IMPLICIT BIAS OF GRADIENT DESCENT BASED ADVERSARIAL TRAINING ON SEPARABLE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training is a principled approach for training robust neural networks. Despite of tremendous successes in practice, its theoretical properties still remain largely unexplored. In this paper, we provide new theoretical insights of gradient descent based adversarial training by studying its computational properties, specifically on its implicit bias. We take the binary classification task on linearly separable data as an illustrative example, where the loss asymptotically attains its infimum as the parameter diverges to infinity along certain directions. Specifically, we show that for any fixed iteration T, when the adversarial perturbation during training has proper bounded ℓ_2-norm, the classifier learned by gradient descent based adversarial training converges in direction to the maximum ℓ_2-norm margin classifier at the rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$, significantly faster than the rate $\mathcal{O}\left(1/\log T\right)$ of training with clean data. In addition, when the adversarial perturbation during training has bounded ℓ_q-norm with $q \geq 1$, the resulting classifier converges in direction to a maximum mixed-norm margin classifier, which has a natural interpretation of robustness, as being the maximum ℓ_2-norm margin classifier under worst-case ℓ_q-norm perturbation to the data. Our findings provide theoretical backups for adversarial training that it indeed promotes robustness against adversarial perturbation.

1 INTRODUCTION

Deep neural networks have achieved remarkable success on various tasks, including visual and speech recognitions, with intriguing generalization abilities to unseen data (Krizhevsky et al., 2012; Hinton et al., 2012). One salient feature of deep models is its overparameterization, with the number of parameters several orders of magnitude larger than the training sample size. As a consequence of such overparameterization, it is likely that the empirical loss function, in addition to being non-convex, can have substantial amount of global minimizers (Choromanska et al., 2015), while only a small subset of global minimizers have the desired generalization properties (Brutzkus et al., 2018).

Contrary to the worst-case reasoning above, researchers have observed that simple first-order algorithm such as Stochastic Gradient Descent (SGD) [1], performs surprisingly well in practice, even without any explicit regularization terms in the objective function (Zhang et al., 2017). Inspired by classical computational learning theories, one plausible explanation of such a remarkable phenomenon is that the training algorithm enjoys some implicit bias. That is, the training algorithm tends to converge to certain kinds of solutions (Neyshabur et al., 2015b;c), and SGD converges to low-capacity solutions with the desired generalization property (Brutzkus et al., 2018). Recently, some exciting works have related the implicit bias to specific first-order algorithms (Wilson et al., 2017), stopping time (Hoffer et al., 2017), and optimization geometry (Gunasekar et al., 2018a; Keskar et al., 2017). Some practical suggestions based on these findings have also been proposed to further improve the generalization ability of deep networks (Neyshabur et al., 2015a).

Despite the aforementioned phenomenal success achieved by deep neural networks, it is observed that adversarially constructed small perturbation to the input can potentially fool the network into making wrong predictions with high confidence (Szegedy et al., 2014; Goodfellow et al., 2015). This issue raises serious concerns about using neural network for some security-sensitive tasks (Papernot et al., 2017). Researchers have devised various mechanisms to generate and defend against adversarial perturbations (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017;

[1] In conjunction with Dropout (Srivastava et al., 2014) and Batch Normalization (Ioffe and Szegedy, 2015)

Athalye et al., 2018; Xie et al., 2018; Papernot et al., 2016). However, most of the defense mechanisms are heuristic or ad-hoc, which lack principled theoretical justification (Carlini and Wagner, 2016; He et al., 2017). Inspired by literatures in robust optimization (Wald, 1939; Ben-Tal et al., 2009), Feige et al. (2015); Madry et al. (2018) formalize the notion of achieving adversarial robustness (i.e., having small adversarial risk) as solving the following minimax optimization problem

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}^{\mathrm{E}}_{\mathrm{adv}}(\theta) = \min_{\theta\in\mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_{\delta\in\Delta} \ell(\theta, x+\delta, y)\Big], \tag{1}$$

where Δ is the set that each sample could be contaminated by arbitrary perturbation chosen within this set. As a common practice, adversarial training refers to the finite-sample empirical version of (1) without access to the underlying distribution $\mathcal{D}$ that

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}_{\mathrm{adv}}(\theta) = \min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{N} \max_{\delta_i\in\Delta} \ell(\theta, x_i+\delta, y_i). \tag{2}$$

A commonly adopted approach to solving (2) is the the Gradient Descent based Adversarial Training (GDAT) method. At each iteration, GDAT first solves the inner maximization problem (approximately) for adversarial perturbations, and then uses the gradient of the loss function evaluated at the perturbed samples to perform a gradient descent step on the parameter θ. A natural question is then how adversarial training helps the trained model in achieving adversarial robustness. Some recent theoretical results partially answer this question, such as deriving adversarial risk bound (Athalye et al., 2018), relating it to the distributionally robust optimization (Sinha et al., 2018), and characterizing trade-offs between robustness and accuracy via regularization (Zhang et al., 2019).

Yet, all existing results neglect the algorithmic effect during the training process in promoting adversarial robustness. Inspired by the significant role of algorithmic bias in the generalization of neural networks, it is natural to ask

Does gradient descent based adversarial training enjoy any implicit bias property?
If so, does the implicit bias provide insights on how adversarial training promotes robustness?

Motivated by these questions, in this paper, we study the algorithmic effect of adversarial training by investigating the implicit bias of GDAT. Due to current technical limits in directly analyzing deep neural networks, we analyze a simpler model, with the key characteristics that the model overfits the training data while being able to generalize well. Specifically, we take the binary classification with linearly separable data as an example. This helps us focus on the effect of implicit bias without dealing with complicated structures of neural networks.

Main Contributions. We summarize our main theoretical findings below.

• Our first part of result shows an interesting interplay between adversarial perturbation and implicit bias of the gradient descent (GD). By exploiting this interplay, we show a property of adversarial training that is not known in the literature before: adversarial training accelerates convergence. Specifically, when the perturbation is bounded by ℓ_2-norm, i.e., $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_2 \le c\}$, with proper choice of c, the gradient descent based adversarial training is directionally convergent that $\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_2$, where u_2 is the maximum ℓ_2-norm margin hyperplane (i.e., standard SVM) of the training data. In addition, when the perturbation level c is set according to T appropriately, the rate of convergence is $\widetilde{\mathcal{O}}(1/\sqrt{T})$[2], which is exponentially faster than the rate $\mathcal{O}\left(1/\log T\right)$ when we use standard clean training, i.e., training with clean data using gradient descent. Based on this, we establish that the convergence of training loss on clean data using GDAT is almost exponentially faster than standard clean training using GD.

• Our second part of result shows that adversarial training adapts the implicit bias of gradient descent for different adversarial perturbation geometry. Specifically, when the perturbation is bounded by ℓ_q-norm for $q \ge 1$, i.e., $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_q \le c\}$, with proper choice of c, the gradient descent based adversarial training is directionally convergent that $\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_{2,q}$, where $u_{2,q}$ is the maximum mixed-norm margin hyperplane of the training data. We further reveal natural interpretation of robustness that we obtain the maximum ℓ_2-norm margin classifier under worst-case ℓ_q-norm perturbation.

Notations. For two vectors $x, y \in \mathbb{R}^d$, $\langle x, y\rangle = \sum_{j=1}^d x_j y_j$ denotes their Euclidean inner product. For a vector $\theta \in \mathbb{R}^d$, $||\theta||_p$ defined by $||\theta||_p^p = \sum_{j=1}^d |\theta_j|^p$ denotes its p-norm for $p \in [1, \infty)$, and

[2] $\widetilde{\mathcal{O}}$ hides logarithmic factor.

$||\theta||_\infty = \max_{j\in[d]} |\theta_j|$, where $[d] = \{1, \ldots, d\}$. For any general norm $||\cdot||$, we denote its dual norm by $||x||_* = \max_{||y||\leq 1} \langle x, y\rangle$. The sign function is $\mathrm{sign}(v) = \mathbb{1}_{(v\geq 0)} - \mathbb{1}_{(v<0)}$. For a linear subspace $L \in \mathbb{R}^d$, we denote its orthogonal subspace by $L^\perp$.

2 BACKGROUND

We consider a binary classification problem using a dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{-1, +1\}$. We aim to learn a linear decision boundary $f(x) = \langle \theta, x\rangle$ and its associated classifier $\widehat{y}(x) = \mathrm{sign}\,(f(x))$, by solving the empirical risk minimization problem:

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta; \mathcal{S}) = \min_{\theta\in\mathbb{R}^d} \sum_{i=1}^n \ell(y_i x_i^\top \theta), \text{ where } \ell(\cdot) \text{ is some loss function.} \tag{3}$$

In what follows, we suppress the explicit presentation of $\mathcal{S}$ when the context is clear, and we focus on the exponential loss $\ell(r) = \exp(-r)$. We point out that our analysis can be further extended to other smooth loss functions with tight exponential tail such as logistic loss.

We assume the dataset $\mathcal{S}$ is linearly separable, i.e., there exists $\overline{u}$ such that $\min_{i\in[n]} y_i x_i^\top \overline{u} > 0$. Under this assumption, one notable feature of problem (3) is that there is no finite minimizer, and $\mathcal{L}(\theta) \to 0$ only if $||\theta||_2 \to \infty$ along certain directions. In fact, there is a polyhedral cone $\mathcal{C}$, such that for any $u \in \mathcal{C}$, we have $\lim_{a\to\infty} \mathcal{L}(a\overline{u}) = 0$.

Several recent results have studied the implicit bias of gradient descent algorithm on separable dataset. Soudry et al. (2018) study the implicit bias of the gradient descent algorithm (GD) on (3), and show that $\lim_{t\to\infty} ||\theta^t||_2 = \infty$, while θ^t converges in direction to the maximum ℓ_2-norm margin classifier (i.e., the standard SVM). Ji and Telgarsky (2018) further study the convergence of risk and parameter without separability condition. (Ji and Telgarsky, 2019) and (Gunasekar et al., 2018b) study the implicit bias for training deep linear network and linear convolutional networks, respectively. Gunasekar et al. (2018a) also analyze the implicit bias of steepest descent in general norm $||\cdot||$, and show that θ^t converges in direction to the maximum $||\cdot||_*$-norm margin hyperplane.

Throughout this paper, we assume the perturbation set is an ℓ_q-norm ball with radius c, i.e., $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_q \leqslant c\}$. Under the general framework of adversarial training in (2), we aim to minimize the empirical adversarial risk

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}_{\mathrm{adv}}(\theta) = \min_{\theta\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \max_{\delta_i\in\Delta} \exp\left(-y_i(x_i + \delta_i)^\top \theta\right). \tag{4}$$

Note that, given any θ, the inner maximization problem in (4) admits a closed form solution. Then the gradient descent based adversarial training (GDAT) algorithm runs iteratively that at the t-th iteration, we first solve the inner maximization problem by deriving the worst adversarial perturbation of each sample. It is not difficult to see that for each sample, the worst perturbation is $\delta_i^t = cy_i\delta_t^*$, where $\delta_t^* = \mathrm{argmin}_{\delta:||\delta||_q\leq 1} \langle \delta, \theta^t\rangle$. Then, letting each sample's perturbed counterpart be $(\widetilde{x}_i^t, y_i) = (x_i + \delta_i^t, y_i)$, we take gradient of the loss function evaluated at the perturbed samples and perform a gradient descent step, i.e., $\theta^{t+1} = \theta^t - \eta^t \nabla_\theta \mathcal{L}\left(\theta^t; \{(\widetilde{x}_i^t, y_i)\}_{i=1}^n\right))$, where $\eta^t > 0$ is some prespecified stepsize. We present the outline of GDAT in Algorithm 1.

Algorithm 1 Gradient Descent based Adversarial Training (GDAT) with ℓ_q-norm Perturbation

Input: Number of iterations T, perturbation level c, stepsizes $\{\eta^t\}_{t=0}^T$, samples $\{x_i, y_i\}_{i=1}^n$.
Initialize: $\theta^0 \leftarrow 0$.
for $t = 0, \ldots, T-1$ **do**
  for $i = 1, \ldots, n$ **do**
    Compute $\delta_i^t = cy_i\,\mathrm{argmin}_{||\delta||_q\leq 1} \langle \delta, \theta^t\rangle$
    Let $(\widetilde{x}_i^t, y_i) \leftarrow (x_i + \delta_i^t, y_i)$.
  end for
  $\theta^{t+1} \leftarrow \theta^t - \frac{\eta^t}{n} \sum_{i=1}^n \exp\left(-y_i \widetilde{x}_i^\top \theta^t\right)(-y_i\widetilde{x}_i)$.
end for

3 THEORETICAL RESULTS

In this section, we show that the GDAT algorithm possesses implicit bias, which depends on the perturbation set during training. We provide explicit characterization of the implicit bias, and further conclude that such implicit bias indeed promotes robustness against adversarial perturbation.

Let us start with some definitions. Consider a dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{-1, +1\}$. Given $p, q > 0$ such that $1/p + 1/q = 1$, the ℓ_q-norm margin of H_θ on $\mathcal{S}$ is defined as $\gamma_q(\theta) = \min_{i\in[n]} y_i x_i^\top \theta / ||\theta||_p$. Note that for $x_i \in \mathbb{R}^d$, $|\theta^\top x|/||\theta||_p$ measures the ℓ_q distance between x_i and the hyperplane $H_\theta = \{x \in \mathbb{R}^d : \theta^\top x = 0\}$. Since $y_i \in \{-1, +1\}$, when H_θ correctly classifies all samples, $\gamma_q(\theta)$ measures the minimal ℓ_q distance between the samples in $\mathcal{S}$ and H_θ. Given that $\gamma_q(\theta)$ is scale-invariant with respect to θ, without loss of generality, we restrict $||\theta||_p = 1$. We also identify the hyperplane H_θ by its normal vector θ.

Definition 3.1. *For $p, q > 0$ with $1/p + 1/q = 1$, the maximum ℓ_q-norm margin hyperplane u_q of $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{-1, +1\}$ and its associated ℓ_q-norm margin γ_q are defined as*

$$u_q \in \operatorname*{argmax}_{||\theta||_p=1} \min_{i\in[n]} y_i x_i^\top \theta, \quad \gamma_q = \max_{||\theta||_p=1} \min_{i\in[n]} y_i x_i^\top \theta. \tag{5}$$

We denote $\mathrm{SV}(\mathcal{S})$ as the support vectors of $\mathcal{S}$, i.e., $\mathrm{SV}(\mathcal{S}) = \operatorname{argmin}_{(x,y)\in\mathcal{S}} \langle u_q, yx \rangle$.

By the separability assumption, u_q is an optimal hyperplane that correctly classifies all samples with the maximal margin $\gamma_q > 0$. Next, by the notion of margin defined above, we characterize the landscape of empirical adversarial risk in (4) based on the perturbation level c.

Proposition 3.1. *Let $p, q > 0$ satisfy $1/p + 1/q = 1$. Given a nonnegative scalar c, where $0 \le c < \gamma_q = \max_{||\theta||_p \le 1} \min_{i\in[n]} y_i x_i^\top \theta$, problem (4) has infimum 0 but does not admit a finite minimizer. When $c > \gamma_q$, problem (4) has a unique finite minimizer $\widehat{\theta}(c)$, and is equivalent to the standard clean training with explicit ℓ_p-norm regularization. That is, there exists $\lambda(c) > 0$ such that*

$$\widehat{\theta}(c) = \operatorname*{argmax}_{\theta\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \exp(-y_i x_i^\top \theta) + \lambda(c)||\theta||_p.$$

It is not difficult to see that for $c < \gamma_q$, any perturbed dataset $\widetilde{\mathcal{S}} = \{(\widetilde{x}_i, y_i)\}_{i=1}^n$, with $||x_i - \widetilde{x}_i||_q \leqslant c$ for all i, is still linearly separable, which directly follows from the definition of γ_q above. On the other hand, when $c > \gamma_q$, by the definition of γ_q, there exists some perturbed dataset $\widetilde{\mathcal{S}} = \{(\widetilde{x}_i, y_i)\}_{i=1}^n$, with $||x_i - \widetilde{x}_i||_q \leqslant c$ for all i, such that $\widetilde{\mathcal{S}}$ is no longer linearly separable.

3.1 Adversarial Perturbation with Bounded ℓ_2-Norm

In this subsection, we analyze both the empirical adversarial risk convergence and the parameter convergence of the case when the perturbation set Δ in (4) is an ℓ_2-norm ball with radius c.

Adversarial Risk Convergence. We first analyze the convergence of empirical adversarial risk (4) using GDAT. One substantial roadblock of minimizing (4) is its non-smoothness, in the sense that $\mathcal{L}_{\mathrm{adv}}(\theta)$ is not differentiable at the origin, and its Hessian $\nabla^2 \mathcal{L}_{\mathrm{adv}}(\theta)$ explodes around the origin. To address the challenge, our key observation is that, by the next lemma, at each iteration, there exists an acute angle between the update on θ^t and the maximum ℓ_2-norm margin hyperplane u_2. This gives a lower bound on $||\theta^t||_2$.

Lemma 3.1. *Take $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_2 \le c\}$ in problem (4). Given $c < \gamma_2$, we have that $\langle -\nabla \mathcal{L}_{\mathrm{adv}}(\theta), u_2 \rangle \ge \mathcal{L}_{\mathrm{adv}}(\theta)(\gamma_2 - c) > 0$ for any $\theta \in \mathbb{R}^d$.*

We highlight that despite its simple proof, Lemma 3.1 and its generalization to ℓ_q-perturbation is a crucial step for analyzing both adversarial risk and implicit bias. In addition, our techniques here can also be adapted to simplify the proof of Lemma 10 in (Gunasekar et al., 2018a), which, in comparison, is more technically involved.

Since we initialize GDAT (Alg. 1) using $\theta^0 = 0$, any perturbation inside Δ will have no effect on the adversarial loss. Hence we take clean samples as adversarial examples at the first iteration of GDAT. From Lemma 3.1, we have the following simple corollary showing that our whole solution path $\{\theta^t\}_{t=1}^T$ is bounded away from the origin.

Corollary 3.1. *Let $\theta^0 = 0$ in Algorithm 1 with $q = 2$, we have: $||\theta^t||_2 \ge \eta^0 \gamma_2$ for all $t \ge 1$.*

By Corollary 3.1, we bypass the non-differentiability issue at the origin and also control the Hessian $\nabla^2 \mathcal{L}_{\mathrm{adv}}(\theta)$ throughout the entire training process. Similar to (Ji and Telgarsky, 2018), in the next theorem, we show that the loss $\mathcal{L}_{\mathrm{adv}}(\theta)$, although not uniformly smooth, is locally $\mathcal{L}_{\mathrm{adv}}(\theta)$-smooth. Consequently, by the smoothness based analysis of the gradient descent algorithm, we establish the convergence of the empirical adversarial risk.

Theorem 3.1. *Suppose $||x_i||_2 \leq 1$ for all $i = 1 \dots n$. For GDAT (Alg. 1) with ℓ_2-norm perturbation, i.e., $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_2 \leqslant c\}$, we set $c < \gamma_2$, $\eta^0 = 1$ and $\eta^t = \eta \leq \min\{\frac{\gamma_2/e}{(1+c)^3\gamma_2+2c(1+c)}, 1\}$ for $t \geq 1$, then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\max_{\delta_i\in\Delta}\exp\left(-y_i(x_i+\delta_i)^\top\theta^t\right) = \mathcal{O}\left(\frac{\log^2 t}{t\eta(\gamma_2-c)^2}\right). \tag{6}$$

In comparison with the standard clean training using GD (Ji and Telgarsky, 2018), this theorem states that we pay an extra $(\gamma_2 - c)^{-2}$ factor in the risk convergence of adversarial training. However, this direct comparison is too pessimistic since we compare the adversarial risk with the standard risk (corresponding to $\Delta = \{0\}$). Interestingly, as seen later in Corollary 3.2, we prove that the convergence of standard risk in GDAT is significantly faster than its counterpart in the standard clean training using GD.

Parameter Convergence. We then show that if we set the perturbation level $c < \gamma_2$ in the GDAT algorithm, GDAT with ℓ_2-norm perturbation possesses the same implicit bias as the standard clean training using GD, i.e., we have $\lim_{t\to\infty}\frac{\theta^t}{||\theta^t||_2} = u_2$. Intuitively, GDAT with ℓ_2-norm perturbation searches for a decision hyperplane that is robust to ℓ_2-norm perturbation. Since the learned decision hyperplane in the standard clean using GD converges to u_2, which is already the most robust decision hyperplane against ℓ_2-norm perturbation to the data, GDAT retains the implicit bias of standard clean training using GD.

Surprisingly, even though both GDAT in the adversarial training and GD in the standard clean training converge in directions to u_2, their rates of directional convergence are significantly different as shown later. Specifically, letting the perturbation level c depend on the total number of iterations T in the GDAT algorithm, the directional error after T iterations in GDAT algorithm can be significantly smaller than the error of GD in the standard clean training.

We first show that the projection of θ^t onto the orthogonal subspace of $\mathrm{span}(u_2)$ is bounded.

Lemma 3.2. *Define $\alpha(\mathcal{S}) = \min_{||\xi||_2=1,\xi\in\mathrm{span}(u_2)^\perp}\max_{(x,y)\in\mathrm{SV}(\mathcal{S})}\langle\xi, yx\rangle$, where we assume $\mathrm{SV}(\mathcal{S})$ spans $\mathbb{R}^d$. Let $\theta_\perp$ be the projection of vector θ onto $\mathrm{span}(u_2)^\perp$. Then there exists a constant K that only depends on $\alpha(\mathcal{S})$ and $\log n$, such that $||\theta_\perp^t||_2 \leq K$ for any $t \geq 0$ in the GDAT algorithm.*

Note that the same $\alpha(\mathcal{S})$ is defined in (Ji and Telgarsky, 2019) and proved to be positive with probability 1 if the data is sampled from absolutely continuous distribution. We then show in the next lemma that $||\theta^t||_2$ goes to infinity, where we provide a refined analysis to establish the acceleration of the directional convergence in comparison with the standard clean training.

Lemma 3.3. *Under the same conditions in Theorem 3.1, and let $\alpha = \alpha(\mathcal{S})$ defined in Lemma 3.2. Then for all $t \geq 0$, we have*

$$||\theta^t||_2 \geq \log\left(\frac{t\eta(\gamma_2-c)^2}{n^{1+1/\alpha}\log^2 t}\right)/(\gamma_2-c).$$

Lemma 3.3 provides the key insight to establish the acceleration of directional convergence. Specifically, it allows us to set c depending on the total number of iterations T, so that $||\theta^T||_2$ is sublinear in T, in comparison with being logarithmic in T in standard clean training as in Ji and Telgarsky (2018). We are now ready to present the main theorem for parameter convergence.

Theorem 3.2 (Speed-up of Parameter Convergence)**.** *Under same conditions in Theorem 3.1, and let $\alpha = \alpha(\mathcal{S})$ and K be defined in Lemma 3.2. In GDAT with ℓ_2-norm perturbation, let c and total number of iterations T satisfy $\gamma_2 - c = \left(\frac{n^{1+1/\alpha}\log T}{\eta T}\right)^{1/2}$, and define $\overline{\theta}^T = \frac{\theta^T}{||\theta^T||_2}$. We have*

$$1 - \left\langle\overline{\theta}^T, u_2\right\rangle = \mathcal{O}\left(\frac{n^{(1+1/\alpha)/2}K\log T}{\sqrt{\eta}\sqrt{T}}\right). \tag{7}$$

One might argue that the polynomial dependence on sample size n in (7) is too pessimistic, making the GDAT unfavorable in comparison with the standard clean training. We show that this is not an issue by a direct comparision of iteration complexity to achieve $||\overline{\theta}^T - u_2||_2 \leq \epsilon$ for a given precision $\epsilon > 0$. Specifically, given $\epsilon > 0$, to achieve $||\overline{\theta}^T - u_2||_2 \leq \epsilon$, GDAT needs $\widetilde{\mathcal{O}}\left(n^{(1+1/\alpha)}\epsilon^{-2}\right)$ number

of iterations. In comparison, the standard clean training by GD needs $\widetilde{\mathcal{O}}\left(n\exp\left(\epsilon^{-1}\right)\right)$ number of iterations (Ji and Telgarsky, 2018), which has exponential dependence on precision ϵ.

Finally, by Theorem 3.1 and Lemma 3.3, we show that the empirical clean risk after T iterations of GDAT is almost exponentially smaller than its counterpart in the standard clean training.

Corollary 3.2 (Speed-up of Clean Risk Convergence)**.** *Under the same conditions in Theorem 3.2, we have*

$$\mathcal{L}(\theta^T) = \mathcal{O}\left(\exp\left(-\mu\sqrt{T}/\log T\right)\right),$$

where μ is a constant dependent on η, α, n.

Note that the empirical clean risk decreases at the rate of $\mathcal{O}\left(\exp(-\sqrt{T})\right)$ up to a logarithmic factor in the exponent. In comparison, using standard clean training with GD, we only have $\mathcal{L}(\theta^T) = \mathcal{O}\left(1/T\right)$ (Soudry et al., 2018).

3.2 Adversarial Perturbation with Bounded ℓ_q-Norm

In this subsection, we generalize our results to the case where the perturbation set is some bounded ℓ_q-norm ball. To facilitate our discussion, we first define a robust version of SVM.

Definition 3.2. *For a given separable dataset $\mathcal{S}$ with ℓ_q-norm margin γ_q and $c < \gamma_q$, letting $1/p + 1/q = 1$, the robust SVM against ℓ_q-norm perturbation parameterized by c is*

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{2}||\theta||_2^2 \quad \textit{s.t.} \quad y_i x_i^\top \theta \geq c||\theta||_p + 1, \forall i = 1, \ldots, n. \tag{8}$$

Remark 3.1 (Maximum Mixed-norm Margin)**.** *Note that problem* (8) *is equivalent to solving for a maximum mixed-norm margin hyperplane. Specifically, by the KKT condition of* (8)*, there exists $\eta(c) > 0$, such that* (8) *is equivalent to the following problem:*

$$\min_{\theta\in\mathbb{R}^d} ||\theta||_2 + \eta(c)||\theta||_p \ \text{ s.t. } \ y_i x_i^\top \theta \geq 1, \forall i = 1, \ldots, n. \tag{9}$$

Now define $||\cdot|| = ||\cdot||_2 + \eta(c)||\cdot||_p$, it is clear that $||\cdot||$ defines a norm which is a mixture of ℓ_2 and ℓ_p norm. Let $||\cdot||_*$ be its dual norm. Then we have that the solution to (9) *is the maximum $||\cdot||_*$-norm margin hyperplane.*

Note that the constraint in (8) is equivalent to $\min_{||\delta_i||_q\leq c} y_i(x_i+\delta_i)^\top\theta \geq 1, \forall i = 1, \ldots, n$. By a simple scaling argument, in the following lemma, we see the robust nature of (8).

Lemma 3.4. *Under the same notations in Definition 3.2, problem* (8) *is equivalent to:*

$$\gamma_{2,q}(c) = \max_{||\theta||_2=1} \min_{i\in[n]} \min_{||\delta_i||_q\leqslant c} y_i(x_i+\delta_i)^\top\theta. \tag{10}$$

We denote the (unique) solution to problem (10) as $u_{2,q}(c)$. In what follows, we surpress explicit presentation of c when the context is clear.

The equivalent formulation (10) provides a clear interpretation on the robustness of (10). In particular, the robust SVM against ℓ_q-norm perturbation parameterized by c is in fact the SVM problem on the the dataset $\mathcal{S}(c,q)$, which is generated from $\mathcal{S}$ by placing a ℓ_q-norm ball with radius c around each samples, i.e., $\mathcal{S}(c,q) = \{(x,y) : \exists i \in [n], \text{ s.t.}, ||x - x_i||_p \leqslant c, y = y_i\}$. In other words, $u_{2,q}$ is the maximum ℓ_2-norm margin classifier under worst case ℓ_q-norm perturbation bounded by c.

In the remaining part of this section, we first analyze the convergence of the empirical adversarial risk, and then establish the implicit bias of GDAT with ℓ_q perturbation for $q \in [1,\infty]$. Our analysis for $q \in \{1,\infty\}$ is based on approximation argument. For ease of presentation, we only discuss when $q \in (1,\infty)$ in the main text, and defer the discussion for $q \in \{1,\infty\}$ in Appendix D.

Adversarial Risk Convergence. Our analysis is similar to the analysis for GDAT with ℓ_2 perturbation, where we use similar techniques to address issues such as non-differentiability at the origin and Hessian explosion of $\mathcal{L}_{\text{adv}}(\theta)$ around the origin.

Theorem 3.3. *Suppose $||x_i||_2 \leq 1$ for $i = 1, \ldots, n$, and let $\frac{1}{p} + \frac{1}{q} = 1$. In the GDAT with ℓ_q-norm perturbation, setting $c < \gamma_q$ and letting $M_p = \left[(1+c\sqrt{d})^2 + \frac{c(p-1)}{\gamma_{2,q}} d^{\frac{3p-2}{2p-2}}\right]\exp\left(-\gamma_{2,q}^2 + c\sqrt{d}\right)$, set $\eta^0 = 1$ and $\eta^t = \eta \leq \min\{\frac{1}{M_p}, 1\}$ for $t \geq 1$. We have that*

$$\frac{1}{n}\sum_{i=1}^{n} \max_{\delta_i\in\Delta} \exp\left(-y_i(x_i+\delta_i)^\top\theta^t\right) = \mathcal{O}\left(\frac{\log^2 t}{t\eta\gamma_{2,q}^2}\right). \tag{11}$$

We point out here that (6) is a special case of (11). In particular, by the definition of $\gamma_{2,q}(c)$, we have that $\gamma_{2,2}(c) = \gamma_2 - c$, which recovers bound (6) from (11).

Parameter Convergence. We show that if we set $c < \gamma_q$ in the GDAT algorithm with stepsizes specified in Theorem 3.3, with ℓ_q perturbation, the algorithm still possesses implicit bias property, i.e., θ^t still has directional convergence, and the limiting direction depends on the perturbation set Δ.

Theorem 3.4 (Implicit Bias of GDAT with ℓ_q-norm Perturbation)**.** *Under the same conditions in Theorem 3.3, define* $\overline{\theta}^t = \frac{\theta^t}{||\theta^t||_2}$*, then we have:*

$$1 - \left\langle \overline{\theta}^t, u_{2,q} \right\rangle = \mathcal{O}\left(\frac{\log n}{\log t}\right)$$

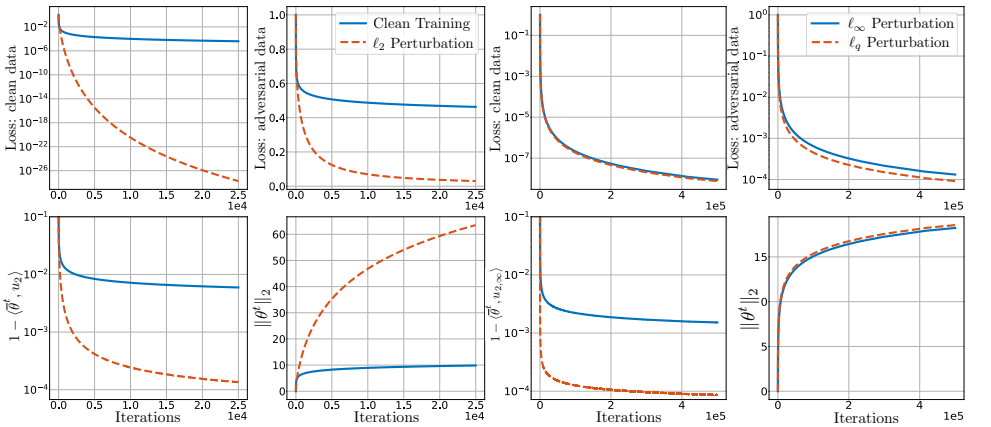
Combining Theorem 3.4 and Lemma 3.4, we conclude that GDAT with ℓ_q-norm perturbation indeed promotes robustness against ℓ_q perturbation. Using GDAT with ℓ_q-norm perturbation will result in a classifier which is the maximum ℓ_2-norm margin classifier under worst case ℓ_q-norm perturbations to the samples bounded by c. The learned classifier will have ℓ_q-norm margin at least c. As we increase perturbation level c to γ_q, the learned classifier will converge to maximum ℓ_q-norm margin classifier.

4 Numerical Experiment

In this section, we first conduct numerical experiments on linear classifiers to backup our theoretical findings. We further empirically extend our method to neural networks, where our numerical results demonstrate that our theoretical results can be potentially generalized.

Linear Classifiers. We investigate the empirical performance of the GDAT algorithm on linear classifiers, with training set $\mathcal{S} = \{((-0.5, 1), +1), ((-0.5, -1), -1), ((-0.75, -1), -1), ((2, 1), +1)\}$. It is straightforward to verify that the maximum ℓ_2-norm margin classifier is $u_2 = (0, 1)$.

Considering ℓ_2-norm perturbations, we first run standard clean training with GD, and GDAT with ℓ_2-norm perturbation ($c = 0.95\gamma_2$), for 2.5×10^4 number of iterations. In both GD and GDAT we take constant stepsizes, with $\eta = 1$ and $\eta = 0.1$, respectively. By Figure 1(a), we see that the convergence rate of adversarial loss using GDAT is similar to the convergence rate of clean loss using GD. However, when we directly compare the clean losses of GDAT and GD, GDAT clearly demonstrates an exponential speed-up in comparison with GD, which is consistent with Corollary 3.2. Additionally, as pointed out by Theorem 3.2, GDAT also enjoys significant speed-up in terms of the directional convergence of θ^t to u_2. We also compare the norm growth $||\theta^t||_2$, and observe that the norm generated by GDAT grows much faster than the norm generated by GD, which is also in alignment with our discussions in Section 3.1.



(a). Clean Training v.s. GDAT (ℓ_2 perturbation) (b). GDAT (ℓ_∞ perturbation) v.s. GDAT (ℓ_q perturbation)

Figure 1: GDAT of Linear Classifiers.

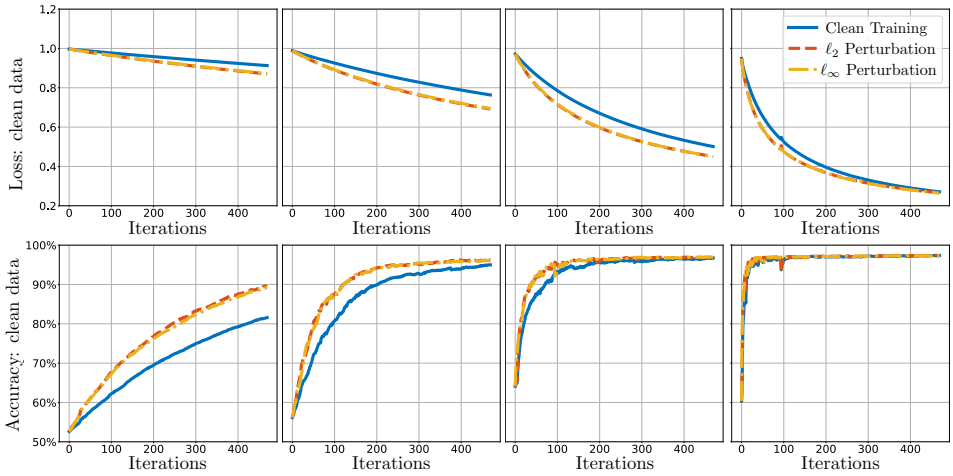
We further run GDAT with ℓ_∞-norm perturbation ($c = 0.5$). By Lemma 3.4, we have that $u_{2,\infty} = (0, 1)$. Note that the Hausdorff distance between ℓ_q-norm ball and ℓ_∞-norm ball distance goes to zero as q goes to infinity. Thus, we have that (10) for $q = 1000$ is a close approximation of (10) for $q = \infty$. We run two versions of GDAT, where one uses ℓ_q-norm perturbation with $q = 1000$, and the other uses ℓ_∞-norm perturbation. We run both algorithms with stepsize $\eta = 0.1$ for 5.0×10^5

number of iterations, and we present the results in Figure 1(b). We find that the two training methods behave similarly. In addition, the empirical directional convergence rates of θ^t just differ slightly.

Neural Networks. It is seen above that GDAT with ℓ_2-norm perturbations converges significantly faster than GD for linear classifiers in adversarial training. A natural question is whether this is still the case on adversarial training of more complicated neural networks. We conduct experiments on neural network with one hidden layer. We take the two classes from MNIST dataset with label "2" and "9" to form our training set $\mathcal{S}$. We also vary the width of the hidden layer in $\{64\times 64, 128\times 128, 256\times 256\}$.

One major difference from the case of linear classifiers is that we cannot solve the inner maximization problem of (2) exactly as it does not admits a closed-form solution. Instead, we solve the inner problem approximately using projected gradient descent with 20 iterations and stepsize 0.01. We test two versions of GDAT, where one adopts ℓ_2-norm perturbations ($c = 2.8$), and the other uses ℓ_∞-norm perturbations ($c = 0.1$). For standard clean training and the outer minimization problem in (2), we use the stochastic gradient descent algorithm with batch size 128 and constant stepsize 10^{-5}.

We compare the loss and classification accuracy, which are evaluated using the clean training samples, of standard clean training and GDAT. By Figure 2, we see that GDAT indeed accelerates the convergence of both loss and classification accuracy on clean training samples. The performance gap is most obvious when the width of the hidden layer is small, and reduces gradually as we increase the width of the hidden layer. We argue that such reduction comes from the fact that as network width increases, the margin on the samples outputted by the hidden layer also increases. As suggested by Theorem 3.2, in this case, a larger perturbation level c should be used. We conduct additional experiments with various perturbation level in Appendix E to empirically verify our argument.



(a). Width = 64×64 (b). Width = 128×128 (c). Width = 256×256 (d). Width = 512×512

Figure 2: GDAT of Neural Network on MNIST Dataset.

5 Discussions

We investigate the implicit bias of GDAT for linear classifier. There are several plausible natural extensions. For example, we can represent a linear classifier using a **deep linear network**, which is significantly overparameterized. Some recent results characterize the implicit bias of gradient descent for training deep linear networks (Ji and Telgarsky, 2019) and linear convolutional networks (Gunasekar et al., 2018b). Motivated by these results, investigating the implicit bias of GDAT in training deep linear networks worths future investigations.

Meanwhile, investigating implicit bias in **deep nonlinear networks** is a more important and challenging direction: (1) For linear classifiers, adding adversarial perturbations during training can be understood as a form of regularization, which explains the faster convergence in training. Although observed empirically, the potential acceleration of adversarial training is not yet understood in the current literature, to the best of our knowledge. (2) The notion of margin for neural networks still lacks proper definition, which we need to define to facilitate investigations on the effect of adversarial training in promoting robustness. (3) Ultrawide nonlinear networks have been shown to evolve similarly to linear networks using gradient descent (Ghorbani et al., 2019; Lee et al., 2019). We shall further investigate if our results on linear classifiers can be extended to wide nonlinear networks.

REFERENCES

ATHALYE, A., CARLINI, N. and WAGNER, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*.

BEN-TAL, A., EL GHAOUI, L. and NEMIROVSKI, A. (2009). *Robust Optimization*, vol. 28. Princeton University Press.

BRUTZKUS, A., GLOBERSON, A., MALACH, E. and SHALEV-SHWARTZ, S. (2018). SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*.

CARLINI, N. and WAGNER, D. (2016). Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311* .

CARLINI, N. and WAGNER, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*.

CHOROMANSKA, A., HENAFF, M., MATHIEU, M., AROUS, G. B. and LECUN, Y. (2015). The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*.

FEIGE, U., MANSOUR, Y. and SCHAPIRE, R. (2015). Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*.

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191* .

GOODFELLOW, I., SHLENS, J. and SZEGEDY, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018a). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*.

GUNASEKAR, S., LEE, J. D., SOUDRY, D. and SREBRO, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*.

HE, W., WEI, J., CHEN, X., CARLINI, N. and SONG, D. (2017). Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies*.

HINTON, G., DENG, L., YU, D., DAHL, G., MOHAMED, A.-R., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P. and KINGSBURY, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* **29**.

HOFFER, E., HUBARA, I. and SOUDRY, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*.

IOFFE, S. and SZEGEDY, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*.

JI, Z. and TELGARSKY, M. (2018). Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* .

JI, Z. and TELGARSKY, M. (2019). Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*.

KESKAR, N. S., MUDIGERE, D., NOCEDAL, J., SMELYANSKIY, M. and TANG, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

LEE, J., XIAO, L., SCHOENHOLZ, S. S., BAHRI, Y., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720* .

MADRY, A., MAKELOV, A., SCHMIDT, L., TSIPRAS, D. and VLADU, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

MOOSAVI-DEZFOOLI, S.-M., FAWZI, A. and FROSSARD, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

NEYSHABUR, B., SALAKHUTDINOV, R. R. and SREBRO, N. (2015a). Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*.

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015b). In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*.

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015c). Norm-based capacity control in neural networks. In *Conference on Learning Theory*.

PAPERNOT, N., MCDANIEL, P., GOODFELLOW, I., JHA, S., CELIK, Z. B. and SWAMI, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*.

PAPERNOT, N., MCDANIEL, P., WU, X., JHA, S. and SWAMI, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy*.

SINHA, A., NAMKOONG, H. and DUCHI, J. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* **19** 2822–2878.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15** 1929–1958.

SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I. and FERGUS, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

WALD, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics* **10** 299–326.

WILSON, A. C., ROELOFS, R., STERN, M., SREBRO, N. and RECHT, B. (2017). The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*.

XIE, C., WU, Y., VAN DER MAATEN, L., YUILLE, A. and HE, K. (2018). Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411* .

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

ZHANG, H., YU, Y., JIAO, J., XING, E. P., GHAOUI, L. E. and JORDAN, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*.

A PROOF OF PROPOSITION 3.1

Proof. Suppose $c < \gamma_q$. Letting $\theta_\alpha = \alpha u_q$ for $\alpha > 0$, we have

$$\begin{aligned}\mathcal{L}_{\text{adv}}(\theta_\alpha) &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \theta_\alpha + c||\theta_\alpha||_p\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\alpha y_i x_i^\top u_q + c\alpha\right) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\alpha\gamma_q + c\alpha\right).\end{aligned}$$

Letting $\alpha \to \infty$, we obtain $\lim_{\alpha\to\infty} \mathcal{L}_{\text{adv}}(\theta_\alpha) = 0$, which implies $\inf_{\theta\in\mathbb{R}^d} \mathcal{L}_{\text{adv}}(\theta) = 0$. Note that $\mathcal{L}(\theta)$ does not admit any finite minimizer since $\mathcal{L}_{\text{adv}}(\theta) > 0$ for any $\theta \in \mathbb{R}^d$.

If $c > \gamma_q$, by the definition of maximum ℓ_q-norm margin, for any $\theta \in \mathbb{R}^d$, there exists $(y_i, x_i) \in \mathcal{S}$ for some $i \in [n]$ such that $y_i x_i^\top \theta \leqslant \gamma_q ||\theta||_p$. Hence, $\mathcal{L}_{\text{adv}}(\theta) \geqslant \exp\left(n^{-1}(c-\gamma_q)||\theta||_p\right)$. Then it is easy to see that $\mathcal{L}_{\text{adv}}(\theta)$ has bounded sublevel set and hence a finite minimizer $\widehat{\theta}$. Since $\mathcal{L}_{\text{adv}}(\theta)$ is convex, we examine its first-order KKT condition, given by

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \widehat{\theta} + c||\widehat{\theta}||_p\right)\left(-y_i x_i + c\partial||\widehat{\theta}||_p\right) \ni 0. \tag{12}$$

Consider the regularized problem with regularization parameter η:

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \theta\right) + \eta||\theta||_p.$$

Its first-order KKT condition is

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \theta\right)(-y_i x_i) + \eta\partial||\theta||_p \ni 0. \tag{13}$$

Looking at (12) and (13) together, by taking $\eta = \frac{c}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \widehat{\theta} + c||\widehat{\theta}||_p\right)$, we have that the solution to the adversarial training problem $\widehat{\theta}$ is also the solution to the regularized problem. □

To facilitate our later discussions, we point out that by the conjugacy of ℓ_p-norm and ℓ_q-norm, (4) has the following equivalent form that

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}_{\text{adv}}(\theta) = \min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \theta + c||\theta||_p\right). \tag{14}$$

In fact, one can verify that the GDAT algorithm is equivalent to gradient descent algorithm on (14).

B PROOFS FOR SECTION 3.1

Proof of Lemma 3.1. Recall we have $\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{||\delta||_2\leq c} \exp\left(-y_i(x_i+\delta_i)^\top\theta\right)$. For each sample $(x_i, y_i) \in \mathcal{S}$, given a classifier θ, the worse case perturbation is $\widetilde{\delta}_i = \text{argmax}_{||\delta||_2\leq c} \exp\left(-y_i(x_i+\delta)^\top\theta\right) = \text{argmin}_{||\delta||_2\leq c} y_i\delta^\top\theta$. The corresponding loss is $\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right)$.

Since for a fixed δ_i, the function $\exp\left(-y_i(x_i+\delta_i)^\top\theta\right)$ is convex in θ, hence the gradient of $\mathcal{L}_{\text{adv}}(\theta)$ is

$$-\nabla\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right) y_i(x_i+\widetilde{\delta}_i).$$

Then from the definition of u_2 (5), we have

$$\langle -\nabla \mathcal{L}_{\text{adv}}(\theta), u_2 \rangle = \sum_{i=1}^{n} \exp\left(-y_i(x_i + \widetilde{\delta}_i)^\top \theta\right) \left\langle y_i(x_i + \widetilde{\delta}_i), u_2 \right\rangle \tag{15}$$

$$\geq \sum_{i=1}^{n} \exp\left(-y_i(x_i + \widetilde{\delta}_i)^\top \theta\right) (\langle y_i x_i, u_2 \rangle - c) \tag{16}$$

$$\geq \sum_{i=1}^{n} \exp\left(-y_i(x_i + \widetilde{\delta}_i)^\top \theta\right) (\gamma_2 - c) = \mathcal{L}_{\text{adv}}(\theta)(\gamma_2 - c), \tag{17}$$

where in the second inequality holds since $||\widetilde{\delta}_i||_2 \leqslant c$ and $||u_2||_2 = 1$. □

Proof of Corollary C.1. Since $\mathcal{L}_{\text{adv}}(\theta)$ is not differentiable at $\theta^0 = 0$, we use subgradient (note that $\mathcal{L}_{\text{adv}}(\theta)$ is convex) at 0. Specifically, we take $\nabla \mathcal{L}_{\text{adv}}(\theta^0) = \frac{1}{n}\sum_{i=1}^{n} z_i \in \partial \mathcal{L}_{\text{adv}}(\theta^0)$. Then we have $\left\langle \theta^1, u_2 \right\rangle = \frac{\eta^0}{n}\sum_i \langle z_i, u_2 \rangle \geq \eta^0 \gamma_2$, where the last inequality uses the definition of γ_2.

By Lemma 3.1, we have $\langle \theta^t, u_2 \rangle \geq \eta^0 \gamma_2$ for all $t \geq 1$, which also implies $\langle v, u_2 \rangle \geq \eta^0 \gamma_2$ and hence $||v^t||_2 \geq \eta^0 \gamma_2$ for $v \in \left[\theta^t, \theta^{t+1}\right]$. □

Proof of Theorem 3.1. For simplicity, we let $z_i = y_i x_i$, where we have $||z_i||_2 \leq 1$ as we assume $\|x_i\|_2 \leq 1$. We have

$$\nabla \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_2\right) \left(-z_i + c\frac{\theta}{||\theta||_2}\right),$$

$$\begin{aligned}
\nabla^2 \mathcal{L}_{\text{adv}}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_2\right) \left(-z_i + c\frac{\theta}{||\theta||_2}\right)\left(-z_i + c\frac{\theta}{||\theta||_2}\right)^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_2\right) c\left(||\theta|| I - \frac{\theta\theta^\top}{||\theta||_2}\right)/||\theta||_2^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_2\right)\left[z_i z_i^\top - 2\frac{c z_i^\top \theta}{||\theta||_2} + c^2\theta\theta^\top/||\theta||_2^2 + cI/||\theta||_2 - c\theta\theta^\top/||\theta||_2^3\right].
\end{aligned}$$

Note that the Hessian expression indicates that the objective is highly non-smooth around origin, and the loss is not even differentiable at origin. However, we shall prove that starting from origin, every iteration generated by GADT stays away from the origin with distance bounded below.

Using Taylor's expansion, and by definition $\theta^{t+1} = \theta^t - \eta^t \nabla \mathcal{L}_{\text{adv}}(\theta^t)$, we have

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||)^2}{2} \max_{v \in [\theta^t, \theta^{t+1}]} \lambda(H(v))_{\max}, \tag{18}$$

where $\lambda(H(v))_{\max}$ denotes the largest eigenvalue of $H(v)$, where

$$H(v) = \frac{1}{n}\sum_{i=1}^{I} n \exp\left(-z_i^\top v + c||v||_2\right)\left[z_i z_i^\top - 2\frac{c z_i^\top v}{||v||_2} + c^2 v v^\top/||v||_2^2 + cI/||v||_2 - c v v^\top/||v||_2^3\right].$$

To upper bound $H(v)$, we need a lower bound on $||v||$, which is readily given by Corollary C.1. That is, $||v||_2 \geq \eta^0 \gamma_2$.

We now analyze (18) for $t \geq 1$, where we show that $\mathcal{L}_{\text{adv}}(\theta^t)$ is locally smooth with parameter proportional to $\mathcal{L}_{\text{adv}}(\theta^t)$, and with proper stepsize, the risk is monotonely decreasing. Note that $z_i z_i^t \leq I$, $-2c\frac{z_i^\top v}{||v||_2} \leq 2cI$, $c^2 v v^\top/||v||_2^2 \leq c^2 I$. Now since $||v^t||_2 \geq \eta^0 \gamma_2$, we have $cI/||v||_2 - c v v^\top/||v||_2^3 \leq \frac{2c}{\eta^0 \gamma_2} I$. Plugging them in, we have

$$\begin{aligned}
H(v) &\leq \frac{1}{n}\sum_{i} \exp\left(-z_i^\top v + c||v||_2\right)\left(1 + 2c + c^2 + \frac{2c}{\eta^0 \gamma_2}\right) I \\
&= L_{adv}(v)\left(1 + 2c + c^2 + \frac{2c}{\eta^0 \gamma_2}\right) I,
\end{aligned}$$

and (18) reduces to

$$\begin{aligned}\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) \leq & \mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \\ & + \frac{(\eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||)^2}{2} \left[(1+c)^2 + \frac{2c}{\eta^0 \gamma_2}\right] \max\left\{\mathcal{L}_{\mathrm{adv}}(\theta^t), \mathcal{L}_{\mathrm{adv}}(\theta^{t+1})\right\}.\end{aligned} \tag{19}$$

Suppose $\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) > \mathcal{L}_{\mathrm{adv}}(\theta^t)$, and let $M = \left[(1+c)^2 + \frac{2c}{\eta^0\gamma_2}\right]$. We have

$$\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||)^2}{2} M \mathcal{L}_{\mathrm{adv}}(\theta^{t+1}),$$

which implies

$$\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) \leq \left(1 - \frac{M(\eta^t)^2}{2} ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2\right)^{-1} \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2\right). \tag{20}$$

Meanwhile, if we choose η^t satisfying

$$\eta^t M = \eta^t \mathcal{L}_{\mathrm{adv}}(\theta^t) \left[(1+c)^2 + \frac{2c}{\eta^0\gamma_2}\right] \leq 1, \tag{21}$$

then we have the right hand side of (20) is upper bounded by $\mathcal{L}_{\mathrm{adv}}(\theta^t)$, and we have

$$\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) \leq \left(1 - \frac{M(\eta^t)^2}{2} ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2\right)^{-1} \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2\right) < \mathcal{L}_{\mathrm{adv}}(\theta^t),$$

which is clearly a contradiction. Hence, if η^t satisfies (21), by (19) we have

$$\begin{aligned}\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) &\leq \mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||)^2}{2} \left[(1+c)^2 + \frac{2c}{\eta^0\gamma_2}\right] \mathcal{L}_{\mathrm{adv}}(\theta^t) \\ &\leq \mathcal{L}_{\mathrm{adv}}(\theta^t) - \frac{\eta^t}{2} ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2,\end{aligned} \tag{22}$$

where the last inequality holds by the choice of η^t in (21).

Note that if (21) holds for $t = 1$ for $\eta^1 = \eta$, by induction it is easy to see that with constant stepsize $\eta^t = \eta$ for $t \geq 1$, (21) holds for all $t \geq 1$. Hence for $t \geq 1$, we choose stepsize η such that $\eta \mathcal{L}_{\mathrm{adv}}(\theta^1) \left[(1+c)^2 + \frac{2c}{\eta^0\gamma_2}\right] \leq 1$. Note that $\mathcal{L}_{\mathrm{adv}}(\theta^1) = \frac{1}{n}\sum_{i=1}^n \exp\left(-z_i^\top \theta^1 + c||\theta^1||_2\right) \leq \exp\left((1+c)\eta^0\right)$ since $||\theta_1|| \leq \eta^0$. Then we only require

$$\begin{aligned}\eta &\leq \exp\left(-(1+c)\eta^0\right) \cdot \frac{\eta^0\gamma_2}{(1+c)^2\eta^0\gamma_2 + 2c} \\ &= \exp\left(-(1+c)\eta^0\right) \cdot \frac{\eta^0(1+c)\gamma_2/(1+c)}{(1+c)^2\eta^0\gamma_2 + 2c} \\ &\leq \frac{\gamma_2/e}{(1+c)^3\gamma_2 + 2c(1+c)},\end{aligned}$$

where in the last inequality we take $\eta^0 = 1$ and use basic inequality $\exp(-x)x \leq e^{-1}$ for $x \geq 1$. In summary, we choose $\eta^0 = 1$ and $\eta^t = \eta = \min\{\frac{\gamma_2/e}{(1+c)^3\gamma_2 + 2c(1+c)}, 1\}$ for $t \geq 1$, then by previous argument, we have (22) holds for all $t \geq 1$.

Now we are ready to apply the standard smoothness-based analysis of gradient descent using (22), take any $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}||\theta^{t+1} - \theta||_2^2 &= ||\theta^t - \theta||_2^2 - 2\eta^t \left\langle \nabla \mathcal{L}_{\mathrm{adv}}(\theta^t), \theta^t - \theta \right\rangle + (\eta^t)^2 ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \\ &\leq ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta)\right) + (\eta^t)^2 ||\nabla \mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \\ &\leq ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta)\right) + 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta^{t+1})\right) \\ &= ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) - \mathcal{L}_{\mathrm{adv}}(\theta)\right),\end{aligned}$$

where the first inequality holds by the convexity of $\mathcal{L}_{\mathrm{adv}}(\theta)$, and the second inequality holds by (22). Now sum up the above inequality from $s = 1$ to $t - 1$. By $\eta^t = \eta \leq 1 = \eta^0$ and $\mathcal{L}_{\mathrm{adv}}(\theta^{s+1}) \leq \mathcal{L}_{\mathrm{adv}}(\theta^s)$, we have

$$\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta) \leq \frac{1}{2t\eta} ||\theta^1 - \theta||_2^2 \leq \frac{1}{t\eta}\left(||\theta||_2^2 + ||\theta^1||_2^2\right).$$

Now since θ is arbitrary, letting $\theta = \frac{\log(t)}{\gamma_2 - c} \cdot u_2$, we have

$$||\theta||_2^2 + ||\theta^1||_2^2 \leq \frac{\log^2 t}{(\gamma_2 - c)^2} + (1+c)^2,$$

and

$$\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top u_2 \cdot \frac{\log t}{\gamma_2 - c} + c \cdot \frac{\log t}{\gamma_2 - c}\right) \leq \frac{1}{t},$$

which yields

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \frac{1}{t} + \left(\frac{\log^2 t}{(\gamma_2 - c)^2} + (1+c)^2\right) = \mathcal{O}\left(\frac{\log^2 t}{t\eta(\gamma_2 - c)^2}\right).$$

□

Proof of Lemma 3.2. For simplicity, we let $z_i = y_i x_i$ and $\ell_i(\theta) = \exp\left(-z_i^\top \theta + c||\theta||_2\right)$. Define

$$\alpha = \min_{||\xi||_2 = 1, \xi \in \text{span}(u_2)^\perp} \max_{i \in \text{SV}(\mathcal{S})} \langle \xi, z_i \rangle$$

where $\text{SV}(\mathcal{S})$ denotes the set of support vectors. It has been shown in Ji and Telgarsky (2019) (Lemma 2.10) that $\alpha > 0$ with probability 1 if the data is sampled from absolutely continuous distribution.

We have

$$\begin{aligned}
\left\langle \nabla \mathcal{L}_{\text{adv}}(\theta^t), \theta_\perp^t \right\rangle &= \frac{1}{n}\left\langle \sum_{i=1}^{n} \exp\left(-z_i^\top \theta^t + c||\theta^t||_2\right)\left(-z_i + c\frac{\theta^t}{||\theta^t||_2}\right), \theta_\perp^t \right\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta^t)\left\langle -z_i, \theta_\perp^t \right\rangle + \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta^t)\left\langle c\frac{\theta^t}{||\theta^t||_2}, \theta_\perp^t \right\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta^t)\left\langle -z_i, \theta_\perp^t \right\rangle \\
&\geq \frac{1}{n}\left[\ell_j(\theta^t)\left\langle -z_j', \theta_\perp^t \right\rangle + \sum_{\left\langle z_i, \theta_\perp^t \right\rangle \geq 0, i \neq j} \ell_i(\theta^t)\left\langle -z_i, \theta_\perp^t \right\rangle\right], \qquad (23)
\end{aligned}$$

where $z_j' \in S$ is arbitrary, by definition of α: $\left\langle -z_j', \theta_\perp^t \right\rangle \geq \alpha||\theta_\perp^t||_2$.

We bound the first term as

$$\begin{aligned}
\ell_j(\theta^t)\left\langle -z_j', \theta_\perp^t \right\rangle &\geq \exp\left(-(z_j')^\top \theta^t + c||\theta^t||_2\right)\alpha||\theta_\perp^t||_2 \\
&= \exp\left(-(z_j')^\top \theta_\perp^t - (z_j')^\top \theta_{u_2}^t + c||\theta^t||_2\right)\alpha||\theta_\perp^t||_2 \\
&\geq \exp\left(-\left\langle \theta^t, \gamma_2 u_2 \right\rangle\right)\exp\left(\alpha||\theta_\perp^t||_2\right)\alpha||\theta_\perp^t||_2 \exp\left(c||\theta^t||_2\right),
\end{aligned}$$

where the second inequality uses $\left\langle z_j', u_2 \right\rangle \geq \gamma_2$.

On the other hand, we can bound the second term in (23) as

$$\begin{aligned}
\frac{1}{n}\sum_{\left\langle z_i, \theta_\perp^t \right\rangle \geq 0, i \neq j} \ell_i(\theta^t)\left\langle -z_i, \theta_\perp^t \right\rangle &\geq \frac{1}{n}\sum_{\left\langle z_i, \theta_\perp^t \right\rangle \geq 0, i \neq j} \exp\left(-z_i^\top \theta^t + c||\theta^t||_2\right)\left\langle -z_i, \theta_\perp^t \right\rangle \\
&= \frac{1}{n}\sum_{\left\langle z_i, \theta_\perp^t \right\rangle \geq 0, i \neq j} \exp\left(-z_i^\top \theta_{u_2}^t - z_i^\top \theta_\perp^t + c||\theta^t||_2\right)\left\langle -z_i, \theta_\perp^t \right\rangle \\
&\geq \exp\left(-\left\langle \theta^t, \gamma_2 u_2 \right\rangle\right)\exp(c||\theta^t||_2)\exp(-z_i^\top \theta_\perp^t)\left\langle -z_i, \theta_\perp^t \right\rangle \\
&\geq \exp\left(-\left\langle \theta^t, \gamma_2 u_2 \right\rangle\right)\exp(c||\theta^t||_2)(-\frac{1}{e}),
\end{aligned}$$

where in the last inequality holds since $\langle \theta^t, u_2 \rangle \geq 0$, $\left\langle z_i, \theta_{u_2}^t \right\rangle = z_i^\top \left(u_2^\top \theta^t\right) u_2 \geq \gamma_2 \langle \theta^t, u_2 \rangle$ and $-x\exp(-x) \geq -\frac{1}{e}$ for $x \geq 0$.

Plugging the two bounds above into (23), we have

$$\left\langle \nabla \mathcal{L}_{\text{adv}}(\theta^t), \theta_\perp^t \right\rangle \geq \exp\left(-\left\langle \theta^t, \gamma_2 u_2 \right\rangle\right)\exp(c||\theta^t||_2)\left[\frac{1}{n}\exp\left(\alpha||\theta_\perp^t||_2\right)\alpha||\theta_\perp^t||_2 - \frac{1}{e}\right],$$

which is non-negative when $||\theta_\perp^t||_2 \geq K' = \frac{1+\log n}{\alpha}$.

Supposing $||\theta_\perp^t||_2 \geq K'$, by gradient descent update, we have,

$$\begin{aligned} ||\theta_\perp^{t+1}||_2^2 &= ||\theta_\perp^t||_2^2 - 2\eta^t \left\langle \nabla \mathcal{L}_{\text{adv}}(\theta^t), \theta_\perp^t \right\rangle + (\eta^t)^2 ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||^2 \\ &\leq ||\theta_\perp^t||_2^2 + 2\eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 \\ &\leq ||\theta_\perp^t||_2^2 + 2\left(\mathcal{L}_{\text{adv}}(\theta^t) - \mathcal{L}_{\text{adv}}(\theta^{t+1})\right), \end{aligned} \tag{24}$$

where the last inequality uses (22).

Now let t_0 satisfy $||\theta_\perp^{t_0-1}||_2 < K'$ and $||\theta_\perp^{t_0-1}||_2 \geq K'$. Define $t_1 = \min\{s \geq t_0 : ||\theta_\perp^s||_2 < K'\}$, when $||\theta_\perp^s||_2 \geq K'$ for all $s \geq t_0$ we define $t_1 = \infty$. That is for any $t \in \{t_0, \ldots, t_1 - 1\}$, we have $||\theta_\perp^t||_2 \geq K'$. then for any s such that $t_0 \leq s < t_1$, summing (24) up from t_0 to $s-1$ yields:

$$\begin{aligned} ||\theta_\perp^s||_2^2 &\leq ||\theta_\perp^{t_0}||_2^2 + 2\left(\mathcal{L}_{\text{adv}}(\theta^{t_0}) - \mathcal{L}_{\text{adv}}(\theta^s)\right) \\ &\leq ||\theta_\perp^t||_2^2 + 2\exp(1+c) \\ &\leq ||\theta_\perp^t||_2^2 + 18, \end{aligned}$$

where we use $\mathcal{L}_{\text{adv}}(\theta^t) \leq \mathcal{L}_{\text{adv}}(\theta^1) \leq \exp(1+c)$ and $c < 1$. This inequality shows that for $\theta^t \in \{\theta^{t_0}, \ldots, \theta^{t_1-1}\} \subset \{\theta : ||\theta_\perp||_2 \geq K'\}$,

$$||\theta_\perp^t||_2 \leq ||\theta_\perp^{t_0}||_2 + 18.$$

Then, we only need to bound $||\theta_\perp^{t_0}||_2$ to conclude the proof, where t_0 is the first time θ^t enters $\{\theta : ||\theta_\perp||_2 \geq K'\}$. We have

$$\theta_\perp^{t_0} = \theta_\perp^{t_0-1} + \eta^{t_0-1} P_\perp \left(\frac{1}{n} \sum_{i=1}^n \ell_i(\theta^{t_0-1})(z_i - c\frac{\theta^{t_0-1}}{||\theta^{t_0-1}||_2}) \right),$$

where $P_\perp(\cdot)$ denotes the projection onto $\text{span}(u_2)^\perp$. Note that t_0 is the first time θ^t (re)-enters the region $\{\theta : ||\theta_\perp||_2 \geq K'\}$, and thus $||\theta_\perp^{t_0-1}||_2 < K'$. We have

$$||\theta_\perp^{t_0}||_2 \leq K' + \eta^{t_0-1}(1+c) \leq K' + 1 + c < K' + 2,$$

where the last inequality we use $c < \gamma_2 \leq 1$.

In summary, we have shown that for any t such that $||\theta_\perp^t||_2 \geq K'$, we have $||\theta_\perp^t||_2 \leq K' + 20$, and we conclude that $||\theta_\perp^t||_2 = K' + 20 = K$ for all $t \geq 0$. Note that K only depends $\alpha(\mathcal{S})$ and sample size n. □

Proof of Lemma 3.3. To obtain a lower bound on $||\theta^t||_2$, we first denote $\theta^t = \theta_u^t + \theta_\perp^t$, where θ_u^t denotes the projection of θ onto $\text{span}(u_2)$, and $\theta_\perp^t$ denotes the projection of θ onto $\text{span}(u_2)^\perp$. We have

$$\frac{1}{n} \sum_{i=1}^n \exp(-z_i^\top \theta_u^t - z_i^\top \theta_\perp^t) \leq \frac{\log^2 t}{t\eta(\gamma_2 - c)^2} \exp(-c||\theta^t||_2).$$

Let us assume that $||\theta_\perp^t||$ is bounded so that $\exp(||\theta_\perp^t||) \leq M$, which will be verified immediately. Choosing an arbitrary support vector z_i, we have $0 < \langle z_i, \theta_u^t \rangle = \langle z_i, u_2 \rangle \langle \theta^t, u_2 \rangle = \gamma_2 \langle \theta^t, u_2 \rangle = \gamma_2 ||\theta_u^t||_2 \leq \gamma_2 ||\theta^t||_2$, hence the previous inequality becomes:

$$\exp(-\gamma_2 ||\theta^t||_2) \leq \frac{n \log^2 t}{t\eta(\gamma_2 - c)^2} \exp(-c||\theta^t||_2) M,$$

which is equivalent to

$$||\theta^t||_2 \geq \log\left(\frac{t\eta(\gamma_2 - c)^2}{nM\log^2 t}\right) / (\gamma_2 - c). \tag{25}$$

Now we only need to show that $||\theta_\perp^t|| \leq M$ for all t for some M. Since we have shown in Lemma 3.2 that $||\theta^t||_2 \leq K$, we choose $M = e^K \leq \exp\left(\frac{20+\log n}{\alpha}\right) = \mathcal{O}(n^{\frac{1}{\alpha}})$, and the lower bound (25) becomes

$$||\theta^t||_2 \geq \log\left(\frac{t\eta(\gamma_2 - c)^2}{n^{1+1/\alpha}\log^2 t}\right) / (\gamma_2 - c), \tag{26}$$

which concludes our proof. □

Proof of Theorem 3.2. We denote $\theta^t = \theta_u^t + \theta_\perp^t$, where θ_u^t denotes the projection of θ onto span(u_2), and $\theta_\perp^t$ denotes the projection of θ onto span$(u_2)^\perp$. Combine Lemma 3.2 and Lemma 3.3, we have

$$\begin{aligned}
1 - \left\langle \frac{\theta^t}{||\theta^t||_2}, u_2 \right\rangle &= 1 - \frac{\left\langle \theta_{u_2}^t, u_2 \right\rangle + \langle \theta_\perp^t, u_2 \rangle}{||\theta^t||_2} \leq 1 - \frac{\left\langle \theta_{u_2}^t, u_2 \right\rangle}{||\theta^t||_2} + \frac{K}{||\theta^t||_2} \\
&= 1 - \frac{||\theta_{u_2}^t||_2}{||\theta^t||_2} + \frac{K}{||\theta^t||_2} \leq 1 - \frac{||\theta_{u_2}^t||_2^2}{||\theta^t||_2^2} + \frac{K}{||\theta^t||_2} \\
&= \frac{||\theta_\perp^t||_2^2}{||\theta^t||_2^2} + \frac{K}{||\theta^t||_2} \\
&\leq \frac{K^2}{||\theta^t||_2^2} + \frac{K}{||\theta^t||_2}.
\end{aligned}$$

By our choice of c and T that $\gamma_2 - c = \left(\frac{n^{1+1/\alpha}\log^2 T}{\eta T}\right)^{1/2}$, together Lemma 3.3, the Theorem holds as desired. □

Proof of Corollary 3.2. By Lemma 3.3 and the the choice of parameters that $\gamma_2 - c = \left(\frac{n^{1+1/\alpha}\log^2 T}{\eta T}\right)^{1/2}$, we have:

$$||\theta^T||_2 \geq \left(\frac{\eta T}{n^{(1+1/\alpha)}\log^2 T}\right)^{1/2}.$$

Together with Theorem 3.1, we have

$$\begin{aligned}
\mathcal{L}(\theta^T) &= \mathcal{L}_{\text{adv}}(\theta^T)\exp\left(-c||\theta^T||_2\right) \\
&\leqslant \frac{\log^2 T}{T\eta(\gamma_2 - c)^2}\exp\left(-c\left(\frac{\eta T}{n^{(1+1/\alpha)}\log^2 T}\right)^{1/2}\right) \\
&= \mathcal{O}\left(\exp\left(-c\left(\frac{\eta T}{n^{(1+1/\alpha)}\log^2 T}\right)^{1/2}\right)\right).
\end{aligned}$$

where the last equality holds by the parameter choice $\gamma_2 - c = \left(\frac{n^{1+1/\alpha}\log^2 T}{\eta T}\right)^{1/2}$. Finally, letting $\mu = c\left(\frac{\eta}{n^{1+1/\alpha)}}\right)^{1/2}$, the claim follows immediately. □

C Proofs for Section 3.2

In this section, we consider general ℓ_q-norm perturbations. In short, we show that no matter how small the perturbation is, adversarial training changes the implicit bias of standard clean training using gradient descent, and adapt it to specific norm we choose for adversarial training.

Intuitively, we might expect that under the ℓ_q-norm perturbation the implicit bias of gradient descent algorithm changes to converging in direction to ℓ_q-norm max margin solution $\overline{u}_q$. We provide a counter example here. Consider $\mathcal{S} = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2)\}$ with $x_1 = (10, 1), x_2 = (-10, -1)$ and $y_1 = 1, y_2 = -1$.

It is easy to see that the ℓ_∞-norm max margin solution is $\overline{u}_\infty = (1, 0)$ with $\gamma_\infty = 10$, and the ℓ_2-norm max margin solution is $\overline{u}_2 = (\frac{10}{\sqrt{101}}, \frac{1}{\sqrt{101}})$ with $\gamma_2 = \sqrt{101}$.

Without perturbation, we have that the gradient descent initialized at the origin converges in direction to ℓ_2-norm max margin solution $\overline{u}_2$ with one step. Now we take l_∞-norm perturbation with $c = 0.5$, the negative gradient is given by: $-\nabla\mathcal{L}_{\text{adv}}(\theta) = \frac{\ell_1(\theta)}{2}(z_1 - c\cdot\text{sign}(\theta)) + \frac{\ell_2(\theta)}{2}(z_2 - c\cdot\text{sign}(\theta))$. We initialize gradient descent at the origin with any constant step size. By the symmetry of the training data, we have that θ^t always stays always inside quadrant I, and converges in direction to $\overline{u} = (\frac{\sqrt{361}}{\sqrt{362}}, \frac{1}{\sqrt{362}})$, which is neither $\overline{u}_\infty$ or $\overline{u}_2$, but inside the interior of convex hull of $\overline{u}_\infty$ and $\overline{u}_2$. In fact, $\overline{u}$ exactly equals to the $u_{2,\infty}$ defined in (10).

Proof of Lemma 3.4. We prove that solutions to (10) and the robust SVM against ℓ_q-norm perturbation parameterized by c (8) are equal up to a constant factor. We first have that $\gamma_{2,q}(c)$ in (10) is equivalent to

$$\gamma_{2,q} = \max_{||\theta||_2 \leq 1} \min_{i \in [n]} y_i x_i^\top \theta - c||\theta||_p. \tag{27}$$

We denote the unique solution to (27) as $u_{2,q}$. It is not difficulty to see that

$$y_i x_i^\top u_{2,q} - c||u_{2,q}||_2 \geqslant \gamma_{2,q}, \forall i = 1, \ldots, n.$$

We define $\overline{u}_{2,q} = \frac{u_{2,q}}{\gamma_{2,q}}$, then:

$$y_i x_i^\top \overline{u}_{2,q} - c||\overline{u}_{2,q}||_2 \geqslant 1, \forall i = 1, \ldots, n.$$

It is now clear that $\overline{u}_{2,q}$ is a feasible solution to (8). We denote the optimal solution to (8) as $\overline{u}$, then we have by the optimality of $\overline{u}$ that $||\overline{u}||_2 \leq ||\overline{u}_{2,q}||_2 \leq \frac{||u_{2,q}||_2}{\gamma_{2,q}}$, and feasibility of $\overline{u}$ that

$$y_i x_i^\top (\gamma_{2,q}\overline{u}) - c||\gamma_{2,q}\overline{u}||_2 \geq \gamma_{2,q} \forall i = 1, \ldots, n.$$

Then from previous two inequalities we have $\gamma_{2,q}\overline{u}$ is a feasible solution to (27) with objective value equal to the optimal objective value of (27). Since the optimal solution to (27) is unique, this implies that $\overline{u} = \frac{u_{2,q}}{\gamma_{2,q}}$, which concludes our proof. □

We extend Lemma 3.1 to bounded ℓ_q-norm perturbation set.

Lemma C.1. *Recall the definition of $\gamma_{2,q}$ in* (10)*. For any $c < \gamma_q$, we have that $\langle -\nabla \mathcal{L}_{\text{adv}}(\theta), u_{2,q} \rangle \geq \mathcal{L}_{\text{adv}}(\theta)\gamma_{2,q}$ for all $\theta \in \mathbb{R}^d$.*

Proof. Recall that we have $\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \max_{||\delta||_q \leq c} \exp\left(-y_i(x_i+\delta_i)^\top\theta\right)$. For each sample $(x_i, y_i) \in \mathcal{S}$, given a classifier θ, the worst case perturbation is $\widetilde{\delta}_i = \text{argmax}_{||\delta||_q \leq c} \exp\left(-y_i(x_i+\delta)^\top\theta\right) = \text{argmin}_{||\delta||_q \leq c} y_i \delta^\top \theta$. The corresponding loss is then $\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right)$.

Since for a fixed δ_i, the function $\exp\left(-y_i(x_i+\delta_i)^\top\theta\right)$ is convex in θ, hence the gradient of $\mathcal{L}_{\text{adv}}(\theta)$ is

$$-\nabla \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right) y_i(x_i+\widetilde{\delta}_i).$$

Then by the definition of $u_{2,q}$, we have

$$\langle -\nabla \mathcal{L}_{\text{adv}}(\theta), u_2 \rangle = \sum_{i=1}^n \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right) \left\langle y_i(x_i+\widetilde{\delta}_i), u_{2,q} \right\rangle \tag{28}$$

$$\geq \sum_{i=1}^n \exp\left(-y_i(x_i+\widetilde{\delta}_i)^\top\theta\right) \gamma_{2,q} = \mathcal{L}_{\text{adv}}(\theta)\gamma_{2,q}, \tag{29}$$

where the second inequality holds by $||\widetilde{\delta}_i||_q \leq c$, and the definitions of $u_{2,q}$ and $\gamma_{2,q}$ in Lemma 3.4. □

Note that for $q = 2$, by the fact that $\gamma_{2,2}(c) = \gamma_2 - c$, we immediately have Lemma 3.1 holds.

As a direct corollary of Lemma C.1, we have $||\theta^t||_2$ is bounded away from 0 for all $t \geq 1$.

Corollary C.1. *Let $\theta^0 = 0$ in Algorithm 1, we have: $||\theta^t||_2 \geq \eta^0 \gamma_{2,q}$ for al $t \geq 1$.*

Proof. The proof is similar to Corollary 3.1, we omit the details here. □

Proof of Theorem 3.3. For simplicity, we define $z_i = y_i x_i$ and have $||z_i||_2 \leq 1$ since $\|x_i\|_2 \leq 1$. We have for $\theta \neq 0$

$$\nabla \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \exp\left(-z_i^\top\theta + c||\theta||_p\right)\left(-z_i + c\partial||\theta||_p\right),$$

$$\nabla^2 \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \exp\left(-z_i^\top\theta + c||\theta||_p\right)\left(-z_i + c\partial||\theta||_p\right)\left(-z_i + c\partial||\theta||_p\right)^\top$$

$$+ \frac{1}{n}\sum_{i=1}^n \exp\left(-z_i^\top\theta + c||\theta||_p\right) c\left((1-p)||\theta||_p^{1-2p}(\odot^{p-1}\theta)(\odot^{p-1}\theta)^\top + (p-1)||\theta||_p^{1-p}\text{diag}(\odot^{p-2}\theta)\right),$$

where $\odot^{p-1}\theta$ denotes taking element-wise $(p-1)$-th power of θ.

Note that we have $||\partial||\theta||_p||_q = 1$. By the conjugacy of ℓ_p-norm and ℓ_q-norm with $\frac{1}{p}+\frac{1}{q}=1$, we have$||\theta||_p = \max_{||s||_q \leq 1} \langle \theta, s\rangle$. Hence we upper bound the first term in Hessian $\nabla^2 \mathcal{L}_{\text{adv}}(\theta)$ above by

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_p\right)\left(-z_i + c\partial||\theta||_p\right)\left(-z_i + c\partial||\theta||_p\right)^\top \tag{30}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + c||\theta||_p\right)(1+c\sqrt{d}||\theta||_2)^2. \tag{31}$$

We further have:

$$\begin{aligned}(p-1)||\theta||_p^{p-1}\text{diag}(\odot^{p-2}\theta) &\leq (p-1)\frac{\text{diag}(\odot^{p-2}\theta)}{d^{\frac{p}{p-1}}||\theta||_\infty^{p-1}} \\ &\leq (p-1)d^{\frac{p}{p-1}}\frac{I}{||\theta||_\infty} \\ &\leq (p-1)d^{\frac{3p-2}{2p-2}}\frac{I}{||\theta||_2}.\end{aligned}$$

Together with the fact that $p \geq 1$, we bound the Hessian $\nabla^2 \mathcal{L}_{\text{adv}}(\theta)$ as:

$$\nabla^2 \mathcal{L}_{\text{adv}}(\theta) \leq \mathcal{L}_{\text{adv}}(\theta)\left[(1+c\sqrt{d})^2 + c(p-1)d^{\frac{3p-2}{2p-2}}\frac{1}{||\theta||_2}\right] I.$$

Note that the Hessian expression indicates that the objective is highly non-smooth around origin. However, as shown in Corollary C.1, starting from origin, θ^t always stays away from the origin with distance bounded below.

Using Taylor expansion, and by $\theta^{t+1} = \theta^t - \eta^t \nabla \mathcal{L}_{\text{adv}}(\theta^t)$, we have

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||)^2}{2} \max_{v \in [\theta^t, \theta^{t+1}]} \lambda\left(H(v)\right)_{\max}, \tag{32}$$

where $\lambda\left(H(v)\right)_{\max}$ denotes the largest eigenvalue of $H(v)$, and

$$H(v) = \mathcal{L}_{\text{adv}}(v)\left[(1+c\sqrt{d})^2 + c(p-1)d^{\frac{3p-2}{2p-2}}\frac{1}{||v||_2}\right] I.$$

Since $\eta^0 = 1$, by Corollary C.1, for any $t \geq 1$, we have $||\theta^t||_2 \geqslant \gamma_{2,q}$. Letting $m_p = (1+c\sqrt{d})^2 + c(p-1)d^{\frac{3p-2}{2p-2}}\frac{1}{\gamma_{2,q}}$, and since that $\mathcal{L}_{\text{adv}}(\theta)$ is a convex function, we obtain that

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||)^2}{2} m_p \max\{\mathcal{L}_{\text{adv}}(\theta^{t+1}, \mathcal{L}_{\text{adv}}(\theta^t)\}.$$

We then show by contradiction that we have $\mathcal{L}_{\text{adv}}(\theta^{t+1}) < \mathcal{L}_{\text{adv}}(\theta^t)$. Assume this is not the case, then we have:

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \left(1 - \frac{M(\eta^t)^2}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right)^{-1}\left(\mathcal{L}_{\text{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right)$$

However, if we choose η^t satisfying $\eta^t \leq \frac{2}{m_q \mathcal{L}_{\text{adv}}(\theta^t)}$, we have the right hand side of previous inequality strictly smaller than $\mathcal{L}_{\text{adv}}(\theta^t)$, which is clearly a constradiction. Hence when we choose $\eta^t \leq \frac{2}{m_q \mathcal{L}_{\text{adv}}(\theta^t)}$, we have $\mathcal{L}_{\text{adv}}(\theta^{t+1}) < \mathcal{L}_{\text{adv}}(\theta^t)$ and

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + \frac{(\eta^t ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||)^2}{2} m_p \mathcal{L}_{\text{adv}}(\theta^t). \tag{33}$$

Now by induction, if we choose $\eta^t = \eta \leq \frac{1}{m_q \mathcal{L}_{\text{adv}}(\theta^1)}$ for $t \geq 1$, then we have (33) holds for all $t \geq 1$. Note that we have an upper bound of $\mathcal{L}_{\text{adv}}(\theta^1)$, which is

$$\begin{aligned}\mathcal{L}_{\text{adv}}(\theta^1) &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i(x_i + \widetilde{\delta}_i)^\top \theta^1)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i(x_i + \widetilde{\delta}_i)^\top \theta_u^1 - y_i(x_i + \widetilde{\delta}_i)^\top \theta_\perp^1\right) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\gamma_{2,q}^2 + (1+c\sqrt{d})\right) = \exp\left(-\gamma_{2,q}^2 + (1+c\sqrt{d})\right),\end{aligned} \tag{34}$$

where $\widetilde{\delta}_i$ denotes the worst case perturbation to x_i, and θ_u^1 denotes projection of θ^1 onto $\mathrm{span}(u_{2,q})$, and $\theta_\perp$ denotes projection of θ^1 onto $\mathrm{span}(u_{2,q})^\perp$.

In summary, we have that if

$$\eta^t = \eta \leq \min\{\frac{1}{M_p}, 1\} \text{ for all } t \geq 1, \text{ where } M_p = m_p \exp\left(-\gamma_{2,q}^2 + (1 + c\sqrt{d})\right), \tag{35}$$

we have

$$\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\mathrm{adv}}(\theta^t) - \eta||\nabla\mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 + \frac{(\eta||\nabla\mathcal{L}_{\mathrm{adv}}(\theta^t)||)^2}{2} m_p \mathcal{L}_{\mathrm{adv}}(\theta^t) \tag{36}$$

$$\leq \mathcal{L}_{\mathrm{adv}}(\theta^t) - \frac{\eta}{2}||\nabla\mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \tag{37}$$

where the last inequality holds since $\eta m_p \mathcal{L}_{\mathrm{adv}}(\theta^t) \leq \eta m_p \mathcal{L}_{\mathrm{adv}}(\theta^1) \leq 1$. Now for any $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}
||\theta^{t+1} - \theta||_2^2 &= ||\theta^t - \theta||_2^2 - 2\eta^t \left\langle \nabla\mathcal{L}_{\mathrm{adv}}(\theta^t), \theta^t - \theta \right\rangle + (\eta^t)^2 ||\nabla\mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \\
&\leq ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta)\right) + (\eta^t)^2 ||\nabla\mathcal{L}_{\mathrm{adv}}(\theta^t)||_2^2 \\
&\leq ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta)\right) + 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta^{t+1})\right) \\
&= ||\theta^t - \theta||_2^2 - 2\eta^t \left(\mathcal{L}_{\mathrm{adv}}(\theta^{t+1}) - \mathcal{L}_{\mathrm{adv}}(\theta)\right),
\end{aligned}$$

where the first inequality holds by the convexity of $\mathcal{L}_{\mathrm{adv}}(\theta)$, and the second inequality holds by (37).

Summing up the above inequality from $s = 1$ to $t - 1$ and by $\eta^t = \eta \leq 1 = \eta^0$ together with $\mathcal{L}_{\mathrm{adv}}(\theta^{s+1}) \leq \mathcal{L}_{\mathrm{adv}}(\theta^s)$, we have

$$\mathcal{L}_{\mathrm{adv}}(\theta^t) - \mathcal{L}_{\mathrm{adv}}(\theta) \leq \frac{1}{2t\eta}||\theta^1 - \theta||_2^2 \leq \frac{1}{t\eta}\left(||\theta||_2^2 + ||\theta^1||_2^2\right) \tag{38}$$

Since θ is arbitrary, by choosing $\theta = \frac{\log(t)}{\gamma_{2,q}} \cdot u_{2,q}$, we have

$$||\theta||_2^2 + ||\theta^1||_2^2 \leq \frac{\log^2 t}{\gamma_{2,q}^2} + (1 + c\sqrt{d})^2,$$

and

$$\mathcal{L}_{\mathrm{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n \exp\left(-\min_{||\delta_i||_q \leq c} (z_i + \delta_i)^\top u_{2,q} \frac{\log t}{\gamma_{2,q}}\right) \leq \frac{1}{t},$$

which yields

$$\mathcal{L}_{\mathrm{adv}}(\theta^t) \leq \frac{1}{t} + \frac{1}{t\eta}\left(\frac{\log^2 t}{\gamma_{2,q}^2} + (1 + c\sqrt{d})^2\right) = \mathcal{O}\left(\frac{\log^2 t}{t\eta\gamma_{2,q}^2}\right). \tag{39}$$

□

Parameter Convergence: Intuition. Before we formally prove the implicit bias of GDAT, we provide some intuitions here for better understanding. We claim that $\overline{u}_\infty = \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2}$ is in the same direction as the solution to

$$\min_\theta \frac{1}{2}||\theta||_2 + \eta(c)||\theta||_p, \quad \text{s.t.} \quad z_i^\top \theta \geq 1, \forall i = 1, \ldots n. \tag{40}$$

Note that θ^t is a conic combination of $\{z_i - ca||\theta^t||_p\}_{i\in[n]}$, and $\partial||\theta^t||_p$ only depends on the direction of θ^t. Hence by normalizing the norm of θ^t and using $\lim_{t\to\infty}||\theta^t||_2 = \infty$, if the limit $\overline{u}_\infty = \lim_{t\to\infty}\frac{\theta^t}{||\theta^t||_2}$ exists, it satisfies the following condition under proper scaling that

$$\begin{aligned}
\theta &= \sum_{i=1}^n a_i(z_i - c\partial||\theta^t||_p), \\
\text{s.t.} \quad & a_i \geq 0, z_i^\top\theta \geq 1, \forall i = 1, \ldots n, \\
& a_i(z_i^\top\theta - 1) = 0, \forall i = 1, \ldots n.
\end{aligned}$$

Defining $a = (a_1, \ldots, a_n)$ and $(\widehat{\theta}, a) = \left((||\theta||_p c + 1)\theta, (||\theta||_p c + 1)a\right)$, it is easy to see that $(\widehat{\theta}, a)$ is a solution to the following system

$$\theta = \sum_{i=1}^n a_i(z_i - c\partial||\theta^t||_p), \tag{41}$$

$$\text{s.t.:} \quad a_i \geq 0, z_i^\top\theta \geq c||\theta||_p + 1, \forall i = 1, \ldots n. \tag{42}$$

$$a_i(z_i^\top\theta - c||\theta||_p - 1) = 0, \forall i = 1, \ldots n. \tag{43}$$

Notice that the above set of equations (41)-(43) is exactly the first-order KKT condition of the following optimization problem

$$\min_{\theta} \frac{1}{2}||\theta||_2^2 \quad \text{s.t.} \quad z_i^\top \theta \geq c||\theta||_p + 1, \forall i = 1, \dots n. \tag{44}$$

(44) has a robust reformulation as maximizing the ℓ_2-norm margin under the worse case ℓ_q-norm perturbation bounded by c that

$$\min_{\theta} \frac{1}{2}||\theta||_2^2 \quad \text{s.t.} \quad \min_{||\delta_i||_q \leq c} (z_i + \delta_i)^\top \theta \geq 1, \forall i = 1, \dots n,$$

or equivalently

$$\max_{\theta} \min_{i=1,\dots,n} \min_{||\delta_i||_\infty \leq c} \frac{y_i(x_i + \delta_i)^\top \theta}{||\theta||_q}. \tag{45}$$

We note that (45) is a Support Vector Machine problem over an uncoutable data set that is generated by norm-bounded perturbation $\mathcal{S}_(c, q) = \{(x, y) : \text{ where } \exists i \in [n], ||x - x_i||_q \leq c, y = y_i\}$. By the separability and $c < \gamma_q$, we have that $\mathcal{S}_(c, q)$ is well defined.

By the first-order KKT condition we have that (44) is equivalent to

$$\min_{\theta} ||\theta||_2 + \eta(c)||\theta||_p \quad \text{s.t.} \quad z_i^\top \theta \geq 1, \forall i = 1, \dots n.$$

for some proper $\eta(c)$ that depends on c. Hence in summary, if $\overline{u}_\infty = \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2}$ exists, it is in the same direction as the solution to the mixed (ℓ_2, ℓ_1)-norm max margin solution of (40).

Claim: In general, for ℓ_q-norm perturbation bounded by c, θ^t converges in direction to the solution to

$$\min_{\theta} \frac{1}{2}||\theta||_2^2 \quad \text{s.t.} \quad \min_{||\delta_i||_q \leq c} (z_i + \delta_i)^\top \theta \geq 1, \forall i = 1, \dots n.$$

or

$$\min_{\theta} ||\theta||_2 + \eta(c)||\theta||_p \quad \text{s.t.} \quad z_i^\top \theta \geq 1, \forall i = 1, \dots n.$$

for some proper $\eta(c)$ that depends on c.

Proof of Theorem 3.4. Recall that in Theorem 3.3 we showed in (36) the following recursion

$$\begin{aligned}
\mathcal{L}_{\text{adv}}(\theta^{t+1}) &\leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + \frac{(\eta||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||)^2}{2} m_p \mathcal{L}_{\text{adv}}(\theta^t) \\
&\leq \exp\left(-\eta \frac{||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2}{\mathcal{L}_{\text{adv}}(\theta^t)} + m_p \frac{\eta^2}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right) \\
&\leq \exp\left(-\eta\gamma_{2,q}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2 + m_p \frac{\eta^2}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right).
\end{aligned}$$

where the last inequality holds by Lemma C.1.

Applying the previous inequality recursively from $s = 1$ to $t - 1$, we have

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \exp\left(-\eta\gamma_{2,q} \sum_{s=1}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 + \sum_{s=1}^{t-1} m_p \frac{\eta^2}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2\right).$$

Now since in the proof of Theorem 3.3 we showed that $\eta m_p < 1$ (35), combining the above inequality this with (37), we have

$$\sum_{s=1}^{t-1} m_p \frac{\eta^2}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2 = \sum_{s=1}^{t-1} \frac{\eta}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2 = \mathcal{L}_{\text{adv}}(\theta^1) - \mathcal{L}_{\text{adv}}(\theta^t) \leq \mathcal{L}_{\text{adv}}(\theta^1).$$

Combining this inequality with the upper bound on $\mathcal{L}_{\text{adv}}(\theta^1)$ in (34), we have

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \exp\left(-\eta\gamma_{2,q} \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 - \gamma_{2,q}^2 + (1 + c\sqrt{d})\right).$$

Now for all $i \in [n]$, we have:

$$\exp\left(-\min_{||\delta_i||_q \leq c} y_i(x_i + \delta_i)^\top \theta^t\right) \leq n \exp\left(-\eta\gamma_{2,q} \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 - \gamma_{2,q}^2 + (1 + c\sqrt{d})\right),$$

which yields

$$\min_{||\delta_i||_q \le c} y_i(x_i+\delta_i)^\top \theta^t \ge \eta\gamma_{2,q}\sum_{s=0}^{t-1}||\nabla\mathcal{L}_{\text{adv}}(\theta^s)||_2 + \gamma_{2,q}^2 - (1+c\sqrt{d}) - \log n.$$

Dividing both sides by $||\theta||_2$, and since $\lim_{t\to\infty}\mathcal{L}_{\text{adv}}(\theta^t)=0$, we have $\lim_{t\to\infty}||\theta^t||_2=\infty$. Hence,

$$\lim_{t\to\infty}\min_{||\delta_i||_q\le c} y_i(x_i+\delta_i)^\top \frac{\theta^t}{||\theta^t||_2} \ge \lim_{t\to\infty}\eta\gamma_{2,q}\sum_{s=0}^{t-1}\frac{||\nabla\mathcal{L}_{\text{adv}}(\theta^s)||_2}{||\theta^t||_2} - \frac{1+c\sqrt{d}+\log n}{||\theta^t||_2} \tag{46}$$

$$\ge \gamma_{2,q},$$

where the last inequality holds by $||\theta^t||_2 \le \eta\sum_{s=0}^{t-1}||\nabla\mathcal{L}_{\text{adv}}(\theta^s)||_2$.

Hence in summary, we have

$$\min_{||\delta_i||_q\le c} y_i(x_i+\delta_i)^\top \lim_{t\to\infty}\frac{\theta^t}{||\theta^t||_2} \ge \gamma_{2,q}.$$

Hence, we have $\lim_{t\to\infty}\theta^t/||\theta^t||_2$ is a solution to (10), but notice that the solution to (10) is unique since a multiple of its optimal solution would be the solution to (8) that

$$\min_{\theta\in\mathbb{R}^d}\frac{1}{2}||\theta||_2^2 \quad \text{s.t.} \quad \min_{\delta_i\in\Delta_i(q)} y_i(x_i+\delta_i)^\top\theta \ge 1, \forall i=1,\dots,n,$$

which is a convex program with strongly convex objective. By this fact, we conclude that $\lim_{t\to\infty}\frac{\theta^t}{||\theta^t||_2} = u_{2,q}$. To further get the rate of convergence, we use the convergence of adversarial risk in (39), and establish the lower bound on $||\theta^t||_2$: $||\theta^t||_2 = \Omega(\log t)$. Combining this with (46), the claim follows immediately. □

D ℓ_∞-NORM PERTURBATION

Recall that the robust SVM against ℓ_∞-norm perturbation parameterized by c is formulated as

$$\gamma_{2,\infty} = \max_\theta \min_{i=1,\dots,n}\min_{||\delta_i||_\infty\le c}\frac{y_i(x_i+\delta_i)^\top\theta}{||\theta||_2}, \tag{47}$$

and its associated max-margin classifier is

$$u_{2,\infty} = \operatorname*{argmax}_{||\theta||_2=1}\min_{i=1,\dots,n}\min_{||\delta_i||_\infty\le c} y_i(x_i+\delta_i)^\top\theta.$$

It is easy to see that for $c<\gamma_\infty$, both $\gamma_{2,\infty}$ and $u_{2,\infty}$ are well defined, and $\gamma_{2,\infty}>0$.

Before showing parameter convergence, we first prove that the adversarial risk goes to zero. To avoid analyzing ℓ_∞-perturbation directly, which can go messy. For $\lambda>0$, we define a smooth approximation of ℓ_1-norm that

$$h_\lambda(\theta_j) = \sqrt{\theta_j^2+\lambda}, \quad \text{and} \quad H_\lambda(\theta_j) = \sum_{j=1}^d h_\lambda(\theta_j).$$

Note that as $\lambda\to 0$, $H_\lambda(\theta)\to||\theta||_1$ uniformly. We then define a smoothified version of (47) that we let perturbation set be $\Delta_i(\lambda)=\{\delta:\forall j\in[d], |\delta_j|\le c\frac{h_\lambda(\theta_j)}{|\theta_j|}\}$, and the corresponding $\gamma_{2,\infty}$ and $u_{2,\infty}$ become

$$\gamma_{2,\lambda} = \max_\theta\min_{i=1,\dots,n}\min_{\delta_i\in\Delta_i(\lambda)}\frac{y_i(x_i+\delta_i)^\top\theta}{||\theta||_2}, \tag{48}$$

$$u_{2,\lambda} = \operatorname*{argmax}_{||\theta||_2=1}\min_{i=1,\dots,n}\min_{\delta_i\in\Delta_i(\lambda)} y_i(x_i+\delta_i)^\top\theta. \tag{49}$$

Note that the Hausdorff distance between $\Delta_i(\lambda)$ and $\{\delta:||\delta||_\infty\le c\}$ converges to 0 as λ goes to 0. It can be seen that when $\lambda\to 0$, the smoothified problem (48) reduces to (47). That is, $\lim_{\lambda\to 0}\gamma_{2,\lambda}=\gamma_{2,\infty}$ and $\lim_{\lambda\to 0}u_{2,\lambda}=u_{2,\infty}$.

Theorem D.1. *Let perturbation set be $\Delta_i(\lambda)=\{\delta:\forall j\in[d], |\delta_j|\le c\frac{h_\lambda(\theta_j)}{|\theta_j|}\}$, and let its associated adversarial risk be*

$$\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^n\max_{\delta_i\in\Delta_i(\lambda)}\exp\left(-y_i(x_i+\delta_i)^\top\theta\right).$$

For $c<\gamma_{2,\lambda}$, letting $\eta=\frac{1}{(1+2c\lambda^{-1/2})^2}$, we have

$$\mathcal{L}_{\text{adv}}(\theta^t) \le \mathcal{O}\left(\frac{\log^2 t(1+2c\lambda^{-1/2})^2}{t\gamma_{2,\lambda}}\right).$$

Proof. By the definition of perturbation set that $\Delta_i = \{\delta : \forall j \in [d], |\delta_j| \leq c\frac{h_\lambda(\theta_j)}{|\theta_j|}\}$, we have

$$\mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i x_i^\top \theta + cH_\lambda(\theta)\right).$$

By some simple calculation, we have

$$\nabla H_\lambda(\theta) = \left(\frac{\theta_1}{\sqrt{\theta_1^2+\lambda}}, \ldots, \frac{\theta_d}{\sqrt{\theta_d^2+\lambda}}\right), \nabla^2 H_\lambda(\theta) = diag\left(\frac{\lambda}{(\theta_1^2+\lambda)^{3/2}}, \ldots, \frac{\lambda}{(\theta_d^2+\lambda)^{3/2}}\right).$$

Then, it holds that

$$\nabla \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + cH_\lambda(\theta)\right)\left(-z_i + c\nabla H_\lambda(\theta)\right),$$

$$\nabla^2 \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-z_i^\top \theta + cH_\lambda(\theta)\right)\left(z_i z_i^\top + c^2 \nabla H_\lambda(\theta)\nabla H_\lambda(\theta)^\top - 2z_i^\top \nabla H_\lambda(\theta) + c\nabla^2 H_\lambda(\theta)\right).$$

It can be verified that $\nabla^2 \mathcal{L}_{\text{adv}}(\theta) \leq (1+\frac{2c}{\sqrt{\lambda}})^2 \mathcal{L}_{\text{adv}}(\theta) I$. By Talyer expansion, we have

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + (1+\frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2}{2} \max\{\mathcal{L}_{\text{adv}}(\theta^t), \mathcal{L}_{\text{adv}}(\theta^{t+1}\}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2. \tag{50}$$

Now we show that $\mathcal{L}_{\text{adv}}(\theta^{t+1}) \geq \mathcal{L}_{\text{adv}}(\theta^t)$ does not hold when $\eta \leq \frac{1}{(1+2c\lambda^{-1/2})^2 \mathcal{L}_{\text{adv}}(\theta^t)}$. Suppose the contrary holds. By (50), we have

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \left(1 - \frac{\eta^2 ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2}{2}(1+\frac{2c}{\sqrt{\lambda}})^2\right)^{-1}\left(\mathcal{L}_{\text{adv}}(\theta^t) - \eta||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right) < \mathcal{L}_{\text{adv}}(\theta^t).$$

where the last inequality holds by $\eta = \frac{1}{(1+2c\lambda^{-1/2})^2 \mathcal{L}_{\text{adv}}(\theta^t)}$. Hence we obtain a contradiction.

Note that $\mathcal{L}_{\text{adv}}(\theta^0) = 1$, and if $\eta \leq \frac{1}{(1+2c\lambda^{-1/2})^2}$, $\eta \leq \frac{1}{(1+2c\lambda^{-1/2})^2 \mathcal{L}_{\text{adv}}(\theta^t)}$ holds for $t = 0$, and $\mathcal{L}_{\text{adv}}(\theta^1) \leq 1$. Consequently, we can inductively show that $\mathcal{L}_{\text{adv}}(\theta^t) \leq 1$ for all t, and $\eta \leq \frac{1}{(1+2c\lambda^{-1/2})^2 \mathcal{L}_{\text{adv}}(\theta^t)}$ always holds if we let $\eta = \frac{1}{(1+2c\lambda^{-1/2})^2}$.

By the choice of η, we obtain the following recursion taht

$$\mathcal{L}_{\text{adv}}(\theta^{t+1}) \leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + (1+\frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2 \mathcal{L}_{\text{adv}}(\theta^t)}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 \tag{51}$$

$$= \mathcal{L}_{\text{adv}}(\theta^t) - \frac{\eta}{2}||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2. \tag{52}$$

Using the previous recursion we have that for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
||\theta^{t+1} - \theta||_2^2 &= ||\theta^t - \theta||_2^2 - 2\eta\left\langle \nabla \mathcal{L}_{\text{adv}}(\theta^t), \theta^t - \theta\right\rangle + \eta^2||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 \\
&\leq ||\theta^t - \theta||_2^2 - 2\eta\left(\mathcal{L}_{\text{adv}}(\theta^t) - \mathcal{L}_{\text{adv}}(\theta)\right) + 2\eta\left(\mathcal{L}_{\text{adv}}(\theta^t) - \mathcal{L}_{\text{adv}}(\theta^{t+1})\right) \\
&= ||\theta^t - \theta||_2^2 - 2\eta\left(\mathcal{L}_{\text{adv}}(\theta^{t+1}) - \mathcal{L}_{\text{adv}}(\theta)\right),
\end{aligned}$$

where the second inequality holds by convexity and (51). Summing up the previous inequality from $s = 0$ to $s = t-1$ and by $\mathcal{L}_{\text{adv}}(\theta^{s+1}) \leq \mathcal{L}_{\text{adv}}(\theta^s)$, we have

$$\mathcal{L}_{\text{adv}}(\theta^t) - \mathcal{L}_{\text{adv}}(\theta) \leq \frac{1}{2t\eta}||\theta||_2^2.$$

Taking $\theta = \frac{\log t}{\gamma_{2,\lambda}} u_{2,\lambda}$, we have

$$\begin{aligned}
\mathcal{L}_{\text{adv}}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \max_{\delta_i \in \Delta_i(\lambda)} \exp\left(-y_i(x_i+\delta_i)^\top \theta\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \max_{\delta_i \in \Delta_i(\lambda)} \exp\left(-y_i(x_i+\delta_i)^\top \frac{\log t}{\gamma_{2,\lambda}} u_{2,\lambda}\right) \leq \frac{1}{t}.
\end{aligned}$$

where the last inequality holds by $\max_{\delta_i \in \Delta_i} y_i(x_i+\delta_i)^\top u_{2,\lambda} \geq \gamma_{2,\lambda}$. Hence we obtain

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \frac{1}{t} + \frac{\log^2 t}{t\gamma_{2,\lambda}\eta} = \mathcal{O}\left(\frac{\log^2 t(1+2c\lambda^{-1/2})^2}{t\gamma_{2,\lambda}}\right).$$

□

Before showing parameter convergence, we need the following lemma which is a generalization of Lemma 10 in Gunasekar et al. (2018a), but with much simpler proof.

Lemma D.1. *Fix $c < \gamma_{2,\lambda}$, for any $\theta \in RR^d$, we have*

$$||\nabla \mathcal{L}_{\text{adv}}(\theta)||_2 \geq \mathcal{L}_{\text{adv}}(\theta)\gamma_{2,\lambda}.$$

Proof.

$$-\nabla \mathcal{L}_{\text{adv}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i \widetilde{x}_i\right) y_i \widetilde{x}_i.$$

where $\widetilde{x}_i = \text{argmin}_{x_i' - x_i \in \Delta_i(\lambda)} \, y_i (x_i')^\top \theta$. Then by the definition of $\gamma_{2,\lambda}$ and $u_{2,\lambda}$ (48), we have

$$\langle y_i \widetilde{x}_i, u_{2,\lambda} \rangle \geq \gamma_{2,\lambda}$$

From which we obtain $\langle -\nabla \mathcal{L}_{\text{adv}}(\theta), u_{2,\lambda} \rangle \geq \mathcal{L}_{\text{adv}}(\theta)\gamma_{2,\lambda}$, the claim follows by Cauchy-Schwarz inequality. □

Theorem D.2. *Under the same setting as in Theorem D.1, we have*

$$\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_{2,\lambda}.$$

Proof. Recall that in Theorem D.1 we showed in (51) that

$$\begin{aligned}
\mathcal{L}_{\text{adv}}(\theta^{t+1}) &\leq \mathcal{L}_{\text{adv}}(\theta^t) - \eta ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 + (1 + \frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2 \mathcal{L}_{\text{adv}}(\theta^t)}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2 \\
&\leq \exp\left(-\eta \frac{||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2}{\mathcal{L}_{\text{adv}}(\theta^t)} + (1 + \frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right) \\
&\leq \exp\left(-\eta\gamma_{2,\lambda} ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2 + (1 + \frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^t)||_2^2\right),
\end{aligned}$$

where the last inequality holds by Lemma D.1. Applying the previous inequality recursively from $s = 0$ to $t - 1$, we have

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \exp\left(-\eta\gamma_{2,\lambda} \sum_{t=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 + \sum_{s=0}^{t-1} (1 + \frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2\right).$$

Now by (51), we have

$$\sum_{s=0}^{t-1} (1 + \frac{2c}{\sqrt{\lambda}})^2 \frac{\eta^2}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2 = \sum_{s=0}^{t-1} \frac{\eta}{2} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2^2 = \mathcal{L}_{\text{adv}}(\theta^0) - \mathcal{L}_{\text{adv}}(\theta^t) \leq 1,$$

which yields

$$\mathcal{L}_{\text{adv}}(\theta^t) \leq \exp\left(-\eta\gamma_{2,\lambda} \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 + 1\right).$$

Next for all $i \in [n]$, we have

$$\begin{aligned}
\exp\left(-\min_{\delta_i \in \Delta_i(\lambda)} y_i (x_i + \delta_i)^\top \theta^t\right) &= \exp\left(-y_i x_i^\top \theta + cH_\lambda(\theta^t)\right) \\
&\leq n \exp\left(-\eta\gamma_{2,\lambda} \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 + 1\right),
\end{aligned}$$

which implies

$$\min_{\delta_i \in \Delta_i(\lambda)} y_i (x_i + \delta_i)^\top \theta^t \geq \eta\gamma_{2,\lambda} \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2 - 1 - \log n.$$

Dividing both sides by $||\theta||_2$, and since $\lim_{t\to\infty} \mathcal{L}_{\text{adv}}(\theta^t) = 0$, we have $\lim_{t\to\infty} ||\theta^t||_2 = \infty$. Hence,

$$\lim_{t\to\infty} \min_{\delta_i \in \Delta_i(\lambda)} y_i (x_i + \delta_i)^\top \frac{\theta^t}{||\theta^t||_2} \geq \lim_{t\to\infty} \eta\gamma_{2,\lambda} \sum_{s=0}^{t-1} \frac{||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2}{||\theta^t||_2} - \frac{1 + \log n}{||\theta^t||_2} \geq \gamma_{2,\lambda},$$

where the last inequality holds by $||\theta^t||_2 \leq \eta \sum_{s=0}^{t-1} ||\nabla \mathcal{L}_{\text{adv}}(\theta^s)||_2$.

In summary, we have

$$\min_{\delta_i \in \Delta_i(\lambda)} y_i(x_i + \delta_i)^\top \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} \geq \gamma_{2,\lambda}.$$

Hence $\lim_{\theta^t} ||\theta^t||_2$ is a solution to (48). Note that the solution to (48) is unique since it is equivalent to

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}||\theta||_2^2 \quad \text{s.t.} \quad \min_{\delta_i \in \Delta_i(\lambda)} y_i(x_i + \delta_i)^\top \theta \geq 1, \forall i = 1, \ldots, n.$$

We thus conclude that $\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_{2,\lambda}$. □

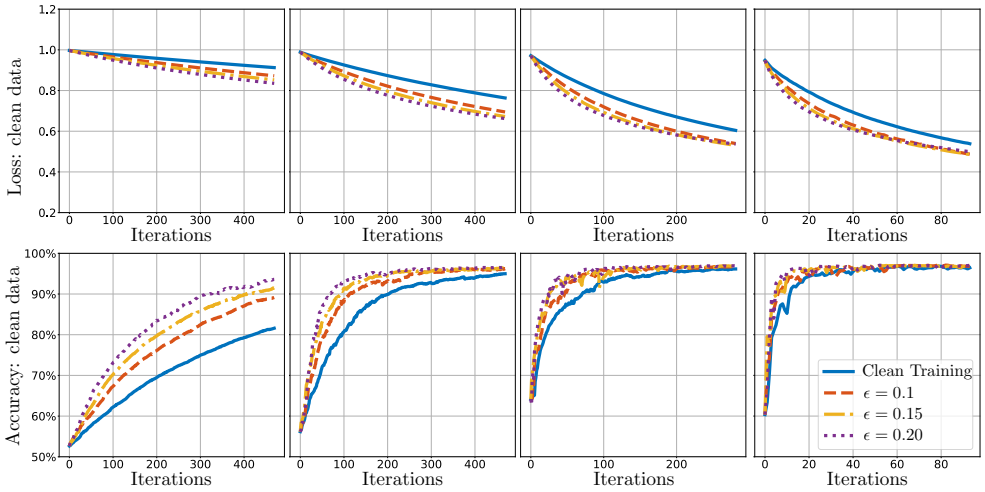
To summarize, we have shown that for all $\lambda > 0$, $\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_{2,\lambda}$. The ℓ_∞-norm perturbation corresponds to the case when $\lambda \to 0$, it is natural to conclude that for ℓ_∞ perturbation, we have $\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} = u_{2,\infty}$. The discussion for $q = 1$ follows similar argument, hence we omit the details here.

E Additional Experiments on Perturbation Level and Speed-up

We provide additional experiments on the connection of perturbation level c and the speed-up effect of adversarial training for neural networkds. We run GDAT with ℓ_∞-norm perturbation. The setup of the experiments is exactly the same as the setup in Section 4. We will vary the perturbation level c used in GDAT algorithm in $\{0.1, 0.15, 0.2\}$.

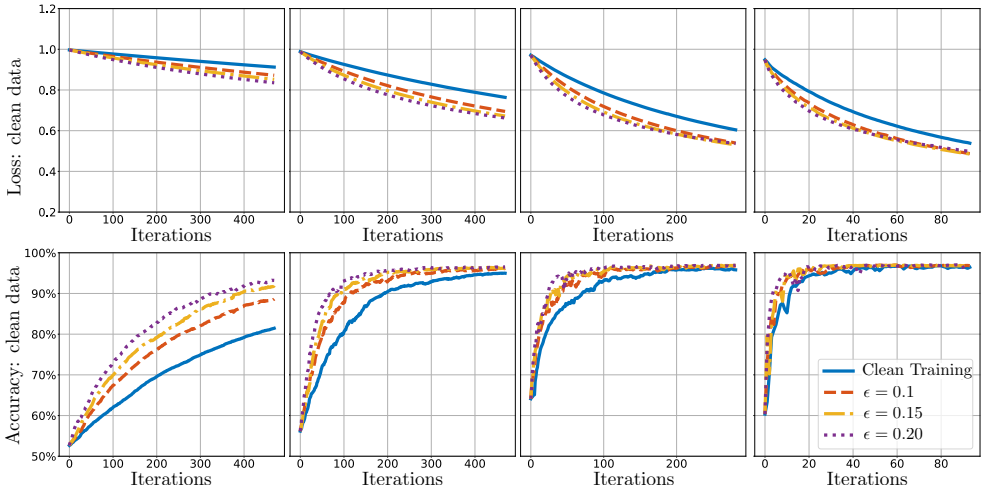
From Figure 3 we could see that GDAT indeed accelerates convergence of loss and accuracy on clean training samples. Moreoever, the acceleration effect is stronger when we use larger perturbation level, and this relationship is consistent across different width of hidden layer.

Similar speed-up effects on the test loss and test accuracy evaluated on clean test samples are also observed for GDAT. From Figure 4, we see that the speed-up effects become stronger when we use larger perturbation level, and this relationship is consistent across different width of hidden layer. Traditionally, the benefit of adversarial training is understood as two fold: 1. it improves the robustness of the learning algorithm, i.e., the solution has better loss toward adversarilly perturbed seample; 2. it has better generalization ability. Our experiments demonstrate a third property of adverserial training that is not known in literature before, i.e., adversarial training accelerates convergence.



(a). Width = 64×64 (b). Width = 128×128 (c). Width = 256×256 (d). Width = 512×512

Figure 3: GDAT with Different Perturbation Level: Clean Training Loss



(a). Width $= 64 \times 64$ (b). Width $= 128 \times 128$ (c). Width $= 256 \times 256$ (d). Width $= 512 \times 512$

Figure 4: GDAT with Different Perturbation Level: Clean Test Loss