AdAUC: End-to-end Adversarial AUC Optimization Against Long-tail Problems

Wenzheng Hou ¹ ²  Qianqian Xu ¹  Zhiyong Yang ²  Shilong Bao ³ ⁴  Yuan He ⁵  Qingming Huang ¹ ² ⁶ ⁷

Abstract

It is well-known that deep learning models are vulnerable to adversarial examples. Existing studies of adversarial training have made great progress against this challenge. As a typical trait, they often assume that the class distribution is overall balanced. However, long-tail datasets are ubiquitous in a wide spectrum of applications, where the amount of head class instances is larger than the tail classes. Under such a scenario, AUC is a much more reasonable metric than accuracy since it is insensitive toward class distribution. Motivated by this, we present an early trial to explore adversarial training methods to optimize AUC. The main challenge lies in that the positive and negative examples are tightly coupled in the objective function. As a direct result, one cannot generate adversarial examples without a full scan of the dataset. To address this issue, based on a concavity regularization scheme, we reformulate the AUC optimization problem as a saddle point problem, where the objective becomes an instance-wise function. This leads to an end-to-end training protocol. Furthermore, we provide a convergence guarantee of the proposed algorithm. Our analysis differs from the existing studies since the algorithm is asked to generate adversarial examples by calculating the gradient of a min-max problem. Finally, the extensive experimental results show the performance and robustness of our algorithm in three long-tail datasets.

1. Introduction

Deep learning has recently achieved significant progress on various machine learning tasks, such as computer vision (Voudodimos et al., 2018) and natural language processing (Strubell et al., 2019; Sorin et al., 2020). However, recent work shows that deep learning models are vulnerable to adversarial attack (Szegedy et al., 2014; Biggio et al., 2013). For example, images with human imperceptible perturbations (i.e., adversarial examples) can easily fool even the well-trained models. The existence of adversarial examples has raised big security threats to deep neural networks, which impels extensive efforts to improve the adversarial robustness (Madry et al., 2018; Zhang et al., 2019b). For example, one can resist the adversarial examples by means of adversarial training (AT). Specifically, AT could be formulated as a min-max problem, where the inner maximization problem is employed to generate adversarial examples, and the outer minimization problem is used to learn the model under the adversarial noise. In this way, AT could be easily applied to most of the modern architectures in deep learning, making it one of the most effective measures against adversarial attack.

The prior art of adversarial training methods focuses on balanced benchmark datasets. On top of this, the learning objective is to increase the overall accuracy. However, real-
world datasets usually exhibit a long-tail distribution that
the proportion of the majority classes examples significantly
dominate the others. For such long-tail problems, accuracy
(ACC) is considered to be a less appropriate performance
metric than another metric named AUC (Area Under the
ROC Curve). Specifically, AUC is the probability of observ-
ing a positive instance with a higher score than a negative
one. It is well-known to be insensitive to class distributions
and costs (Fawcett, 2006; Hand & Till, 2001). Comparing
AUC with ACC, we then ask:

Can we improve the adversarial robustness of AUC by em-
ploying the traditional ACC-based methods?

Unfortunately, the answer might be negative. As shown in
Fig.1, the adversarial examples generated by minimizing
ACC (ACC drops from 0.8 to 0.6), may fail to attack AUC
(AUC here increases from 0.714 to 0.762). In this sense, the
model trained on such adversarial examples cannot improve
the adversarial robustness of AUC. Therefore, more atten-
tion should be paid to AUC when studying the adversarial
robustness against long-tail problems.

Inspired by this fact, we present a very early trial to study
adversarial training in AUC optimization with an end-to-
end framework.

Existing AT methods can be easily implemented in an end-
to-end manner since the inner maximization problem for
generating the adversarial examples can be solved instance-
wisely. However, this is not the case for AUC optimization.
Specifically, in the expression of AUC, every positive in-
stance is coupled with all the negative instances and vice
versa. For a binary class classification problem, this means
that we need to spend $O(n^{-} \cdot T)$ time for generating the
adversarial example of a positive instance and $O(n^{+} \cdot T)$
for a negative one, where $n^{-}$, $n^{+}$ are the number of positive
and negative instances, and $T$ is complexity for calculating
the gradient for a single positive-negative instance pair. In
this sense, we can hardly implement such a naive training
method on top of even the simplest deep learning fram-
work.

To solve the challenge, this paper proposes an end-to-end ad-
versarial AUC optimization framework with a convergence
guarantee. Specifically, our contribution is as follows:

First, based on a reformulation technique and a concavity
regularizer, we show that the original problem is equivalent
to a min-max problem where the objective function can be
expressed in an instance-wise manner.

Second, we propose an AT algorithm to optimize the min-
max problem, where we alternately invoke a projected-
gradient-descent-like protocol to generate the adversarial
examples, and a stochastic gradient descent-ascent protocol
to train the model parameters. Meanwhile, we also present
a convergence analysis to show the correctness of our al-
gorithm. The proof here is non-trivial since we have to simulta-
neously estimate the gradient of the min player and the max player.

Finally, we conduct a series of empirical analyses of our
proposed algorithm on long-tail datasets. The results demon-
strate the effectiveness of our proposed method.

2. Related Work

2.1. AUC Optimization

As a motivating study, (Cortes & Mohri, 2003) investigates
the inconsistency between AUC maximization and error
rate minimization, which shows the necessity to study di-
rect AUC optimization methods. After that, a series of
algorithms are designed for off-line AUC optimization (Her-
schtal & Raskutti, 2004; Joachims, 2005). To extend the
scalability of AUC optimization, researchers start to explore
the online and stochastic optimization extensions of the AUC
maximization problem. (Zhao et al., 2011) makes the
first attempt for this direction based on the reservoir sam-
pling technique. (Gao et al., 2013) proposes a one-pass AUC
optimization algorithm based on the squared surrogate loss.
After that, (Ying et al., 2016) reformulates the minimiza-
tion problem of the pairwise square loss into an equivalent
stochastic saddle point problem, where the objective func-
tion could be expressed in an instance-wise manner. On top
of the reformulation framework, (Natole et al., 2018) pro-
poses an accelerated version with a faster convergence rate
and (Liu et al., 2019) explores its extension in deep neural
networks. Meanwhile, many researchers provide theoretical
guarantees for AUC optimization algorithms from differ-
ent aspects, such as generalization analysis (Agarwal et al.,
2005; Clémenc¸on et al., 2008; Usunier et al., 2005) and
Beyond the optimization algorithms and theoretical supports
for AUC, in practice, AUC optimization demonstrates its
effectiveness in various class-imbalanced tasks, such as dis-
 ease prediction (Westcott et al., 2019; Gola et al., 2020; Ren
et al., 2018), rare event detection (Feizi, 2020; Robles et al.,
2020) and etc.

Compared with the existing study, we present a very early
trial for the adversarial training problem.

2.2. Adversarial Training

For a long time, machine learning models have proved vul-
nerable to adversarial examples (Biggio et al., 2013;
Szegedy et al., 2014; Goodfellow et al., 2014). Numerous
defenses have been proposed to address the security con-
cern raised by the issue (Athalye & Carlini, 2018; Athalye
et al., 2018). Among such studies, adversarial training is one
of the most popular methods (Kurakin et al., 2017; Madry
AdAUC: End-to-end Adversarial AUC Optimization Against Long-tail Problems

et al., 2018; Zhang et al., 2019b). The majority of studies in this direction follows the min-max formulation proposed in (Madry et al., 2018), which so far has been improved in various way (Shafahi et al., 2020; Cai et al., 2018; Tramèr et al., 2017; Pang et al., 2019; Wang et al., 2019b; Zhang et al., 2020; Maini et al., 2020; Tramèr & Boneh, 2019). Furthermore, due to the heavy computational burden of AT, accelerating the training procedure of AT becomes increasingly urgent. Recently, there has been a new wave to explore the acceleration of AT, which includes reusing the computations (Shafahi et al., 2019; Zhang et al., 2019a), adaptive adversarial steps (Wang et al., 2019a) and one-step training (Wong et al., 2019). Besides the practical improvements, there are also some recent advances in theoretical investigations from the perspective of optimization (Wang et al., 2019a; Bai et al., 2022), generalization (Xing et al., 2021; Tu et al., 2019), and consistency (Bao et al., 2020).

In this paper, we will present an AT algorithm on top of the AUC optimization. As shown in the introduction, the complicated expression of AUC brings new elements into our model formulation and theoretical analysis.

3. Preliminaries

In this section, we briefly introduce the AUC optimization problem and the adversarial training framework.

3.1. AUC Optimization Problem

Let $X$ be the feature set. Based on (Hanley & McNeil, 1982), AUC of a scoring function $h_\theta : X \rightarrow [0, 1]$ is equivalent to the probability that a positive instance is predicted with a higher score compared to a negative instance:

$$\text{AUC}(h_\theta) = \Pr \left( h_\theta(x^+) \geq h_\theta(x^-) \middle| y^+ = 1, y^- = 0 \right),$$

where $(x^+, y^+)$ and $(x^-, y^-)$ represent positive and negative examples, respectively, and $\theta$ is the model parameters. By employing a differentiable loss $\ell$ as the surrogate loss, the unbiased estimation of $\text{AUC}(h_\theta)$ could be expressed as:

$$\hat{\text{AUC}}(h_\theta) = 1 - \frac{1}{n^+n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \frac{\ell(h_\theta(x_i^+) - h_\theta(x_j^-))}{n^+n^-},$$

where $n^+$ and $n^-$ denote the number of positive and negative examples, respectively. Then AUC maximization problem is equivalent to the following minimization problem:

$$\text{OP0} \quad \min_{\theta} \mathcal{L}(\theta, x, y) := \frac{1}{n^+n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \ell(h_\theta(x_i^+) - h_\theta(x_j^-)).$$

3.2. Adversarial Training Framework

Adversarial training is one of the most effective defensive strategies against adversarial examples (Goodfellow et al., 2015; Madry et al., 2018), the key idea of which is to directly optimize the model performance based on the perturbed examples. Generally speaking, the adversarial training framework can be formalized as

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{|\delta_i|_\infty \leq \epsilon} \ell(h_\theta(x_i^0 + \delta_i), y_i).$$

Here $\delta_i$ is the perturbation on clean feature vector $x_i^0$, and $x_i = x_i^0 + \delta_i$ is the resulting adversarial example for the instance $(x_i^0, y_i), i = 1, 2, \cdots, n$. The inner maximization problem generates such adversarial examples by trying to hurt the model performance (by maximizing the loss $\ell(h_\theta(x_i^0 + \delta_i), y_i)$). The constraint $|\delta_i|_\infty \leq \epsilon$ makes sure that the adversarial perturbation is small enough to be imperceptible. In this sense, the adversarial example lives in $X_i = \{x \mid \|x - x_i^0\|_\infty \leq \epsilon \}$. Finally, the outer minimization problem is to find a robust model that can resist the adversarial perturbation.

For the inner maximization problem, K-PGD (Madry et al., 2018) is a widely used attack method that perturbs the clean examples $x_i^0$ iteratively with a total of $K$ steps. At the end of each iteration, the example will be projected to the $\epsilon$-ball of $x_i^0$. Specifically, the adversarial examples generated in $k + 1$ step are as follows:

$$x_i^{k+1} = \text{Proj}(x_i^k + \beta \cdot \text{sign}(\nabla_x \ell(h_\theta(x_i^k), y_i))),$$

where $\text{Proj}$ is the projection function, and $\beta$ is the step size. For the outer minimization problem, gradient descent is
we need to reformulate the optimization problem. A more detailed introduction of adversarial attack methods is shown in the Appendix B.

4. Methodology

Before entering into the methodology, we summarize some useful notations in Tab.3.1 to make our argument easier to follow.

4.1. Reformulation of Optimization Problem

A naive idea to perform AUC adversarial training is to directly combine (OP0) with the standard AT framework (Madry et al., 2018), resulting in the following problem:

\[
\min_{\theta, \delta_1, \delta_2, \ldots, \delta_n} \sum_{i=1}^{n^n} \sum_{j=1}^{n^n} \ell(h_\theta(x_i^+ + \delta_i) - h_\theta(x_j^- + \delta_j))
\]

where

\[
g(\theta, a, b, \alpha, (x_i, y_i)) = (1 - p)(h_\theta(x_i) - a)^2\mathbb{I}_{[y_i=1]} + p(h_\theta(x_i) - b)^2\mathbb{I}_{[y_i=0]} + 2(1 + \alpha)\left(p h_\theta(x_i)\mathbb{I}_{[y_i=0]} - (1 - p) h_\theta(x_i)\mathbb{I}_{[y_i=1]}\right) - p(1 - p)\alpha^2.
\]

Thus, we can verify that \( g \) is \( \mu \)-strongly concave w.r.t. \( \alpha \) in \( \Omega_\alpha \), i.e., for any \( \alpha_1, \alpha_2 \in \Omega_\alpha \), it holds that

\[
g(\theta, a, b, \alpha_1, (x, y)) \leq g(\theta, a, b, \alpha_2, (x, y)) + 
\left(\nabla_\alpha g(\theta, a, b, \alpha_2, (x, y)), \alpha_1 - \alpha_2\right) - \frac{\mu}{2}||\alpha_1 - \alpha_2||^2.
\]

And we can also verified that \( g \) is locally strongly convex in \( \Omega_{\alpha,b} \) w.r.t. \( a \) and \( b \).

Hence, if we in turn construct an AT problem based on Prop.1, we can obtain the following optimization problem with ease:

\[
\min_{\theta} \max_{\delta_1, \delta_2, \ldots, \delta_n} \sum_{i=1}^{n^n} \sum_{j=1}^{n^n} \ell(h_\theta(x_i^+ + \delta_i) - h_\theta(x_j^- + \delta_j))
\]

Proposition 1. The empirical risk of AUC in (OP0) is equivalent to

\[
\mathcal{L}(\theta, x, y) = \min_{a, b} \max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} g(\theta, a, b, \alpha, (x_i, y_i)),
\]

where

\[
g(\theta, a, b, \alpha, (x_i, y_i)) = (1 - p)(h_\theta(x_i) - a)^2\mathbb{I}_{[y_i=1]} + p(h_\theta(x_i) - b)^2\mathbb{I}_{[y_i=0]} + 2(1 + \alpha)\left(p h_\theta(x_i)\mathbb{I}_{[y_i=0]} - (1 - p) h_\theta(x_i)\mathbb{I}_{[y_i=1]}\right) - p(1 - p)\alpha^2.
\]

where \( a, b, \alpha \in \mathbb{R} \) are learnable parameters, and \( p = \Pr(y = 1) \).

Remark 1. According to (Ying et al., 2016), \( a, b, \alpha \) has the following closed-form solution: \( a = \mathbb{E}[h_\theta(x)|y = 1] \), \( b = \mathbb{E}[h_\theta(x)|y = 0] \) and \( \alpha = \mathbb{E}[h_\theta(x)|y = 0] - \mathbb{E}[h_\theta(x)|y = 1] \), where \( \mathbb{E} \) is a shorthand for sample mean. If the score \( h_\theta \) is normalized to the set \([0, 1]\), we can restrict \( a, b \) and \( \alpha \) to the following bounded domains:

\[
\begin{align*}
\Omega_{a,b} &= \{a, b \in \mathbb{R}|0 \leq a, b \leq 1\}, \\
\Omega_\alpha &= \{\alpha \in \mathbb{R}|\alpha \leq 1\}.
\end{align*}
\]

The good news here is that in the new loss function \( g \), positive samples and negative samples are independent of each other. However, the bad news is that the min-max-min-max problem is still hardly tractable. Through a careful investigation, if we can swap the order of \( \min_{\alpha} \) and \( \max_{a, b} \), we can then obtain a min-max problem which can be solved in an end-to-end fashion. To realize the idea, we could resort to the von Neumann’s Minimax theorem (Neumann, 1928; Sion, 1958):

**Theorem 1.** Let \( X \subset \mathbb{R}^n \) and \( Y \subset \mathbb{R}^m \) be compact convex sets. If \( f : X \times Y \to \mathbb{R} \) is a continuous function that is concave-convex, i.e.

\[
\begin{align*}
f(\cdot, y) : X &\to \mathbb{R} \text{ is concave for fixed } y, \\
f(x, \cdot) : Y &\to \mathbb{R} \text{ is convex for fixed } x.
\end{align*}
\]

Then we have that \( \max_{y \in Y} \min_{x \in X} f(x, y) = \min_{x \in X} \max_{y \in Y} f(x, y) \).

Moreover, we resort to the definition of weak-concavity (Böhm & Wright, 2021; Liu et al., 2021):
Definition 1. $f(x): \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be a $\gamma$-weakly concave ($\gamma > 0$) function w.r.t. $x$, if

$$f(x) - \frac{\gamma}{2}||x||_2^2$$

is a concave function w.r.t. $x$.

In the following proposition, we find a surrogate objective function $f(w, \alpha, x_i + \delta_i)$ such that the resulting optimization problem could be reformulated as a min-max problem:

Proposition 2. Define:

$$r(a,b, x) = \max_{\alpha} \frac{1}{n} \sum_{i=1}^n g(\theta, a, b, \alpha, (x_i + \delta_i, y_i)),$$

$$f(w, \alpha, x_i + \delta_i) = g(\theta, a, b, \alpha, (x_i + \delta_i, y_i)) - \gamma ||x_i + \delta_i||_2$$

If $r(a,b, x)$ is $\gamma_s$-weakly concave w.r.t. $\delta_1, \delta_2, \cdots, \delta_n$, then for all $\gamma > \gamma_s$, we have the following problem:

$$\min_{\theta} \max_{\alpha, a, b, \delta} \max_{\alpha, \delta} \frac{1}{n} \sum_{i=1}^n f(w, \alpha, (x_i + \delta_i))$$

is equivalent to:

$$(\text{OP}) \min_{w} \max_{\alpha, \delta} \max_{\alpha, \delta} \frac{1}{n} \sum_{i=1}^n [f(w, \alpha, x_i + \delta_i)]$$

$$= \min_w \max_{\alpha} \max_{\delta} \frac{1}{n} \sum_{i=1}^n [f(w, \alpha, x_i + \delta_i)],$$

where $w = (\theta, a, b)$. Moreover, $f(w, \alpha, x_i + \delta_i)$ is strongly concave w.r.t. $\delta_i$.

Remark 2. According to (Böhm & Wright, 2021; Liu et al., 2021), if $\max_{\alpha} \frac{1}{2} \sum_{i=1}^n [f(w, \alpha, x_i + \delta_i)]$ is smooth and $L$ gradient Lipschitz, then it is also $L$-weakly concave. Hence, the weakly concavity assumption is much weaker than the strongly-concave assumption (Wang et al., 2019a).

In this sense, we could turn to optimize (OP) in the next subsection.

4.2. Training Strategy

In this subsection, we continue to propose an adversarial AUC optimization framework to solve (OP). Specifically, we design solutions for the inner maximization problem and the outer min-max problem, respectively.

Inner Maximization Problem: Adversarial Attack. In this paper, we choose K-PGD (Madry et al., 2018) to generate adversarial examples. To better control the quality of adversarial examples, we introduce the First-Order Stationary Condition (FOSC) (Wang et al., 2019a) about the inner maximization problem, which is as follows:

$$c(x^k) = \max_{x \in \mathcal{X}} (x - x^k, \nabla_x f(w, \alpha, x^k))$$

When $c(x^k) = 0$, the optimization problem reaches the convergence state. Specifically, such a condition can be achieved when a) $\nabla_x f(w, \alpha, x^k) = 0$, or b) $x^k - x^0 = \epsilon \cdot \text{sign}(\nabla_x f(w, \alpha, x^k))$. Here a) implies $x^k$ is a stationary point in the inner maximization problem, b) shows that local maximum point of $f(w, \alpha, x^k)$ reaches the boundary of $\mathcal{X}$. The proof process is shown in Lem.1.

Outer min-max Problem. For the outer min-max problem, we apply Stochastic Gradient Descent Ascent (SGDA) to solve the problem. At each iteration, SGDA performs stochastic gradient descent over the parameter $w$ with the stepsize $\eta_w$, and stochastic gradient ascent over the parameter $\alpha$ with the stepsize $\eta_\alpha$.

The total training strategy is presented in Alg.1. This is an extension of the algorithm proposed in (Wang et al., 2019a), where the outer level minimization problem now becomes a min-max problem. The value of FOSC can imply the adversarial strength of adversarial examples, whereas a small FOSC value implies a high adversarial strength. Due to this fact, through the FOSC value of current epoch $c_t$, we can dynamically control the strength of adversarial examples. Specifically, in the initial stage of training, the value of $c_t$ is large, which means the generated adversarial examples are not so hard. In the later stage of training, the value of $c_t$ is 0, which allows model to be trained on much stronger adversarial examples. Consequently, such an algorithm will allow the model to learn from in a progressive manner. Here, we use $M_c$ to mask the examples that satisfy the condition in Line 9 in Alg.1. By doing so, we can ensure that the FOSC values of the adversarial examples generated by
Algorithm 1 are made in (Sinha et al., 2018), and the last three w.r.t. \(\alpha\) Alg. 1 are all less than \(\hat{L}\) least Lipschitz conditions as follows: assumptions needed to the analysis. The same definition as \(L\) stochastic gradient of \(\nabla f\). The adversarial AUC optimization framework. Next, we provide a convergence analysis of our proposed descent-ascent on \(w\) and \(\alpha\) respectively.

4.3. Convergence Analysis

Next, we provide a convergence analysis of our proposed adversarial AUC optimization framework. We first give the definition and description of some notations. In detail, let \(x^*_i(w, \alpha) = \arg\max_{x_i \in \mathcal{X}_i} f(w, \alpha, x_i)\) where \(\mathcal{X}_i = \{x \mid \|x - x^0_i\|_\infty \leq \epsilon\}\). And

\[
L(w, \alpha) = \frac{1}{n} \sum_{i=1}^{n} \max_{x_i \in \mathcal{X}_i} f(w, \alpha, x_i) = \frac{1}{n} \sum_{i=1}^{n} f(w, \alpha, x_i^*).
\]

Then \(\hat{x}_i(w, \alpha)\) is a \(\delta\)-approximate solution to \(x^*_i(w, \alpha)\), if it satisfies that

\[
\max_{x \in \mathcal{X}_i} \langle x - \hat{x}_i(w, \alpha), \nabla_x f(w, \alpha, \hat{x}_i(w, \alpha)) \rangle \leq \delta. \tag{6}
\]

Furthermore, let \(\nabla L(\alpha)\) denote the gradient of \(L(w, \alpha)\) w.r.t. \(\alpha\). And let \(g(\alpha) = \frac{1}{M} \sum_{i \in B} \nabla_x f(w, \alpha, x_i)\) be the stochastic gradient of \(L(w, \alpha)\) w.r.t. \(\alpha\), where \(B\) is mini-batch and \(M = |B|\). Meanwhile, let \(\nabla_x f(w, \alpha, \hat{x}(w, \alpha))\) be the gradient of \(f(w, \alpha, \hat{x}(w, \alpha))\) w.r.t. \(\alpha\), and let \(\hat{g}(\alpha) = \frac{1}{M} \sum_{i \in B} \nabla f(w, \alpha, x_i)\) be the approximate stochastic gradient of \(L(w, \alpha)\) w.r.t. \(\alpha\). And for \(w\), we have the same definition as \(\alpha\). In addition, let

\[
\Phi(w) = \max_{\alpha} L(w, \alpha), \quad \nabla \Phi(w) = \nabla_w L(w, \alpha^*(w)).
\]

Then, before giving the convergence analysis, we list some assumptions needed to the analysis.

**Assumption 1.** The function \(f(w, \alpha, x)\) satisfies the gradient Lipschitz conditions as follows:

\[
\begin{align*}
\sup_{\alpha, w} \|\nabla_w f(w, \alpha, x) - \nabla_w f(w', \alpha, x)\|_2 &\leq L_{ww} \|w - w'\|_2, \\
\sup_{\alpha, w} \|\nabla_x f(w, \alpha, x) - \nabla_x f(w', \alpha, x)\|_2 &\leq L_{wx} \|x - x'\|_2, \\
\sup_{\alpha, w} \|\nabla_x f(w, \alpha, x) - \nabla_x f(w, \alpha', x)\|_2 &\leq L_{xx} \|x - x'\|_2, \\
\sup_{\alpha, w} \|\nabla_x f(w, \alpha, x) - \nabla_x f(w, \alpha', x')\|_2 &\leq L_{w} \|\alpha - \alpha'\|_2, \\
\sup_{w, \alpha} \|\nabla_x f(w, \alpha, x) - \nabla_x f(w', \alpha, x')\|_2 &\leq L_{wa} \|x - x'\|_2, \\
\sup_{w, \alpha} \|\nabla_x f(w, \alpha, x) - \nabla_x f(w', \alpha', x)\|_2 &\leq L_{wa} \|\alpha - \alpha'\|_2.
\end{align*}
\]

where \(L_{w}, L_{ww}, L_{wx}, L_{xx}, L_{w} \) and \(L_{wa}\) are positive constants.

**Remark 3.** The first three gradient Lipschitz conditions in Asm.1 are made in (Sinha et al., 2018), and the last three gradient Lipschitz conditions are made in (Liu et al., 2019). Meanwhile, for the overparameterized deep neural network, the loss function is semi-smooth (Allen-Zhu et al., 2019; Du et al., 2019), which helps to justify Asm.1.

**Assumption 2.** \(\|\nabla_w f(w, \alpha, x)\|_2\) is upper bounded by \(l_w\).

**Remark 4.** Asm.2 is widely used in minimax optimization problems (Sinha et al., 2018).

**Assumption 3.** \(f(w, \alpha, x)\) is locally \(\mu\)-strongly concave in \(X_i\) for all \(i \in [n]\), i.e. for any \(x_1, x_2 \in X_i\), it holds that

\[
f(w, \alpha, x_1) \leq f(w, \alpha, x_2) + \langle \nabla_x f(w, \alpha, x_2), x_1 - x_2 \rangle - \frac{\mu}{2} \|x_1 - x_2\|^2.
\]

**Remark 5.** The strongly concave assumption is equivalent to weakly concave assumption of \(r\), which is much easier to be achieved.

**Assumption 4.** The stochastic gradient \(g(\alpha)\) satisfies

\[
\begin{align*}
E[g(\alpha) - \nabla L(\alpha)] = 0, \quad &E\left[\|g(\alpha) - \nabla L(\alpha)\|_2^2\right] \leq \sigma^2, \\
E[g(w) - \nabla L(w)] = 0, \quad &E\left[\|g(w) - \nabla L(w)\|_2^2\right] \leq \sigma^2.
\end{align*}
\]

**Remark 6.** Asm.4 is a common assumption used to analyze the optimization algorithm based on stochastic gradient (Lin et al., 2020; Liu et al., 2019).

**Theorem 2.** Under the above assumptions and let the stepsizes be chosen as \(\eta_w = \Theta(1/\kappa L)\), \(\eta_\alpha = \Theta(1/L)\) and \(\kappa \leq \frac{1}{T}\), then we have the following inequality:

\[
\frac{1}{T + 1} \sum_{t=0}^{T} E\left[\|\nabla \Phi(w_t)\|_2^2\right] \leq 360\kappa^2 L^3 \Delta_\Phi + 13\kappa L^2 D^2 + \frac{26\kappa\sigma^2}{M} + h_\delta + h_\Delta
\]

where \(h_\delta = \frac{1024}{253} (3\kappa L \delta + 6\kappa^4 L \delta + \frac{L \delta}{\kappa} + L^2 \sqrt{\frac{3}{\mu}})\) and \(h_\Delta = \frac{384\kappa^4 L^2 \Delta}{253}\). The condition number \(\kappa = L/\mu\). \(\Delta_\Phi\) is a common approximation of the inner maximization problem instead of the optimal solution \(x^*\), and the fourth term \(h_\Delta\) is due to \(X\) being a bounded set. Since all of the residuals are small, we can find an \(e\)-stationary point within a finite number of epochs.

5. Experiments

In this section, we evaluate the performance of our AdAUC algorithms in three long-tail datasets.
We compare the performance of our proposed algorithm and the AT methods with the classical CE classification loss function when the datasets have the long tail distribution.

We adopt the WideResNet-28 (Zagoruyko & Komodakis, 2016) as the model architecture. The other detailed settings are shown in App.C.1. In Tab.2, NT means Natural Training without adversarial operations, AT$_1$ means adversarial training without FOSC, and AT$_2$ means algorithm in Alg.1.

To validate the robustness of our algorithm, we adopt FSGM (Goodfellow et al., 2014), iterative attack PGD (Madry et al., 2018), C&W (Carlini & Wagner, 2017) and ensemble attack AA (Croce & Hein, 2020) as attack methods.

### 5.2. Dataset Description

#### Binary CIFAR-10-LT Dataset.**

We construct a long-tail CIFAR-10 dataset, where the sample size across different classes decays exponentially and ensure the ratio of sample sizes of the least frequent to the most frequent class is set to 0.01. Then, we label the first 5 classes as the negative class and the last 5 classes as positive, which leads that the ratio of positive class size to negative class size $\rho \approx 1 : 9$.

#### Binary CIFAR-100-LT Dataset.**

We construct a long-tail CIFAR-100 dataset in the same way as CIFAR-10-LT, where we label the first 50 classes as the negative class and the last 50 classes as positive, which leads that the ratio of positive class size to negative class size $\rho \approx 1 : 9$.

![Figure 2. The t-SNE projection of our methods on MNIST-LT dataset. Red points represent the positive examples, and blue points represent the negative examples.](image-url)
class size to negative class size $\rho \approx 1 : 9$.

**Binary MNIST-LT Dataset.** We construct a long-tail MNIST dataset from original MNIST dataset (LeCun et al., 1998) in the same way as CIFAR-10-LT, where the ratio of positive class size to negative class size $\rho \approx 1 : 9$.

### 5.3. Overall Performance

The performance and robustness of all the involved methods on three datasets are shown in Tab.2. Consequently, we have the following observations: 1) On all the datasets, our methods achieve the best or competitive performance evaluated against all adversarial attack methods as shown in Tab.2. 2) Even the AUC optimization with NT has certain robustness (the AUC will not drop to 0 when evaluated against adversarial examples). This is because the decision surface obtained by AUC optimization has a greater tolerance for minority classes than CE. Specifically, the decision surface is far away from the positive examples. To validate this argument, we show the score distribution in Fig.4 w.r.t MNIST-LT. For AdAUC NT, adversarial examples increase the score of the negative examples and decreases the score of the positive examples. The more results of other datasets are shown in the App.C.2, which show a similar trend. Moreover, we show the feature visualization of NT and AT$_2$ of our methods for clean data and adversarial examples. It implies that our AdAUC algorithm could separate the positive and negative instances well in the embedding space in Fig.2.

### 5.4. Convergence Analysis

We report the convergence of test AUC of CE-based methods and our proposed methods in Fig. 3. We can observe that our proposed method performs better both on clean data and adversarial examples. However, due to the high complexity of the outer min-max problem, it could be seen that our method converges slightly slower than CE methods, which is consistent with the analysis of Thm.2.

### 6. Conclusion

In this paper, we initiate the study on adversarial AUC optimization against long-tail problem. The complexity of AUC loss function makes the corresponding adversarial training hardly scalable. To address this issue, we first construct a reformulation of the AT problem of AUC optimization.
By further applying a concavity promoting regularizer, we can reformulate the original problem as a min-max problem where the objective function can be expressed instance-wisely. On top of the reformulation, we construct an end-to-end training algorithm with provable guarantee. Finally, we conduct a series of empirical studies on three long-tail benchmark datasets, the results of which demonstrate the effectiveness of our proposed method.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: U21B2038, 61931008, 6212200758 and 61976202, in part by the Fundamental Research Funds for the Central Universities, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX2021298.

References


AdAUC: End-to-end Adversarial AUC Optimization Against Long-tail Problems


A. Proofs of Main Results

In this section, we provide the proofs of the main results.

A.1. Proof of Proposition 2

Proof. According to (Ying et al., 2016), $a, b, \alpha$ have the following closed-form solution:

$$a = \mathbb{E}[h_\theta(x)|y = 1], \quad b = \mathbb{E}[h_\theta(x)|y = 0], \quad \alpha = \mathbb{E}[h_\theta(x)|y = 0] - \mathbb{E}[h_\theta(x)|y = 1].$$

Then it is easy to check that

$$r(a, b) = \max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} g(\theta, a, b, \alpha, (x_i + \delta_i, y_i))$$

is a strongly-convex problem w.r.t. $(a, b)$. However, $r$ is in general not concave w.r.t. $\delta_i$. In this sense, we then try to find a surrogate objective to induce the concavity w.r.t $\delta_i$. Specifically, we adopt a concavity regularization term $-\gamma ||x_i + \delta_i||^2$, and define a surrogate objective:

$$f(w, \alpha, x_i + \delta_i) = g(\theta, a, b, \alpha, (x_i + \delta_i, y_i)) - \gamma ||x_i + \delta_i||^2.$$

Therefore, if $r$ is $\gamma_a$-weakly concave w.r.t. $\delta_i, i = 1, 2, \ldots, n$ (Liu et al., 2021; Böhм & Wright, 2021), then we can define $\gamma > \gamma_a$ to obtain an objective $f(w, \alpha, x + \delta)$ such that $\max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} f(w, \alpha, x_i + \delta_i)$ is strongly concave w.r.t. $\delta_i, i = 1, 2, \ldots, n$. Above all, with the weakly concavity assumption, we can instead solve the following surrogate problem by the von Neumann’s minimax theorem:

$$(OP) \quad \min_{w} \max_{\alpha} \max_{\delta} \frac{1}{n} \sum_{i=1}^{n} f(w, \alpha, x_i + \delta_i) \quad (7)$$

$$= \min_{w} \max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta} f(w, \alpha, x_i + \delta_i),$$

where $w = (\theta, a, b)$.

A.2. Proof of Lemma 1

Lemma 1. For all $x \in \mathcal{X}$, $c(x^k) = 0$ when 1) $\nabla_x f(w, \alpha, x^k) = 0$, or 2) $x^k - x^0 = \epsilon \cdot \text{sign}(\nabla_x f(w, \alpha, x^k))$.

Proof.

$$c(x^k) = \max_{x \in \mathcal{X}} (x - x^k, \nabla_x f(w, \alpha, x^k)) = \max_{x \in \mathcal{X}} (x - x^0 + x^0 - x^k, \nabla_x f(w, \alpha, x^k)) = \max_{x \in \mathcal{X}} (x - x^0, \nabla_x f(w, \alpha, x^k)) - (x^k - x^0, \nabla_x f(w, \alpha, x^k)) = \epsilon \cdot \left(\|\nabla_x f(w, \alpha, x^k)\|_1 - (x^k - x^0, \nabla_x f(w, \alpha, x^k))\right)$$

This completes the proof.

A.3. Proof of Technical Lemmas

In this subsections, we present seven key lemmas which are important for the proof of Theorem 2.

Lemma 2. Under the Asm.1 and 3, we have $L(\alpha)$ is $L_\alpha$-smooth where $L_\alpha = \frac{L_{\alpha x} L_{\alpha w}}{\mu} + L_{\alpha x x}$. For any $\alpha_1, \alpha_2$, it holds

$$L(\alpha_1) \leq L(\alpha_2) + \langle \nabla L(\alpha_2), \alpha_1 - \alpha_2 \rangle + \frac{L_{\alpha} \|\alpha_1 - \alpha_2\|^2}{2}$$

$$\|\nabla L(\alpha_1) - \nabla L(\alpha_2)\|_2 \leq L_{\alpha} \|\alpha_1 - \alpha_2\|_2. \quad (9)$$

We also have that $L(\alpha)$ is $L_\alpha$-smooth where $L_\alpha = \frac{L_{\alpha x} L_{\alpha w}}{\mu} + L_{\alpha x w}$.
Proof. Here, since we only focus on $\alpha$ and $x$ when the $w$ is fixed, we abbreviate $x^*(w, \alpha)$ as $x^*(\alpha)$ for convenience.

By the Asm.3, we have
\[
f(w, \alpha_2, x^*_i(\alpha_1)) \leq f(w, \alpha_2, x^*_i(\alpha_2)) + \langle \nabla_x f(w, \alpha_2, x^*_i(\alpha_2)), x^*_i(\alpha_1) - x^*_i(\alpha_2) \rangle - \frac{\mu}{2} \|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|^2, \quad (10)
\]
\[
f(w, \alpha_2, x^*_i(\alpha_2)) \leq f(w, \alpha_2, x^*_i(\alpha_1)) + \langle \nabla_x f(w, \alpha_2, x^*_i(\alpha_1)), x^*_i(\alpha_2) - x^*_i(\alpha_1) \rangle - \frac{\mu}{2} \|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|^2. \quad (11)
\]

Since $\langle \nabla_x f(w, \alpha_2, x^*_i(\alpha_2)), x^*_i(\alpha_1) - x^*_i(\alpha_2) \rangle \leq 0$, combining (10) and (11), we obtain
\[
\mu \|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|^2 \leq \langle \nabla_x f(w, \alpha_2, x^*_i(\alpha_1)), x^*_i(\alpha_2) - x^*_i(\alpha_1) \rangle \\
\leq \langle \nabla_x f(w, \alpha_2, x^*_i(\alpha_1)) - \nabla_x f(w, \alpha_1, x^*_i(\alpha_1)), x^*_i(\alpha_2) - x^*_i(\alpha_1) \rangle \\
\leq \|\nabla_x f(w, \alpha_2, x^*_i(\alpha_1)) - \nabla_x f(w, \alpha_1, x^*_i(\alpha_1))\|_2 \|x^*_i(\alpha_2) - x^*_i(\alpha_1)\|_2 \\
\leq L_{\text{xa}} \|\alpha_1 - \alpha_2\|_2 \|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|_2, \quad (12)
\]
where the second inequality holds because $\langle \nabla_x f(w, \alpha_1, x^*_i(\alpha_1)), x^*_i(\alpha_2) - x^*_i(\alpha_1) \rangle \leq 0$.

Then (12) immediately yields
\[
\|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|_2 \leq \frac{L_{\text{xa}}}{\mu} \|\alpha_1 - \alpha_2\|_2. \quad (13)
\]

Then for $i \in [n]$, we have
\[
\|\nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_1)) - \nabla_\alpha f(w, \alpha_2, x^*_i(\alpha_2))\|_2 \leq \|\nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_1)) - \nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_2))\|_2 \\
+ \|\nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_2)) - \nabla_\alpha f(w, \alpha_2, x^*_i(\alpha_2))\|_2 \\
\leq L_{\alpha x} \|x^*_i(\alpha_1) - x^*_i(\alpha_2)\|_2 + L_{\alpha a} \|\alpha_1 - \alpha_2\|_2 \\
\leq \left(\frac{L_{\alpha x} L_{\text{xa}}}{\mu} + L_{\alpha a}\right) \|\alpha_1 - \alpha_2\|_2. \quad (14)
\]

Finally, by the definition of $L(\alpha)$, we have
\[
\|\nabla L(\alpha_1) - \nabla L(\alpha_2)\|_2 \leq \left|\frac{1}{n} \sum_{i=1}^{n} \nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_1)) - \frac{1}{n} \sum_{i=1}^{n} \nabla_\alpha f(w, \alpha_2, x^*_i(\alpha_2))\right|_2 \\
\leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla_\alpha f(w, \alpha_1, x^*_i(\alpha_1)) - \nabla_\alpha f(w, \alpha_2, x^*_i(\alpha_2))\|_2 \\
\leq \left(\frac{L_{\alpha x} L_{\text{xa}}}{\mu} + L_{\alpha a}\right) \|\alpha_1 - \alpha_2\|_2. \quad (15)
\]

This completes the proof.

\[\square\]

Lemma 3. Under the Asm.1, we have $\Phi(w)$ is $L_w$-smooth where $L_w = L_{\text{xa}} + L_{\text{wx}}$. For any $w_1, w_2$, it holds
\[
\Phi(w_1) \leq \Phi(w_2) + \langle \nabla \Phi(w_2), w_1 - w_2 \rangle + \frac{L_w}{2} \|w_1 - w_2\|_2^2 \\
\|\nabla \Phi(w_1) - \nabla \Phi(w_2)\|_2 \leq L_w \|w_1 - w_2\|_2. \quad (16)
\]

Proof. By the Rem.1, we have
\[
L(w_2, \alpha^*(w_1)) \leq L(w_2, \alpha^*(w_2)) + \langle \nabla_{\alpha} L(w_2, \alpha^*(w_2)), \alpha^*(w_1) - \alpha^*(w_2) \rangle - \frac{\mu}{2} \|\alpha^*(w_1) - \alpha^*(w_2)\|_2^2, \quad (17)
\]
\[
L(w_2, \alpha^*(w_2)) \leq L(w_2, \alpha^*(w_1)) + \langle \nabla_{\alpha} L(w_2, \alpha^*(w_1)), \alpha^*(w_2) - \alpha^*(w_1) \rangle - \frac{\mu}{2} \|\alpha^*(w_1) - \alpha^*(w_2)\|_2^2. \quad (18)
\]
We also have
\[ \| \alpha^*(w_1) - \alpha^*(w_2) \|^2 \leq (\nabla_{\alpha} \langle L(w_1, \alpha^*(w_2)) - \alpha^*(w_1) \rangle - \nabla_{\alpha} \langle L(w_2, \alpha^*(w_1)) - \alpha^*(w_2) \rangle) \]
\[ \leq \| \nabla_{\alpha} \langle L(w_1, \alpha^*(w_1)) - \nabla_{\alpha} \langle L(w_1, \alpha^*(w_1)) \rangle \|_2 \| \alpha^*(w_2) - \alpha^*(w_1) \|_2 \]
\[ \leq L_{\alpha w} \| w_2 - w_1 \|_2 \| \alpha^*(w_2) - \alpha^*(w_1) \|_2 \cdot \]  
(19)

Then (19) yields
\[ \| \alpha^*(w_1) - \alpha^*(w_2) \|_2 \leq \frac{L_{\alpha w}}{\mu} \| w_1 - w_2 \|_2 \cdot \]
(20)

Then we have for \( i \in [n] \),
\[ \| \nabla_w L(w_1, \alpha^*(w_1)) - \nabla_w L(w_2, \alpha^*(w_2)) \|_2 \leq \| \nabla_w L(w_1, \alpha^*(w_1)) - \nabla_w L(w_1, \alpha^*(w_2)) \|_2 \]
\[ + \| \nabla_w L(w_1, \alpha^*(w_2)) - \nabla_w L(w_2, \alpha^*(w_1)) \|_2 \]
\[ \leq L_{\alpha w} \| \alpha^*(w_1) - \alpha^*(w_2) \|_2 + L_{\alpha w} \| w_1 - w_2 \|_2 \]
\[ \leq \left( \frac{L_{\alpha w} L_{\alpha w}}{\mu} + L_{\alpha w} \right) \| w_1 - w_2 \|_2 \cdot \]
(21)

Finally, by the definition of \( \Phi(w) \), we have
\[ \| \nabla \Phi(w_1) - \nabla \Phi(w_2) \|_2 \leq \left( \frac{1}{n} \sum_{i=1}^n \nabla \Phi(w_1, \alpha^*(w_1)) - \nabla \Phi(w_2, \alpha^*(w_2)) \right) \]
\[ \leq \frac{1}{n} \sum_{i=1}^n \| \nabla_{\alpha} L(w_1, \alpha^*(w_1)) - \nabla_{\alpha} L(w_2, \alpha^*(w_2)) \|_2 \]
\[ \leq \left( \frac{L_{\alpha w} L_{\alpha w}}{\mu} + L_{\alpha w} \right) \| w_1 - w_2 \|_2 \cdot \]
(22)

This completes the proof.

\[ \square \]

**Lemma 4.** Under Asm.1 and 3, the approximate stochastic gradient \( \hat{g}(\alpha) \) satisfies
\[ \| \hat{g}(\alpha) - g(\alpha) \|_2 \leq L_{\alpha x} \sqrt{\frac{\delta}{\mu}} \]
(23)

We also have
\[ \| \hat{g}(w) - g(w) \|_2 \leq L_{\alpha x} \sqrt{\frac{\delta}{\mu}} \]
(24)

**Proof.** We have
\[ \| \hat{g}(\alpha) - g(\alpha) \|_2 \leq \left( \frac{1}{|B|} \sum_{i \in B} \left( \nabla_{\alpha} f(w, \alpha, \hat{x}_i(\alpha)) - \nabla_{\alpha} f(w, \alpha, x_i^*(\alpha)) \right) \right) \|_2 \]
\[ \leq \frac{1}{|B|} \sum_{i \in B} \| \nabla_{\alpha} f(w, \alpha, \hat{x}_i(\alpha)) - \nabla_{\alpha} f(w, \alpha, x_i^*(\alpha)) \|_2 \]
\[ \leq \frac{1}{|B|} \sum_{i \in B} L_{\alpha x} \| \hat{x}_i(\alpha) - x_i^*(\alpha) \|_2 \cdot \]
(25)

By the Asm.3, we have
\[ \mu \| \hat{x}_i(\alpha) - x_i^*(\alpha) \|_2 \leq (\nabla_{\alpha} f(w, \alpha, \hat{x}_i(\alpha)) - \nabla_{\alpha} f(w, \alpha, x_i^*(\alpha)), \hat{x}_i(\alpha) - x_i^*(\alpha)). \]
(26)
Since the $\hat{x}_i(\alpha)$ is a $\delta$-approximate solution to $x_i^*(\alpha)$, if it satisfies that
\[
\langle x_i^*(\alpha) - \hat{x}_i(\alpha), \nabla_x f(w, \alpha, \hat{x}_i(\alpha)) \rangle \leq \delta. \tag{27}
\]
Furthermore, we have
\[
\langle \hat{x}_i(\alpha) - x_i^*(\alpha), \nabla_x f(w, \alpha, x_i^*(\alpha)) \rangle \leq 0. \tag{28}
\]
Combining (27) and (28), we obtain
\[
\langle \hat{x}_i(\alpha) - x_i^*(\alpha), \nabla_x f(w, \alpha, x_i^*(\alpha)) - \nabla_x f(\alpha, \hat{x}_i)(\alpha) \rangle \leq \delta. \tag{29}
\]
Combining (26) and (29), we obtain
\[
\mu \|\hat{x}_i(\alpha) - x_i^*(\alpha)\|_2^2 \leq \delta. \tag{30}
\]
Combining (25) and (30), we obtain
\[
\|\hat{g}(\alpha) - g(\alpha)\|_2 \leq \omega \sqrt{\frac{\delta}{\mu}}. \tag{31}
\]

**Lemma 5.** $g(w) = \frac{1}{M} \sum_{i=1}^{M} G_w(w_i, \alpha_t, \xi_i)$ and $g(\alpha) = \frac{1}{M} \sum_{i=1}^{M} G(\alpha(w_i, \alpha_t, \xi_i))$ are unbiased and have bounded variance:
\[
E[g(w)] = \nabla_w L(w_t, \alpha_t), \quad E[\|g(w)\|_2^2] = \|\nabla_w L(w_t, \alpha_t)\|_2^2 + \frac{\sigma^2}{M}, \tag{32}
\]
\[
E[g(\alpha)] = \nabla_\alpha L(w_t, \alpha_t), \quad E[\|g(\alpha)\|_2^2] = \|\nabla_\alpha L(w_t, \alpha_t)\|_2^2 + \frac{\sigma^2}{M}.
\]

**Proof.** Since $\hat{g}$ is unbiased, we have
\[
E[g(w)] = \nabla_w L(w_t, \alpha_t), \quad E[g(\alpha)] = \nabla_\alpha L(w_t, \alpha_t).
\]
Furthermore, we have
\[
E[\|g(w) - \nabla_w L(w_t, \alpha_t)\|_2^2] = E\left[\left\| \frac{1}{M} \sum_{i=1}^{M} G_w(w_i, \alpha_t, \xi_i) - \nabla_w L(w_t, \alpha_t) \right\|_2^2 \right]
\leq \frac{\sigma^2}{M}.
\]
\[
E[\|g(\alpha) - \nabla_\alpha L(w_t, \alpha_t)\|_2^2] = E\left[\left\| \frac{1}{M} \sum_{i=1}^{M} G(\alpha(w_i, \alpha_t, \xi_i)) - \nabla_\alpha L(w_t, \alpha_t) \right\|_2^2 \right]
\leq \frac{\sigma^2}{M}.
\]

**Lemma 6.** The iterates $\{w_t\}_{t \geq 1}$ satisfy the following inequality:
\[
E[\Phi(w_t)] \leq E[\Phi(w_{t-1})] - \left(\frac{\eta_w}{2} - 2L_w \eta_w^2\right) E\left[\|\nabla \Phi(w_{t-1})\|_2^2\right]
+ \left(\frac{\eta_w}{2} + 2L_w \eta_w^2\right) E\left[\|\nabla \Phi(w_{t-1}) - \nabla_w L(w_{t-1}, \alpha_{t-1})\|_2^2\right]
+ \frac{L_w \eta_w^2}{\mu} \sigma^2 + \frac{\delta L_w L_w \eta_w^2}{\mu} + L_w \ell \sqrt{\frac{\delta}{\mu}}. \tag{33}
\]
Proof. By Lem.2, we have
\[
\Phi(w_t) \leq \Phi(w_{t-1}) + \langle \nabla \Phi(w_{t-1}), w_t - w_{t-1} \rangle + \frac{L_w}{2} \| w_t - w_{t-1} \|^2.
\]
(34)
Plugging \( w_t - w_{t-1} = -\eta_w \hat{g}(w) \) into (34) yields that
\[
\Phi(w_t) \leq \Phi(w_{t-1}) - \eta_w \| \nabla \Phi(w_{t-1}), \hat{g}(w) \| + \frac{L_w \eta_w^2}{2} \| \hat{g}(w) \|^2
\]
\[
= \Phi(w_{t-1}) - \eta_w \| \nabla \Phi(w_{t-1}) \|^2 - \eta_w \langle \nabla \Phi(w_{t-1}), \hat{g}(w) \rangle + \frac{L_w \eta_w^2}{2} \| \hat{g}(w) \|^2
\]
\[
\leq \Phi(w_{t-1}) - \eta_w \| \nabla \Phi(w_{t-1}) \|^2 - \eta_w \langle \nabla \Phi(w_{t-1}), g(w) \rangle + \frac{L_w \eta_w^2}{2} \| g(w) \|^2 + L_w \eta_w^2 \| \hat{g}(w) - g(w) \|^2
\]
\[
= \Phi(w_{t-1}) - \eta_w \| \nabla \Phi(w_{t-1}) \|^2 - \eta_w \langle \nabla \Phi(w_{t-1}), g(w) - \nabla_w L(w_{t-1}, \alpha_{t-1}) \rangle
\]
\[
+ L_w \eta_w^2 \| \hat{g}(w) - g(w) \|^2 - \eta_w \langle \nabla \Phi(w_{t-1}), g(w) - \nabla_w L(w_{t-1}, \alpha_{t-1}) \rangle
\]
\[
+ \frac{\eta_w}{2} \| \nabla \Phi(w_{t-1}) - \nabla_w L(w_{t-1}, \alpha_{t-1}) \|^2 + L_w \eta_w^2 \| g(w) \|^2 + \frac{\delta L_w L^2}{\mu^2} \eta_w^2 + \eta_w L_w \| \hat{g}(w) - g(w) \| \sqrt{\frac{\delta}{\mu}}.
\]
(35)
By Lem.5, taking the expectation on the both sides yields that
\[
\mathbb{E}[\Phi(w_t)] \leq \mathbb{E}[\Phi(w_{t-1})] - \frac{\eta_w}{2} \mathbb{E}[\| \nabla \Phi(w_{t-1}) \|^2] + \frac{\eta_w^2}{2} \mathbb{E}[\| \nabla \Phi(w_{t-1}) - \nabla_w L(w_{t-1}, \alpha_{t-1}) \|^2] + L_w \eta_w^2 \mathbb{E}[\| \hat{g}(w) \|^2] + \frac{\delta L_w L^2}{\mu^2} \eta_w^2 + \eta_w L_w \mathbb{E}[\| \hat{g}(w) - g(w) \| \sqrt{\frac{\delta}{\mu}}].
\]
(36)
By the Cauchy-Schwarz inequality, we have
\[
\| \nabla_w L(w_{t-1}, \alpha_{t-1}) \|^2 \leq 2 \left( \| \nabla_w L(w_{t-1}, \alpha_{t-1}) - \nabla \Phi(w_{t-1}) \|^2 + \| \nabla \Phi(w_{t-1}) \|^2 \right).
\]
(37)
Plugging (37) into (36) and taking the expectation of both side, yields that
\[
\mathbb{E}[\Phi(w_t)] \leq \mathbb{E}[\Phi(w_{t-1})] - \left( \frac{\eta_w}{2} - 2L_w \eta_w^2 \right) \mathbb{E}[\| \nabla \Phi(w_{t-1}) \|^2] + \left( \frac{\eta_w}{2} + 2L_w \eta_w^2 \right) \mathbb{E}[\| \nabla \Phi(w_{t-1}) - \nabla_w L(w_{t-1}, \alpha_{t-1}) \|^2] + \frac{L_w \eta_w^2 \sigma^2}{M} + \frac{\delta L_w L^2}{\mu^2} \eta_w^2 + \eta_w L_w \mathbb{E}[\| \hat{g}(w) - g(w) \| \sqrt{\frac{\delta}{\mu}}].
\]
This completes the proof.

Lemma 7. Let \( \delta_t = \mathbb{E}[\| \alpha^*(w_t, x_t^t) - \alpha_t(w_t, \hat{x}_t) \|^2] \), the following statements holds that
\[
\delta_t \leq 1 - \frac{\mu}{2L_w \sigma^2} + \frac{L^2 w L^2 \sigma^2 \eta_w^2}{2M} \delta_{t-1} + \frac{L^2 w \eta_w^2}{2M^2} \mathbb{E}[\| \nabla \Phi(w_{t-1}) \|^2] + \frac{2 \sigma^2}{L_w M} + \frac{L^2 w \eta_w^2 \sigma^2}{4M^3} + \frac{L^2 w \eta_w^2 \sigma^2}{4M^4} + \frac{L^2 w \eta_w^2 \sigma^2}{8M^5} + \frac{2 \mu \delta}{\mu^3}.
\]
(38)
Proof.
\[
\delta_t = \mathbb{E}[\| \alpha^*(w_t, x_t^*) - \alpha^*(w_t, \hat{x}_t) + \alpha^*(w_t, \hat{x}_t) - \alpha_t(w_t, \hat{x}_t) \|^2] \leq 2 \mathbb{E}[\| \alpha^*(w_t, x_t^*) - \alpha^*(w_t, \hat{x}_t) \|^2] + 2 \mathbb{E}[\| \alpha^*(w_t, \hat{x}_t) - \alpha_t(w_t, \hat{x}_t) \|^2].
\]
By Young’s inequality, for any $\epsilon_0 \leq 0$, we have

$$
E \left[ \| \alpha^*(w_t, \hat{x}_t) - \alpha_t(w_t, \hat{x}_t) \|^2 \right] \leq \left(1 + \frac{1}{\epsilon_0} \right) E \left[ \| \alpha^*(w_{t-1}, \hat{x}_{t-1}) - \alpha_t(w_{t-1}, \hat{x}_{t-1}) \|^2 \right] + (1 + \epsilon_0) E \left[ \| \alpha^*(w_t, \hat{x}_t) - \alpha^*(w_{t-1}, \hat{x}_{t-1}) \|^2 \right].
$$

(39)

By the Rem.1 and $\eta_\alpha = 1/L_{aw}$ (refer to (Nesterov, 1998) Theorem 2.3.4), we have

$$
E \left[ \| \alpha^*(w_{t-1}, \hat{x}_{t-1}) - \alpha_t(w_{t-1}, \hat{x}_{t-1}) \|^2 \right] \leq \left(1 - \frac{\mu}{L_{aw}} \right) \delta_{t-1} + E \left[ \| \alpha^*(w_{t-1}, \hat{x}_{t-1}) - \alpha^*(w_{t-1}, x^*_{t-1}) \|^2 \right] + \frac{\sigma^2}{L_{aw}^2}.
$$

(40)

By triangle inequality, we have

$$
E \left[ \| \alpha^*(w_{t-1}, \hat{x}_{t-1}) - \alpha_t(w_{t-1}, \hat{x}_t) \|^2 \right] \leq 2E \left[ \| \alpha^*(w_{t-1}, \hat{x}_{t-1}) - \alpha^*(w_{t-1}, x^*_{t-1}) \|^2 \right] + \frac{\sigma^2}{L_{aw}^2}.
$$

(41)

Plugging (41) into (40), yields that

$$
E \left[ \| \alpha^*(w_{t-1}, \hat{x}_t) - \alpha_t(w_t, \hat{x}_t) \|^2 \right] \leq 2 \left(1 - \frac{\mu}{L_{aw}} \right) \delta_{t-1} + 2E \left[ \| \alpha^*(w_{t-1}, \hat{x}_t) - \alpha^*(w_{t-1}, x^*_{t-1}) \|^2 \right] + \frac{\sigma^2}{L_{aw}^2}.
$$

(42)

By triangle inequality, we have

$$
E \left[ \| \alpha^*(w_t, \hat{x}_t) - \alpha^*(w_{t-1}, \hat{x}_{t-1}) \|^2 \right] \leq 2E \left[ \| \alpha^*(w_{t-1}, \hat{x}_t) - \alpha^*(w_{t-1}, x^*_{t-1}) \|^2 \right] + \frac{\sigma^2}{L_{aw}^2}.
$$

(43)

Since $\alpha^*(\cdot)$ is $\frac{L_{aw}}{\mu}$-Lipschitz, we have

$$
E \left[ \| \alpha^*(w_t, \hat{x}_t) - \alpha^*(w_{t-1}, \hat{x}_{t-1}) \|^2 \right] \leq \frac{L_{aw}^2}{\mu^2} E \left[ \| w_t - w_{t-1} \|^2 \right].
$$

(44)

Furthermore, we have

$$
E \left[ \| w_t - w_{t-1} \|^2 \right] = \eta_\alpha^2 E \left[ \| \hat{g}(w) \|^2 \right] = 2\eta_\alpha^2 E \left[ \| g(w) \|^2 \right] + 2\eta_\alpha^2 E \left[ \| g(w) - g(w) \|^2 \right] \leq 2\eta_\alpha^2 \| \nabla \Phi(w_{t-1}, \alpha_{t-1}) \|^2 + \frac{2\eta_\alpha^2 \sigma^2}{\mu} + \frac{2\delta L_{aw}^2 \eta_\alpha^2}{\mu}.
$$

(45)

Plugging (44) and (45) into (43) and taking the exception of both side, yields that

$$
E \left[ \| \alpha^*(w_t, \hat{x}_t) - \alpha^*(w_{t-1}, \hat{x}_{t-1}) \|^2 \right] \leq \frac{8L_{aw}^2 \eta_\alpha^2 \delta_{t-1} + \frac{8L_{aw}^2 \eta_\alpha^2}{\mu^2}}{\mu^2} + \frac{4L_{aw}^2 \eta_\alpha^2 \sigma^2}{\mu^2} + 2E \left[ \| \alpha^*(w_{t-1}, \hat{x}_t) - \alpha^*(w_{t-1}, \hat{x}_{t-1}) \|^2 \right].
$$

(46)
Then, plugging (42) and (46) into (39) and let \( \epsilon_0 = \frac{8(\kappa-1)}{7-6\kappa} \) and \( \kappa \leq \frac{7}{6} \), yields that

\[
\mathbb{E} \left[ \| \alpha^* (w_t, \hat{x}_t) - \alpha_t (w_t, \hat{x}_t) \|^2 \right] \leq \left( \frac{1}{2} - \frac{\mu}{2L_{\alpha w}^2} + \frac{L_{\alpha w}^2 \eta_w^2}{4\mu^3} \right) \delta_{t-1} + \frac{L_{\alpha w}^2 \eta_w^2}{4\mu^3} \mathbb{E} \left[ \| \nabla \Phi (w_{t-1}) \|^2 \right] \\
+ \frac{\sigma^2}{2L_{\alpha w}^2 M} + \frac{L_{\alpha w}^2 \eta_w^2 \sigma^2}{8\mu^3 M} + \frac{L_{\alpha w}^2 \eta_w^2 \delta}{8\mu^3} \\
+ \frac{L_{\alpha w}}{16\mu} \mathbb{E} \left[ \| \alpha^* (w_{t-1}, \hat{x}_t) - \alpha^* (w_{t-1}, \hat{x}_{t-1}) \|^2 \right].
\]

(47)

Since \( \alpha^* (\cdot) \) is \( \frac{L_{\alpha w}}{\mu} \)-Lipschitz, we have

\[
\mathbb{E} \left[ \| \alpha^* (w_{t-1}, \hat{x}_t) - \alpha^* (w_{t-1}, \hat{x}_{t-1}) \|^2 \right] \leq \frac{L_{\alpha w}}{\mu^2} \mathbb{E} \left[ \| \hat{x}_t - \hat{x}_{t-1} \|^2 \right] \leq \frac{L_{\alpha w}}{\mu^2} \Delta,
\]

(48)

\[
\mathbb{E} \left[ \| \alpha^* (w_t, x_t^*) - \alpha^* (w_t, \hat{x}_t) \|^2 \right] \leq \frac{L_{\alpha w}}{\mu^2} \mathbb{E} \left[ \| x_t^* - \hat{x}_t \|^2 \right] \leq \frac{L_{\alpha w} \delta}{\mu^3},
\]

(49)

where \( \Delta \) is the maximum distance between two adversarial samples on the same sample. We can get the value of \( \Delta \) by the diameter if \( \mathcal{X}_t \).

Plugging (46) (47) and (48) into (38), yields that

\[
\delta_t \leq \left( 1 - \frac{\mu}{2L_{\alpha w}} + \frac{L_{\alpha w}^2 \eta_w^2}{2\mu^3} \right) \delta_{t-1} + \frac{L_{\alpha w}^2 \eta_w^2}{4\mu^3} \mathbb{E} \left[ \| \nabla \Phi (w_{t-1}) \|^2 \right] \\
+ \frac{2\sigma^2}{L_{\alpha w}^2 M} + \frac{L_{\alpha w}^2 \eta_w^2 \sigma^2}{4\mu^3 M} + \frac{L_{\alpha w}^2 \eta_w^2 \delta}{4\mu^3} + \frac{L_{\alpha w} \Delta}{8\mu} + \frac{L_{\alpha w}^2 \delta}{\mu^3}.
\]

\( \square \)

**Lemma 8.** Let \( \delta_t = \mathbb{E}[\| \alpha^* (w_t, x_t^*) - \alpha_t (w_t, \hat{x}_t) \|^2] \), the following statements holds that

\[
\mathbb{E}[\Phi(w_t)] \leq \mathbb{E}[\Phi(w_{t-1})] - \frac{\eta_w}{4} \mathbb{E} \left[ \| \nabla \Phi (w_{t-1}) \|^2 \right] + \frac{3\eta_w L^2 \delta_{t-1}}{4} + \frac{L_{\alpha w} \eta_w^2 \sigma^2}{M} + \frac{\delta L_{\alpha w} \eta_w^2}{\mu} + \eta_w L_{\alpha w} \sqrt{\frac{\delta}{\mu}}.
\]

**Proof.** Let \( \eta_w = \frac{1}{10(\kappa+1)^2} \), where \( L \) is the maximum value in the set \( \{ L_{w w}, L_{\alpha w}, L_{x w}, L_{w x}, L_{x \alpha}, L_{\alpha x} \} \), and \( \kappa = \frac{L}{\mu} \), hence

\[
\frac{1}{4} \eta_w \leq \frac{\eta_w}{2} - 2L_{w w} \eta_w \leq \frac{\eta_w}{2} + 2L_{w w} \eta_w \leq \frac{3}{4} \eta_w.
\]

(50)

Then we have

\[
\mathbb{E} \left[ \| \nabla \Phi (w_{t-1}) - \nabla_{w} L(w_{t-1}, \alpha_{t-1}) \|^2 \right] \leq L^2 \delta_{t-1}.
\]

(51)

Combining (50) (51) and Lem.6 yields that

\[
\mathbb{E}[\Phi(w_t)] \leq \mathbb{E}[\Phi(w_{t-1})] - \frac{\eta_w}{4} \mathbb{E} \left[ \| \nabla \Phi (w_{t-1}) \|^2 \right] + \frac{3\eta_w L^2 \delta_{t-1}}{4} + \frac{L_{\alpha w} \eta_w^2 \sigma^2}{M} + \frac{\delta L_{\alpha w} \eta_w^2}{\mu} + \eta_w L_{\alpha w} \sqrt{\frac{\delta}{\mu}}
\]

\( \square \)
A.4. Proof of Theorem 1

Proof. Throughout this subsection, we define $\gamma = 1 - \frac{1}{4\kappa} + \frac{L^2\kappa^3\mu^2}{2}$. Since $\delta_0 \leq D^2$, where $D = |\Omega|$, we have

$$\delta_t \leq \gamma^t D^2 + \frac{\kappa^3\eta^2}{4} \left( \sum_{j=0}^{t-1} \gamma^{t-j} \mathbb{E} \left[ \left\| \nabla \Phi(w_{t-1}) \right\|^2 \right] \right) + \left( \frac{2\sigma^2}{L^2M} + \frac{\kappa^3\eta^2\sigma^2}{4M} + \frac{L\kappa^2\eta^2\delta}{4} + \frac{2\kappa^2\delta}{8} + \kappa^3 \Delta \right) \left( \sum_{j=0}^{t-1} \gamma^{t-j} \right) \right) \right).$$

Combining (52) and Lem.8 yields that

$$\mathbb{E}[\Phi(w_t)] \leq \mathbb{E}[\Phi(w_{t-1})] - \eta \mathbb{E} \left[ \left\| \nabla \Phi(w_t) \right\|^2 \right] + \frac{L\kappa^2\eta^2\delta}{4} + \frac{2\kappa^2\delta}{8} + \kappa^3 \Delta \left( \sum_{j=0}^{t-2} \gamma^{t-j} \right)$$

Summing up (53) over $t = 1, 2, \cdots, T+1$ and rearranging the terms yields that

$$\mathbb{E}[\Phi(w_{T+1})] \leq \mathbb{E}[\Phi(w_0)] - \eta \sum_{t=0}^{T} \mathbb{E} \left[ \left\| \nabla \Phi(w_t) \right\|^2 \right] + (T+1) \left( \frac{L\kappa^2\eta^2\delta}{4M} + \frac{2\kappa^2\delta}{8} + \kappa^3 \Delta \left( \sum_{j=0}^{T-1} \gamma^{t-j} \right) \right)$$

Since $\eta = \frac{1}{10(\kappa+1)^2 \delta}$, we have $\gamma \leq 1 - \frac{1}{4\kappa}$ and $\frac{3\kappa^2\eta^2}{16} \leq \frac{3\eta}{4096\kappa}$. This implies that $\sum_{t=0}^{T} \gamma^t \leq 4\kappa$ and

$$\sum_{t=1}^{T+1} \sum_{j=0}^{t-2} \gamma^{t-j} \mathbb{E} \left[ \left\| \nabla \Phi(w_j) \right\|^2 \right] \leq 4\kappa \sum_{t=0}^{T} \mathbb{E} \left[ \left\| \nabla \Phi(w_t) \right\|^2 \right]$$

Putting these pieces together yields that

$$\mathbb{E}[\Phi(w_{T+1})] \leq \mathbb{E}[\Phi(w_0)] - \frac{253\eta}{1024} \sum_{t=0}^{T} \mathbb{E} \left[ \left\| \nabla \Phi(w_t) \right\|^2 \right] + \eta (T+1) \left( \frac{\sigma^2}{8\kappa M} + \frac{L\delta}{8} + \frac{L^2\mu}{2} \right)$$

Futhermore, we have

$$\frac{\sigma^2}{8\kappa M} + \frac{6\sigma^2\kappa}{M} + \frac{3\sigma^2}{1024M} \leq \frac{6403\kappa\sigma^2}{1024M}$$

(57)
By definition of $\Delta_\Phi$ and plugging (57) into (56), we have
\[
\frac{1}{T + 1} \sum_{t=0}^{T} \mathbb{E} \left[ \| \nabla \Phi(w_t) \|_2^2 \right] \leq \frac{1024 \Delta_\Phi}{253 \eta_w (T + 1)} + \frac{3072 L^2 D^2 \kappa}{253 (T + 1)} + \frac{6403 \kappa \sigma^2}{253 M} + \frac{1024 \Delta_\Phi}{253}
\]
\[
+ \frac{1024}{253} \left( \frac{3 \kappa L \delta}{1024} + 6 \kappa^4 L \delta + \frac{L \delta}{8} + L^2 \sqrt{\frac{\delta}{\mu}} \right) + \frac{384 \kappa^4 L^2 \Delta}{253}
\]
\[
\leq \frac{5 \Delta_\Phi}{\eta_w (T + 1)} + \frac{13 \kappa L^2 D^2}{T + 1} + \frac{26 \kappa \sigma^2}{M} + h_\delta + h_\Delta
\]
\[
\leq \frac{360 \kappa^2 L \Delta_\Phi + 13 \kappa L^2 D^2}{T + 1} + \frac{26 \kappa \sigma^2}{M} + h_\delta + h_\Delta
\]
where $h_\delta = \frac{1024}{253} \left( \frac{3 \kappa L \delta}{1024} + 6 \kappa^4 L \delta + \frac{L \delta}{8} + L^2 \sqrt{\frac{\delta}{\mu}} \right)$ and $h_\Delta = \frac{384 \kappa^4 L^2 \Delta}{253}$ and $\Delta_\Phi = \Phi(w^0) - \min_{w} \Phi(w)$. On the right side of (58), the second term $\frac{26 \kappa \sigma^2}{M}$ is due to random sampling, the third term $h_\delta$ is due to the use of an approximate solution $\hat{x}$ to the inner maximization problem instead of the optimal solution $x^*$, and the fourth term $h_\Delta$ is due to $X$ being a bounded set. And their values are very small.

B. Adversarial Attacks

Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) is a single-step attack that generates adversarial examples through a permutation along the gradient of the loss function with respect to the clean image feature vector as:
\[
x = x + \epsilon \cdot \text{sign} \left( \nabla_x \ell(h_{\theta}(x, y)) \right)
\]
(59)

Projected Gradient Descent (PGD) (Madry et al., 2018) starts from an initialization point that is uniformly sampled from the allowed $\epsilon$-ball centered at the clean image, and it expends FSGM by iteratively applying multiple small steps of permutation updating with respect to the current gradient as:
\[
x^k = \text{Proj} \left( x^{k-1} + \beta \cdot \text{sign} \left( \nabla_x \ell(h_{\theta}(x^{k-1}, y)) \right) \right)
\]
(60)

Carlini & Wagner (C&W) (Carlini & Wagner, 2017) is another powerful attack based on optimization, where an auxiliary variable $\omega$ is induced and an adversarial example constrained by $l_2$ norm is represented by $x' = \frac{1}{2} (\tanh \omega + 1)$. It can be optimized by:
\[
\arg \min \omega \left\{ c \cdot f(x') + \| x' - x \|_2^2 \right\}
\]
(61)

where
\[
f(x) = \max \left\{ \max_{i \neq y} Z(x' - Z(x'_y), -\kappa) \right\}
\]
(62)

and here $\kappa$ controls the confidence of the adversarial examples. It can also be extended to other $l_\infty$.

Auto Attack (AA) (Croce & Hein, 2020) is a combination of multiple attacks that forms a parameter-free and computationally affordable ensemble of attacks to evaluate adversarial robustness. The standard attacks include four selected attacks: APGD, targeted version of APGD-DLR and FAB, and Square Attack.

C. Additional Results

In this section, we provide additional results to further support the conclusions in the main text.

C.1. Experiment Setting

In the section, we provide additional results to further support the conclusions in the main text.
The initial learning rate $\eta_w$ is set as 0.01 with decay $5 \times 10^{-4}$, and the batch size is 128. And the initial learning rate $\eta_\alpha$ is set as 0.1. In the training process, we adopt a learning rate step decay schedules, which cut the learning rate by a constant factor 0.001 every 30 constant number of epochs for all methods.

C.2. Score Distribution

![Score Distribution](image)

**Figure 5.** Distribution of positive and negative example scores of CE on MNIST-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.

![Score Distribution](image)

**Figure 6.** Distribution of positive and negative example scores of our proposed AdAUC on MNIST-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.
Figure 7. Distribution of positive and negative example scores of CE on CIFAR-10-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.

Figure 8. Distribution of positive and negative example scores of AdAUC on CIFAR-10-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.
Figure 9. Distribution of positive and negative example scores of CE on CIFAR-100-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.

Figure 10. Distribution of positive and negative example scores of AdAUC on CIFAR-100-LT dataset. The first row represents the score distribution against different attacks under Natural Training, and the second row represents the score distribution under Adversarial Training.