mathrm{W}\|_F^2+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right)\\
&\qquad\text{(Lemma A.5)}\\
&=\widehat{L}_{rob}(\mathrm{W};S)-\frac{C_x^2}{2\left(\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}}\right)}.
\end{aligned}$$

Now we prove statement 7. For any $\gamma\in(0,1)$, from the definition of the Moreau envelope,

$$\begin{aligned}
&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S)+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2\\
\le&\widehat{L}_{rob}((1-\gamma)\mathrm{W}+\gamma\mathrm{U}^{\mu}(\mathrm{W};S);S)+\frac{1}{2\mu}\|(1-\gamma)\mathrm{W}+\gamma\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2\\
&\qquad\text{($\mathrm{U}^{\mu}(\mathrm{W};S)$ obtains the minimum)}\\
\le&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S)+C_x(1-\gamma)\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F+\frac{\gamma^2}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2.\\
&\qquad\text{(The robust loss is C_x-Lip from Lemma A.2)}
\end{aligned}$$

Simplifying the inequality above, we get

$$\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F \le \frac{2\mu C_x}{1+\gamma}.$$

Let $\gamma \to 1$, we get

$$\|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W};S)\|_F = \frac{1}{\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F \le C_x.$$

□

Next we show a result similar as [Xiao et al., 2024, Lemma A.2]. They use the first-order optimal condition to prove the result, which only holds for the smooth loss functions. Here we give a different proof that doesn't depend on the subgradient, so it can be applied to the robust loss.

Lemma C.2. Let $\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S)$ be any β_2-optimal minimizer of $\min_{\mathrm{U}} \left(\widehat{L}_{rob}(\mathrm{U};S)+\frac{1}{2\mu}\|\mathrm{U}-\mathrm{W}\|_F^2\right)$. For two data sets S and $S^{(i)}$ that differ in only one example and any weight matrix W, we have

$$\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})\|_F \le \frac{2C_x}{n\left(\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}}\right)}$$

and

$$\|\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S)\|_F \le \sqrt{\frac{2\beta_2}{\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}}}}.$$

Proof of Lemma C.2. From strong convexity of the regularized robust loss, we have

$$\begin{aligned}
&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S^{(i)});S)+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})-\mathrm{W}\|_F^2\\
\ge&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S)+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2+(\frac{1}{2\mu}-\frac{HC_x^2}{2\sqrt{m}})\|\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})-\mathrm{U}^{\mu}(\mathrm{W};S)\|_F^2,
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S^{(i)})+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2\\
\ge&\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S^{(i)});S^{(i)})+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})-\mathrm{W}\|_F^2+(\frac{1}{2\mu}-\frac{HC_x^2}{2\sqrt{m}})\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})\|_F^2.
\end{aligned}$$

Adding these two inequalities,

$$\begin{aligned}
&(\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}})\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})\|_F^2\\
\le&\Big[\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S^{(i)});S)-\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S^{(i)});S^{(i)})\Big]+\Big[\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S^{(i)})-\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S)\Big]\\
=&\frac{1}{n}\Big[\ell_{rob}(\mathrm{z}_i,\mathrm{U}^{\mu}(\mathrm{W};S^{(i)}))-\ell_{rob}(\mathrm{z}_i',\mathrm{U}^{\mu}(\mathrm{W};S^{(i)}))\Big]+\frac{1}{n}[\ell_{rob}(\mathrm{z}_i',\mathrm{U}^{\mu}(\mathrm{W};S))-\ell_{rob}(\mathrm{z}_i,\mathrm{U}^{\mu}(\mathrm{W};S))]\\
=&\frac{1}{n}\Big[\ell_{rob}(\mathrm{z}_i,\mathrm{U}^{\mu}(\mathrm{W};S^{(i)}))-\ell_{rob}(\mathrm{z}_i,\mathrm{U}^{\mu}(\mathrm{W};S))\Big]+\frac{1}{n}\Big[\ell_{rob}(\mathrm{z}_i',\mathrm{U}^{\mu}(\mathrm{W};S))-\ell_{rob}(\mathrm{z}_i',\mathrm{U}^{\mu}(\mathrm{W};S^{(i)}))\Big]\\
\le&\frac{2C_x}{n}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})\|_F. \qquad \text{(Lemma A.2)}
\end{aligned}$$

Therefore,

$$\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S^{(i)})\|_F \le \frac{2C_x}{n\left(\frac{1}{\mu}-\frac{HC_x^2}{\sqrt{m}}\right)}. \tag{12}$$

From the strong convexity of the regularized robust loss, we also have

$$\begin{aligned}\beta_2 \geq &\left(\widehat{L}_{rob}(\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S);S)+\frac{1}{2\mu}\|\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2\right)\\ &-\left(\widehat{L}_{rob}(\mathrm{U}^{\mu}(\mathrm{W};S);S)+\frac{1}{2\mu}\|\mathrm{U}^{\mu}(\mathrm{W};S)-\mathrm{W}\|_F^2\right)\\ \geq &(\frac{1}{2\mu}-\frac{HC_x^2}{2\sqrt{m}})\|\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S)\|_F^2.\end{aligned}$$

We get $\|\tilde{\mathrm{U}}^{\mu}(\mathrm{W};S)-\mathrm{U}^{\mu}(\mathrm{W};S)\|_F \leq \sqrt{\frac{\beta_2}{\frac{1}{2\mu}-\frac{HC_x^2}{2\sqrt{m}}}}$. □

Now we show an upper bound that is key to bounding the stability of the weight matrix.

Lemma C.3. For any $\eta \leq \mu \leq \frac{\sqrt{m}}{2HC_x^2}$ and any weight matrices W_1 and W_2,

$$\|\mathrm{W}^1-\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)-\mathrm{W}^2+\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)\|_F^2 \leq \frac{\|\mathrm{W}^1-\mathrm{W}^2\|_F^2}{1-\frac{4HC_x^2\eta}{\sqrt{m}}}.$$

Proof of Lemma C.3. Define $\psi_1(\mathrm{W}) = M^{\mu}(\mathrm{W};S) - M^{\mu}(\mathrm{W}^1;S) - \left\langle\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S), \mathrm{W}-\mathrm{W}^1\right\rangle$. From Lemma C.1, $\psi_1(\mathrm{W})$ is $\frac{1}{\mu}$-smooth. From the standard descent lemma of smooth functions, for any $\eta \leq \mu$,

$$\begin{aligned}\psi_1(\mathrm{W}^2-\eta\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2)) &\leq \psi_1(\mathrm{W}^2)-\frac{\eta}{2}\|\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2)\|_F^2\\ &= \psi_1(\mathrm{W}^2)-\frac{\eta}{2}\|\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)-\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)\|_F^2.\end{aligned}$$

Since $\psi_1(\mathrm{W})+\frac{\|\mathrm{W}\|_F^2}{2\left(\frac{\sqrt{m}}{HC_x^2}-\mu\right)}$ is convex from Lemma C.1,

$$\begin{aligned}&\psi_1(\mathrm{W}^2-\eta\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2))\\ \geq&\psi_1(\mathrm{W}^1)+\left\langle\nabla\psi_1(\mathrm{W}^1), \mathrm{W}^2-\eta\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2)-\mathrm{W}^1\right\rangle-\frac{\|\mathrm{W}^2-\eta\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2)-\mathrm{W}^1\|_F^2}{2\left(\frac{\sqrt{m}}{HC_x^2}-\mu\right)}\\ =&\psi_1(\mathrm{W}^1)-\frac{\|\mathrm{W}^2-\eta\nabla_{\mathrm{W}}\psi_1(\mathrm{W}^2)-\mathrm{W}^1\|_F^2}{2\left(\frac{\sqrt{m}}{HC_x^2}-\mu\right)}\\ \geq&\psi_1(\mathrm{W}^1)-\frac{\|\mathrm{W}^2-\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)-\mathrm{W}^1+\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)\|_F^2}{\frac{\sqrt{m}}{HC_x^2}}.\end{aligned}$$

Combining the two inequalities above,

$$\begin{aligned}&M^{\mu}(\mathrm{W}^2;S)-M^{\mu}(\mathrm{W}^1;S)-\left\langle\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S), \mathrm{W}^2-\mathrm{W}^1\right\rangle\\ =&\psi_1(\mathrm{W}^2)-\psi_1(\mathrm{W}^1)\\ \geq&\frac{\eta}{2}\|\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)-\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)\|_F^2-\frac{\|\mathrm{W}^2-\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)-\mathrm{W}^1+\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)\|_F^2}{\frac{\sqrt{m}}{HC_x^2}}.\end{aligned}$$

Similarly, we can get the counterpart of this equation:

$$\begin{aligned}&M^{\mu}(\mathrm{W}^1;S)-M^{\mu}(\mathrm{W}^2;S)-\left\langle\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S), \mathrm{W}^1-\mathrm{W}^2\right\rangle\\ \geq&\frac{\eta}{2}\|\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)-\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)\|_F^2-\frac{\|\mathrm{W}^1-\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^1;S)-\mathrm{W}^2+\eta\nabla_{\mathrm{W}}M^{\mu}(\mathrm{W}^2;S)\|_F^2}{\frac{\sqrt{m}}{HC_x^2}}.\end{aligned}$$

Adding these two inequalities, we get

$$\begin{aligned}&\left\langle \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S), \mathrm{W}^2 - \mathrm{W}^1 \right\rangle \\ \geq& \eta \|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2 - \frac{2\|\mathrm{W}^1 - \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \mathrm{W}^2 + \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2}{\frac{\sqrt{m}}{HC_x^2}}.\end{aligned}$$

Thus,

$$\begin{aligned}&\|\mathrm{W}^1 - \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \mathrm{W}^2 + \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2 \\ =& \|\mathrm{W}^1 - \mathrm{W}^2\|_F^2 + \eta^2 \|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2 \\ &\quad - 2\eta \left\langle \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S), \mathrm{W}^1 - \mathrm{W}^2 \right\rangle \\ \leq& \|\mathrm{W}^1 - \mathrm{W}^2\|_F^2 + \eta^2 \|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2 \\ &\quad - 2\eta^2 \|\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2 \\ &\quad + \frac{4\eta\|\mathrm{W}^1 - \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \mathrm{W}^2 + \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2}{\frac{\sqrt{m}}{HC_x^2}} \\ \leq& \|\mathrm{W}^1 - \mathrm{W}^2\|_F^2 + \frac{4HC_x^2\eta\|\mathrm{W}^1 - \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^1; S) - \mathrm{W}^2 + \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}^2; S)\|_F^2}{\sqrt{m}}.\end{aligned}$$

We get the desired result by simplifying the above inequality. □

Theorem C.4. (Restatement of Theorem 3.5) For any $\eta \leq \min\{\mu, \frac{\sqrt{m}}{8HC_x^2}\} \leq \frac{\sqrt{m}}{2HC_x^2}$, let W_T and $\mathrm{W}_T^{(i)}$ be the weight matrices returned after running Algorithm 1 with Moreau Envelope using S and $S^{(i)}$ respectively for T iterations. Here S and $S^{(i)}$ only differ in the i-th data. We have

$$\|\mathrm{W}_T - \mathrm{W}_T^{(i)}\|_F^2 \leq e^{1+\frac{8HC_x^2\eta T}{\sqrt{m}}} (T+1)^2 \left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}} \right)^2.$$

Proof of Theorem C.4.

$$\begin{aligned}&\|\mathrm{W}_{t+1} - \mathrm{W}_{t+1}^{(i)}\|_F^2 \\ =& \|\mathrm{W}_t - \eta\frac{1}{\mu}(\mathrm{W}_t - \tilde{\mathrm{U}}^{\mu}(\mathrm{W}_t; S)) - \mathrm{W}_t^{(i)} + \eta\frac{1}{\mu}(\mathrm{W}_t^{(i)} - \tilde{\mathrm{U}}^{\mu}(\mathrm{W}_t^{(i)}; S^{(i)}))\|_F^2 \\ \leq& \left(\|\mathrm{W}_t - \eta\frac{1}{\mu}(\mathrm{W}_t - \mathrm{U}^{\mu}(\mathrm{W}_t; S)) - \mathrm{W}_t^{(i)} + \eta\frac{1}{\mu}(\mathrm{W}_t^{(i)} - \mathrm{U}^{\mu}(\mathrm{W}_t^{(i)}; S^{(i)}))\|_F + \frac{2\eta}{\mu}\sqrt{\frac{2\beta_2}{\frac{1}{\mu} - \frac{HC_x^2}{\sqrt{m}}}} \right)^2 \\ & \hfill \text{(Lemma C.2)} \\ \leq& \Bigg(\|\mathrm{W}_t - \eta\frac{1}{\mu}(\mathrm{W}_t - \mathrm{U}^{\mu}(\mathrm{W}_t; S)) - \mathrm{W}_t^{(i)} + \eta\frac{1}{\mu}(\mathrm{W}_t^{(i)} - \mathrm{U}^{\mu}(\mathrm{W}_t^{(i)}; S))\|_F + \frac{2C_x\eta}{\mu n\left(\frac{1}{\mu} - \frac{HC_x^2}{\sqrt{m}}\right)} \\ &\quad + \frac{2\eta}{\mu}\sqrt{\frac{2\beta_2}{\frac{1}{\mu} - \frac{HC_x^2}{\sqrt{m}}}} \Bigg)^2 \qquad \text{(Lemma C.2)} \\ \leq& \left(\|\mathrm{W}_t - \eta\frac{1}{\mu}(\mathrm{W}_t - \mathrm{U}^{\mu}(\mathrm{W}_t; S)) - \mathrm{W}_t^{(i)} + \eta\frac{1}{\mu}(\mathrm{W}_t^{(i)} - \mathrm{U}^{\mu}(\mathrm{W}_t^{(i)}; S))\|_F + \frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}} \right)^2 \\ \leq& \frac{T+2}{T+1}\|\mathrm{W}_t - \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}_t; S) - \mathrm{W}_t^{(i)} + \eta\nabla_{\mathrm{W}} M^{\mu}(\mathrm{W}_t^{(i)}; S)\|_F^2 + (T+2)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}} \right)^2 \\ & \hfill ((a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2 \text{ for } p > 0.) \\ \leq& \frac{T+2}{T+1} \cdot \frac{1}{1 - \frac{4HC_x^2\eta}{\sqrt{m}}}\|\mathrm{W}_t - \mathrm{W}_t^{(i)}\|_F^2 + (T+2)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}} \right)^2. \qquad \text{(Lemma C.3)}\end{aligned}$$

Define $\gamma = \frac{T+2}{T+1} \cdot \frac{1}{1-\frac{4HC_x^2\eta}{\sqrt{m}}}$, then the inequality above can be written as

$$\|\mathbf{W}_{t+1} - \mathbf{W}_{t+1}^{(i)}\|_F^2 \leq \gamma \|\mathbf{W}_t - \mathbf{W}_t^{(i)}\|_F^2 + (T+2)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2.$$

Dividing both sides by γ^{t+1} and summing up the inequality, we get

$$\begin{aligned}
\|\mathbf{W}_T - \mathbf{W}_T^{(i)}\|_F^2 &\leq \frac{\gamma^T}{\gamma - 1}(T+2)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2 \\
&\leq \gamma^T (T+1)(T+2)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2 \\
&= \left(\frac{T+2}{T+1}\right)^{T+1}\left(\frac{1}{1-\frac{4HC_x^2\eta}{\sqrt{m}}}\right)^T (T+1)^2\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2 \\
&\leq e^{1+\frac{\frac{4HC_x^2\eta}{\sqrt{m}}T}{1-\frac{4HC_x^2\eta}{\sqrt{m}}}}(T+1)^2\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2 && (\tfrac{1}{1-x} \leq e^{\frac{x}{1-x}}) \\
&\leq e^{1+\frac{8HC_x^2\eta T}{\sqrt{m}}}(T+1)^2\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)^2. && (\eta \leq \tfrac{\sqrt{m}}{8HC_x^2})
\end{aligned}$$

□

The proof of Theorem 3.5 is obvious from Theorem C.4 by observing $e^{1+\frac{8HC_x^2\eta T}{\sqrt{m}}} \leq e^9$.

We have the following corollary by combining Theorem C.4, Lemma C.1-6 and Proposition A.3.

Corollary C.5. Assume the width of the networks satisfies $m \geq H^2C_x^4\eta^2T^2$. After T iterations of Algorithm 1 with Moreau Envelope with $\eta \leq \min\{\mu, \frac{\sqrt{m}}{8HC_x^2}\} \leq \frac{\sqrt{m}}{2HC_x^2}$, we have

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathbf{W}_T) \leq \mathbb{E}_{S\sim\mathcal{D}^n} \frac{M^\mu(\mathbf{W}_T; S) + C_x^2\mu}{1 - e^{4.5}C_x(T+1)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right)}$$

and

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{rob}(\mathbf{W}_T) \leq \mathbb{E}_{S\sim\mathcal{D}^n} M^\mu(\mathbf{W}_T; S) + C_x^2\mu + e^{4.5}C_x(T+1)\left(\frac{4C_x\eta}{n} + 4\eta\sqrt{\frac{\beta_2}{\mu}}\right).$$

Now we derive an optimization guarantee of optimizing the Moreau envelope.

Theorem C.6. Assume the width of the networks satisfies $m \geq H^2C_x^4\eta^2T^2$. After running Algorithm 1 with Moreau Envelope for T iterations with $\eta \leq \mu \leq \frac{\sqrt{m}}{2HC_x^2}$, we have

$$\min_{1\leq t\leq T} M^\mu(\mathbf{W}_t; S) \leq \min_{\mathbf{W}}\left(\widehat{L}_{rob}(\mathbf{W}; S) + \frac{2}{\eta T}\|\mathbf{W} - \mathbf{W}_0\|_F^2 + 2\eta(T+1)\frac{\beta_2}{\mu}\right).$$

Proof of Theorem C.6. From Lemma C.1, $M^\mu(\mathbf{W}; S)$ is $\frac{1}{\mu}$ smooth, so

$$\begin{aligned}
&M^\mu(\mathbf{W}_{t+1}; S) \\
\leq &M^\mu(\mathbf{W}_t; S) + \langle \mathbf{W}_{t+1} - \mathbf{W}_t, \nabla_{\mathbf{W}} M^\mu(\mathbf{W}_t; S)\rangle + \frac{1}{2\mu}\|\mathbf{W}_{t+1} - \mathbf{W}_t\|_F^2. \qquad (13)
\end{aligned}$$

From the weakly convex property Lemma C.1-3,

$$
\begin{aligned}
&M^\mu(\mathrm{W};S)\\
&\geq M^\mu(\mathrm{W}_t;S)+\langle\nabla_{\mathrm{W}}M^\mu(\mathrm{W}_t;S),\mathrm{W}-\mathrm{W}_t\rangle-\frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&=M^\mu(\mathrm{W}_t;S)+\langle\nabla_{\mathrm{W}}M^\mu(\mathrm{W}_t;S),\mathrm{W}-\mathrm{W}_{t+1}\rangle+\langle\mathrm{W}_{t+1}-\mathrm{W}_t,\nabla_{\mathrm{W}}M^\mu(\mathrm{W}_t;S)\rangle-\frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&\geq M^\mu(\mathrm{W}_{t+1};S)+\langle\nabla_{\mathrm{W}}M^\mu(\mathrm{W}_t;S),\mathrm{W}-\mathrm{W}_{t+1}\rangle-\frac{1}{2\mu}\|\mathrm{W}_{t+1}-\mathrm{W}_t\|_F^2-\frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(equation (13))}\\
&\geq M^\mu(\mathrm{W}_{t+1};S)+\frac{1}{\eta}\langle\mathrm{W}_{t+1}-\mathrm{W}_t,\mathrm{W}_{t+1}-\mathrm{W}\rangle-\frac{1}{2\eta}\|\mathrm{W}_{t+1}-\mathrm{W}_t\|_F^2-\frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(\mu\geq\eta)\\
&\quad-\frac{1}{\mu}\|\mathrm{U}^\mu(\mathrm{W}_t;S)-\tilde{\mathrm{U}}^\mu(\mathrm{W}_t;S)\|_F\cdot\|\mathrm{W}_{t+1}-\mathrm{W}\|_F\qquad\qquad\text{(Lemma C.1-2.)}\\
&\geq M^\mu(\mathrm{W}_{t+1};S)+\frac{1}{\eta}\langle\mathrm{W}_{t+1}-\mathrm{W}_t,\mathrm{W}_{t+1}-\mathrm{W}\rangle-\frac{1}{2\eta}\|\mathrm{W}_{t+1}-\mathrm{W}_t\|_F^2-\frac{HC_x^2}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&\quad-2\sqrt{\frac{\beta_2}{\mu}}\cdot\|\mathrm{W}_{t+1}-\mathrm{W}\|_F.\qquad\qquad\text{(Lemma C.2)}
\end{aligned}
$$

From the inequality above, for any weight matrix W,

$$
\begin{aligned}
&\|\mathrm{W}_{t+1}-\mathrm{W}\|_F^2\\
=&\|\mathrm{W}_t-\mathrm{W}\|_F^2-\|\mathrm{W}_{t+1}-\mathrm{W}_t\|_F^2+2\langle\mathrm{W}_{t+1}-\mathrm{W}_t,\mathrm{W}_{t+1}-\mathrm{W}\rangle\\
\leq&\|\mathrm{W}_t-\mathrm{W}\|_F^2+\left(2\eta M^\mu(\mathrm{W};S)-2\eta M^\mu(\mathrm{W}_{t+1};S)+\frac{2HC_x^2\eta}{\sqrt{m}}\|\mathrm{W}_t-\mathrm{W}\|_F^2+4\eta\sqrt{\frac{\beta_2}{\mu}}\cdot\|\mathrm{W}_{t+1}-\mathrm{W}\|_F\right).
\end{aligned}
$$

Simplifying the above inequality and combining with Lemma C.1-6 gives us

$$
\begin{aligned}
\|\mathrm{W}_{t+1}-\mathrm{W}\|_F^2&\leq\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)\|\mathrm{W}_t-\mathrm{W}\|_F^2+4\eta\sqrt{\frac{\beta_2}{\mu}}\|\mathrm{W}_{t+1}-\mathrm{W}\|_F\\
&\quad+2\eta\widehat{L}_{rob}(\mathrm{W};S)-2\eta M^\mu(\mathrm{W}_{t+1};S)\\
&\leq\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)\|\mathrm{W}_t-\mathrm{W}\|_F^2+\frac{1}{T+1}\|\mathrm{W}_{t+1}-\mathrm{W}\|_F^2+4\eta^2(T+1)\frac{\beta_2}{\mu}\\
&\quad+2\eta\widehat{L}_{rob}(\mathrm{W};S)-2\eta M^\mu(\mathrm{W}_{t+1};S).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\|\mathrm{W}_{t+1}-\mathrm{W}\|_F^2&\leq\left(1+\frac{1}{T}\right)\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)\|\mathrm{W}_t-\mathrm{W}\|_F^2\\
&\quad+2\eta\left(1+\frac{1}{T}\right)\left(\widehat{L}_{rob}(\mathrm{W};S)-M^\mu(\mathrm{W}_{t+1};S)+2\eta(T+1)\frac{\beta_2}{\mu}\right).\qquad(14)
\end{aligned}
$$

Dividing both sides by $\left(1+\frac{1}{T}\right)^{t+1}\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)^{t+1}$ and summing up the inequality for $t=0,1,\ldots,T-1$, we get

$$
\begin{aligned}
\min_{1\leq t\leq T}M^\mu(\mathrm{W}_t;S)&\leq\widehat{L}_{rob}(\mathrm{W};S)+2\eta(T+1)\frac{\beta_2}{\mu}+\frac{\|\mathrm{W}-\mathrm{W}_0\|_F^2}{2\eta\left(1+\frac{1}{T}\right)\sum_{t=1}^{T}\frac{1}{\left(1+\frac{1}{T}\right)^t\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)^t}}\\
&\leq\widehat{L}_{rob}(\mathrm{W};S)+\frac{2}{\eta T}\|\mathrm{W}-\mathrm{W}_0\|_F^2+2\eta(T+1)\frac{\beta_2}{\mu},
\end{aligned}
$$

where in the last inequality we use $\left(1+\frac{1}{T}\right)\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)\leq\left(1+\frac{1}{T}\right)\left(1+\frac{2}{T}\right)\leq 1+3\frac{T+1}{T^2}$. □

Theorem 3.6. Define $\alpha_2(\eta,T) := \mathcal{O}(C_x^2\frac{\eta T}{n} + C_x\eta T\sqrt{\frac{\beta_2}{\mu}})$. Assume $\alpha_2(\eta,T) < 1$. Then, after $T \geq 8$ iterations of Algorithm Moreau Envelope with step-size $\eta \leq \mu$ on a network of width $m \geq H^2C_x^4\eta^2T^2$, we have

$$\min_{[\frac{9T}{10}]\leq t\leq T} \mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathrm{W}_t) \leq \frac{55\alpha_2(\eta,T)}{1-\alpha_2(\eta,T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + C_x^2\mu + 2\eta(T+1)\frac{\beta_2}{\mu}\right],$$

and

$$\min_{1\leq t\leq T} \mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathrm{W}_t) \leq \frac{1}{1-\alpha_2(\eta,T)}\left[\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + C_x^2\mu + 2\eta(T+1)\frac{\beta_2}{\mu}\right].$$

Proof of Theorem 3.6. Define $t_0 := [\frac{9T}{10}]$ and $k := \frac{1}{\left(1+\frac{1}{T}\right)\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)}$. Dividing both sides of equation (14) by $\left(1+\frac{1}{T}\right)^{t+1}\left(1+\frac{2HC_x^2\eta}{\sqrt{m}}\right)^{t+1}$ and summing up, we have

$$\begin{aligned}
&2\eta\left(1+\frac{1}{T}\right)\sum_{t=1}^{T}k^t\left(\widehat{L}_{rob}(\mathrm{W};S) + 2\eta(T+1)\frac{\beta_2}{\mu}\right) + \|\mathrm{W}-\mathrm{W}_0\|_F^2 \\
\geq& 2\eta\left(1+\frac{1}{T}\right)\sum_{t=t_0}^{T}k^t M^{\mu}(\mathrm{W}_t;S).
\end{aligned}$$

Taking minimum over all weight matrices W,

$$\begin{aligned}
&2\eta\left(1+\frac{1}{T}\right)\sum_{t=t_0}^{T}k^t\left(\widehat{L}_{rob}(\mathrm{W}_t;S) - C_x^2\mu\right) \\
\leq& 2\eta\left(1+\frac{1}{T}\right)\sum_{t=t_0}^{T}k^t M^{\mu}(\mathrm{W}_t;S) && \text{(Lemma C.1-6.)} \\
\leq& \min_{\mathrm{W}}\left(2\eta\left(1+\frac{1}{T}\right)\sum_{t=1}^{T}k^t\left(\widehat{L}_{rob}(\mathrm{W};S) + 2\eta(T+1)\frac{\beta_2}{\mu}\right) + \|\mathrm{W}-\mathrm{W}_0\|_F^2\right) \\
\leq& 2\eta\left(1+\frac{1}{T}\right)\sum_{t=1}^{T}k^t\cdot\min_{\mathrm{W}}\left(\widehat{L}_{rob}(\mathrm{W};S) + \frac{2}{\eta T}\|\mathrm{W}-\mathrm{W}_0\|_F^2 + 2\eta(T+1)\frac{\beta_2}{\mu}\right) \\
=& 2\eta\left(1+\frac{1}{T}\right)\sum_{t=1}^{T}k^t\cdot\left(\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right).
\end{aligned}$$

Taking the expectation on both sides, we get

$$\begin{aligned}
\sum_{t=t_0}^{T}k^t\left(\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S) - C_x^2\mu\right) &\leq \sum_{t=t_0}^{T}k^t\left(\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S) - C_x^2\mu\right) \\
&\leq \sum_{t=1}^{T}k^t\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right).
\end{aligned}$$

Simplifying the equation above, we get

$$\begin{aligned}
\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathrm{W}_t;S) &\leq C_x^2\mu + \frac{\sum_{t=1}^{T}k^t}{\sum_{t=t_0}^{T}k^t}\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right) \qquad (15) \\
&\leq C_x^2\mu + \left(\sum_{r=0}^{9}\left(\frac{1}{k}\right)^{r\frac{T}{10}}\right)\cdot\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right)
\end{aligned}$$

$$\leq C_x^2\mu + \left(\sum_{r=0}^{9} e^{\frac{3r}{10}}\right) \cdot \left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right)$$
$$(m \geq H^2C_x^4\eta^2T^2)$$
$$\leq C_x^2\mu + 55\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right).$$

Equation (4) gives us for any $t \leq T$, $\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathbf{W}_t) \leq \frac{\alpha_2(\eta,T)}{1-\alpha_2(\eta,T)}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathbf{W}_t;S)$. Therefore,

$$\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\varepsilon_{gen}(\mathbf{W}_t) \leq \frac{\alpha_2(\eta,T)}{1-\alpha_2(\eta,T)}\min_{t_0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathbf{W}_t;S)$$
$$\leq \frac{\alpha_2(\eta,T)}{1-\alpha_2(\eta,T)}\left(C_x^2\mu + 55\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + 2\eta(T+1)\frac{\beta_2}{\mu}\right)\right).$$

The proof of the second statement takes a similar approach. Following the same procedure, we can replace t_0 by 0 in equation (15), and get

$$\min_{0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathbf{W}_t;S) \leq \mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + C_x^2\mu + 2\eta(T+1)\frac{\beta_2}{\mu}.$$

Combining with equation (4),

$$\min_{0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathbf{W}_t) \leq \frac{1}{1-\alpha_2(\eta,T)}\min_{0\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}\widehat{L}_{rob}(\mathbf{W}_t;S)$$
$$\leq \frac{1}{1-\alpha_2(\eta,T)}\left(\mathbb{E}_{S\sim\mathcal{D}^n}\Delta_S^{\text{oracle}} + C_x^2\mu + 2\eta(T+1)\frac{\beta_2}{\mu}\right).$$

□

Corollary 3.7. After $T \leq \mathcal{O}(\min\{n^2, \frac{1}{\beta_2^{2/3}}\})$ iterations of Algorithm 1 with Moreau Envelope with step-size $\eta = \mu = \Theta(\frac{1}{C_x^2\sqrt{T}})$ on a network of width $m \geq \Omega(T)$, we have for any weight matrix W,

$$\min_{1\leq t\leq T}\mathbb{E}_{S\sim\mathcal{D}^n}L_{rob}(\mathbf{W}_t) \leq 1.1L_{rob}(\mathbf{W}) + \mathcal{O}\left(\frac{C_x^2\|\mathbf{W}-\mathbf{W}_0\|_F^2}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Proof of Corollary 3.7. Under the conditions of the corollary, we have $m \geq H^2C_x^4\eta^2T^2$, and $\alpha_2(\eta,T) = \mathcal{O}(C_x^2\frac{\eta T}{n} + C_x\eta T\sqrt{\frac{\beta_2}{\mu}})$ can be small enough so that $\frac{1}{1-\alpha_2(\eta,T)} \leq 1.1$. Then it is clear that this corollary is a special case of Theorem 3.6. □

NeurIPS Paper Checklist

A. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We further expand on the claims made in Abstract and Introduction in Section 3.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

B. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discussed the limitations in the Conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

C. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provide the full set of assumptions and a complete and correct proof for each theoretical result. Please see Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

D. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

E. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

F. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

G. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

H. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

I. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics in every respect. The theoretical nature of the results means there are minimal ethical concerns.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

J. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The societal impacts of the paper is overall positive.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

K. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

L. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

M. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

N. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

O. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.