

Achieve Optimal Adversarial Accuracy for Adversarial Deep Learning using Stackelberg Game*

Xiao-Shan Gao, Shuang Liu, and Lijia Yu
 Academy of Mathematics and Systems Science, Chinese Academy of Sciences
 University of Chinese Academy of Sciences

July 19, 2022

Abstract

Adversarial deep learning is to train robust DNNs against adversarial attacks, which is one of the major research focuses of deep learning. Game theory has been used to answer some of the basic questions about adversarial deep learning such as the existence of a classifier with optimal robustness and the existence of optimal adversarial samples for a given class of classifiers. In most previous work, adversarial deep learning was formulated as a simultaneous game and the strategy spaces are assumed to be certain probability distributions in order for the Nash equilibrium to exist. But, this assumption is not applicable to the practical situation. In this paper, we give answers to these basic questions for the practical case where the classifiers are DNNs with a given structure, by formulating the adversarial deep learning as sequential games. The existence of Stackelberg equilibria for these games are proved. Furthermore, it is shown that the equilibrium DNN has the largest adversarial accuracy among all DNNs with the same structure, when Carlini-Wagner’s margin loss is used. Trade-off between robustness and accuracy in adversarial deep learning is also studied from game theoretical aspect.

Keywords. Adversarial deep learning, Stackelberg game, optimal robust DNN, universal adversarial attack, adversarial accuracy, trade-off result.

1 Introduction

A major safety issue for deep learning [22] is the existence of adversarial samples [40], that is, it is possible to make little modifications to an input sample which are essentially imperceptible to the human eye, but the DNN outputs a wrong label or even any label given by the adversary. Existence of adversarial samples makes deep learning vulnerable in safety critical applications and *adversarial deep learning* has become a major research focus of deep learning [44]. The goal of adversarial deep learning is to train robust DNNs against adversarial attacks and well as developing more effective attack methods for generating adversarial samples.

Many adversarial defence models were proposed, including the adversarial training based on robust optimization [24, 49], the gradient masking and obfuscation approaches [1, 48], adversarial parameter attacks [41, 23, 47], universal adversaries [5, 27], randomized smoothing [9], and the

*This work is partially supported by NSFC grant No.12288201 and NKRDG grant No.2018YFA0704705.

adversarial sample detection [7]. Many attack methods are also proposed, including the white-box attacks based on gradient information of the DNN [6, 24, 30], the black-box attacks based on the transferability of the adversaries [31], the poisoning attacks for the input data [17, 36], and the physical world attacks [21, 2]. More details can be found in the survey [44].

Many of the defenses are found to be susceptible to new adversarial attacks, and stronger defences also are proposed against the new adversarial attacks. To break this loop of defences and attacks, a recent line of research based on game theory [14, 38] tries to establish more rigorous foundation for adversarial deep learning by answering questions such as [5, 8, 25, 32]:

Question Q₁: Does there exist a classifier which ensures optimal robustness against any adversarial attack?

Question Q₂: Does there exist optimal adversarial samples for a given class of classifiers and a given set of data distribution?

To answer these questions, the adversarial deep learning was formulated as a simultaneous game between the Classifier and the Adversary. The goal of the Classifier is to train a robust DNN. The goal of the Adversary is to create optimal adversarial samples. A *Nash equilibrium* of the game is a DNN \mathcal{C}^* and an attack \mathcal{A}^* , such that no player can benefit by unilaterally changing its strategy and thus gives an optimal solution to the adversarial deep learning. Existence of Nash equilibria was proved under various assumptions [5, 25, 32].

Despite the great progresses, questions **Q₁** and **Q₂** are not answered satisfactorily. The main reason is that in order for the Nash equilibrium to exist, both the Classifier and the Adversary are either assumed to be a convex set of probability distributions or measurable functions. However, in practice, DNNs with fixed structures are used and Nash equilibria do not exist in this case. In this paper, we will show that questions **Q₁** and **Q₂** can be answered positively for DNNs with a fixed structure by formulating the adversarial deep learning as Stackelberg games.

1.1 Main contributions

A positive answer to question **Q₁** is given by formulating the adversarial deep learning as a Stackelberg game \mathcal{G}_s with the Classifier as the leader and the Adversary as the follower, where the strategy space for the Classifier is a class of DNNs with a given structure, say DNNs with a fixed depth and width. We show that game \mathcal{G}_s has a Stackelberg equilibrium which gives the optimal robust DNN under certain robustness measurement (Refer to Theorem 3.5). We further show that when the Carlini-Wagner margin loss is used as the payoff function, the equilibrium DNN is the optimal defense which has the largest adversarial accuracy among all DNNs with the same structure (Refer to Theorem 4.4). Furthermore, the equilibrium DNN is the same as that of the adversarial training [24]. Thus, our results give another theoretical explanation for the fact that adversarial training is one of the most effective defences against adversarial attacks.

The trade-off property for deep learning means that there exists a trade-off between the robustness and accuracy [42, 45, 49]. We prove a trade-off result from game theoretical viewpoint. Precisely, we show that if a linear combination of the payoff functions of adversarial training and normal training is used as the total payoff function, then the equilibrium DNN has robustness not higher and accuracy no lower than that of the DNN obtained by adversarial training. We also show that trade-off property does not hold if using empirical loss to train the DNNs, that is, the DNNs with the largest adversarial accuracy can be parameterized by elements in an open set of \mathbb{R}^K , where K is the number of parameters, that is, there still exist rooms to improve the accuracy for DNNs with the optimal adversarial accuracy.

Finally, when using the empirical loss for a finite set of samples to train the DNN, we compare \mathcal{G}_s (denoted as \mathcal{G}_1 in this case) with two other games: \mathcal{G}_2 is the Stackelberg game with the Adversary as the leader and \mathcal{G}_3 is the simultaneous game between the Classifier and the Adversary. We show that \mathcal{G}_2 has a Stackelberg equilibrium and \mathcal{G}_3 has a mixed strategy Nash equilibrium. Furthermore, the payoff functions of $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ at their equilibria decrease successively. Existence of Stackelberg equilibrium for \mathcal{G}_2 gives a positive answer to question **Q₂** for DNNs with a given structure.

1.2 Related work

The game theoretical approach to adversarial machine learning was first studied in the seminal work of Dalvi, Domingos, Mausam, and Verma [11], where they formulated adversarial machine learning as a simultaneous game between the Classifier and the Adversary. Quite a number of work has been done along this line, by formulating adversarial machine learning both as a simultaneous game and as a Stackelberg game, which can be found in the nice surveys [50, 20]. These works usually used linear models such as SVM for binary classifications, and used spam email filtering as the main application background.

Game theoretical approach to adversarial deep learning appeared recently and was partially stimulated by the fact that adversarial samples seem inevitable for deep learning [3, 4, 10, 37]. The adversarial training was introduced in [24], which is one of the best practical training method to defend adversaries. In [32, 5, 16, 25, 29, 18, 34], the adversarial deep learning was all formulated as a simultaneous game. In [32], it was shown that the game exists no pure strategy Nash equilibrium, but mixed strategies give more robust classifiers. In [5], it was proved that Nash equilibrium exists when the strategy space for the Classifier is convex and the strategy space for the Adversary is certain probability distributions. In [16, 25], it was proved that Nash equilibria exist and can be approximated by a pure strategy, when the strategy spaces for both the Classifier and Adversary are parameterized by distributions. In [29], the Classifier ensures the robustness of a fixed DNN by adding perturbation to the sample to counteract the Adversary. In [18, 34], methods to compute mixed Nash equilibria were given. In [8], the adversarial deep learning was formulated as a Stackelberg game with the Adversary as the leader, but existence of equilibria was not given. In [13, 19], properties and algorithms for local Stackelberg equilibria were studied. In above work, the adversarial deep learning is modeled as a non-cooperative game. In [35], the cooperative game is used to explain various adversarial attacks and defenses.

Most of the above work formulated adversarial deep learning as a simultaneous game and assume the strategy spaces to be certain convex probability distributions in order to prove the existence of the Nash equilibrium. In this paper, we show that by formulating the adversarial deep learning as a sequential game, Stackelberg equilibria exist for DNNs with a given structure, and the equilibrium DNN is the best defence in that it has the largest adversarial accuracy among all DNNs with the same structure.

The rest of this paper is organized as follows. In section 2, preliminary results are given. In section 3, the adversarial deep learning is formulated as a Stackelberg game and the existence of Stackelberg equilibria is proved. In section 4, it is proved that adversarial training with Carlini-Wagner loss gives the best adversarial accuracy. In section 5, two trade-off results are proved. In section 6, three types of adversarial games are compared when the data set is finite. In section 7, conclusions and problems for further study are given.

2 Preliminaries

2.1 Adversarial training and robustness of DNN

Let $\mathcal{C} : \mathcal{X} \rightarrow \mathbb{R}^m$ be a classification DNN with m labels in $\mathcal{Y} = [m] = \{1, \dots, m\}$ [22]. Without loss of generality, we assume $\mathcal{X} = \mathbb{I}^n$, where $\mathbb{I} = [0, 1]$. Denote $\mathcal{C}_l(x) \in \mathbb{R}$ to be the l -th coordinate of $\mathcal{C}(x)$ for $l \in [m]$, which are called *logits* of the DNN. For $x \in \mathcal{X}$, the classification result of \mathcal{C} is $\hat{\mathcal{C}}(x) = \operatorname{argmax}_{l \in \mathcal{Y}} \mathcal{C}_l(x)$. We assume that Relu is used as the *activation function*, so \mathcal{C} is continuous and piecewise linear. The results are easily generated to any activation functions which are Lipschitz continuous.

To train a DNN, we need first to choose a *hypothesis space* \mathcal{H} for the DNNs, say the set of CNNs or RNNs with certain fixed structure. In this paper, denote $\mathcal{N}_{W,D}$ to be the set of DNNs with width W and depth D and use it as the hypothesis space. For a given hypothesis space \mathcal{H} , the parameter set of DNNs in \mathcal{H} is fixed and is denoted as $\Theta \in \mathbb{R}^K$, where K is the number of the parameters. \mathcal{C} can be written as \mathcal{C}_Θ if the parameters need to be mentioned explicitly, that is,

$$\mathcal{H} = \{\mathcal{C}_\Theta : \mathcal{X} \rightarrow \mathbb{R}^m : \Theta \in \mathbb{R}^K\}. \quad (1)$$

Let the objects to be classified satisfy a distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$. Given a loss function $\mathbf{L} : \mathbb{R}^m \times \mathcal{Y} \rightarrow \mathbb{R}$, the total loss for the data set is

$$\varphi_0(\Theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{L}(\mathcal{C}_\Theta(x), y). \quad (2)$$

Training a DNN \mathcal{C}_Θ is to make the total loss minimum by solving the following optimization problem

$$\Theta^* = \operatorname{argmin}_{\Theta \in \mathbb{R}^K} \varphi_0(\Theta). \quad (3)$$

Given an *attack radius* $\varepsilon \in \mathbb{R}_+$, denote $\mathbb{B}(x, \varepsilon) = \{\bar{x} \in \mathbb{R}^n : \|\bar{x} - x\| \leq \varepsilon\}$. We use ∞ norm if not mentioned otherwise. We will find adversaries for x in $\mathbb{B}(x, \varepsilon)$. Precisely, $\bar{x} \in \mathbb{B}(x, \varepsilon)$ is called an *adversary of x* with label y , if $\hat{\mathcal{C}}(\bar{x}) \neq y$. In order to increase the robustness of a trained DNN, the *adversarial training* [24] is introduced which is to solve the following robust optimization problem

$$\Theta^* = \operatorname{argmin}_{\Theta \in \mathbb{R}^K} \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}(\mathcal{C}_\Theta(\bar{x}), y). \quad (4)$$

Intuitively, the adversarial training is first computing a *most-adversarial sample*

$$x_a = \operatorname{argmax}_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}(\mathcal{F}(\bar{x}), l_x)$$

for x and then minimizing $\mathbf{L}(\mathcal{F}(x_a), y)$ instead of $\mathbf{L}(\mathcal{F}(x), y)$.

Given a DNN \mathcal{C} and an attack radius ε , we define the *adversarial robustness measure* of \mathcal{C} with respect to ε as follows

$$\operatorname{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}(\mathcal{C}(\bar{x}), y) \quad (5)$$

which is the total loss of \mathcal{C} at the most-adversarial samples. \mathcal{C} is more robust if $\operatorname{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon)$ is smaller. Then the adversarial training is to find a DNN in \mathcal{H} with the optimal adversarial robustness measurement which is denoted as

$$\operatorname{AR}_{\mathcal{D}}(\mathcal{H}, \varepsilon) = \min_{\Theta \in \mathbb{R}^K} \operatorname{AR}_{\mathcal{D}}(\mathcal{C}_\Theta, \varepsilon). \quad (6)$$

$\text{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon)$ and $\text{AR}_{\mathcal{D}}(\mathcal{H}, \varepsilon)$ have the following simple properties.

- (1) If $W_1 \geq W_2$ and $D_1 \geq D_2$, then $\text{AR}_{\mathcal{D}}(\mathcal{N}_{W_1, D_1}, \varepsilon) \leq \text{AR}_{\mathcal{D}}(\mathcal{N}_{W_2, D_2}, \varepsilon)$.
- (2) If $\varepsilon_1 \leq \varepsilon_2$, then $\text{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon_1) \leq \text{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon_2)$.

(3) In the optimal case, we have $\text{AR}_{\mathcal{D}}(\mathcal{C}, \varepsilon) = 0$, which means that \mathcal{C} gives the correct label for any $\bar{x} \in \mathbb{B}(x, \varepsilon)$. In this case, we say that \mathcal{C} is *robust* for the attack radius ε . It was proved that there exist robust classifiers for a separated data set [45].

2.2 Bounds and continuity of the DNN

Let $C_{\Theta} : \mathcal{X} \rightarrow \mathbb{R}^m$ be a fully connected feed-forward DNN with depth D , whose l -th hidden layer is

$$x_l = \sigma(W_l x_{l-1} + b_l) \in \mathbb{R}^{n_l}, l = 1, \dots, D, \quad (7)$$

where $n_0 = n$, $n_D = m$, $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$, $b_l \in \mathbb{R}^{n_l}$, $\sigma = \text{Relu}$, $x_0 \in \mathbb{R}^n$ is the input, and $x_D \in \mathbb{R}^m$ is the output. The parameter set is $\Theta = \cup_{l=1}^D (W_l \cup b_l)$. It is easy to show that \mathcal{C} is bounded. For $\varepsilon \in \mathbb{R}_+$, denote $\mathbb{I}_{\varepsilon} = [-\varepsilon, 1 + \varepsilon]$.

Lemma 2.1. *For any DNN $C_{\Theta} : \mathbb{I}_{\varepsilon}^n \rightarrow \mathbb{R}^m$ with width $\leq W$, depth $\leq D$, and $\|\Theta\|_2 \leq E$, there exists an $\Omega(n, m, D, W, E, \varepsilon) \in \mathbb{R}_+$ such that $\|C_{\Theta}(x)\| \leq \Omega(n, m, D, W, E, \varepsilon)$.*

Proof. $C_{\Theta}(x)$ is bounded because $C_{\Theta}(x)$ is continuous on x and Θ , and $[-\varepsilon, 1 + \varepsilon]^n$ and $[-E, E]^n$ are compact. $\Omega(n, m, D, W, E, \varepsilon)$ can be derived from (7). \square

Lemma 2.2. *For any DNN $C_{\Theta} : \mathbb{I}_{\varepsilon}^n \rightarrow \mathbb{R}^m$ with width $\leq W$, depth $\leq D$, and $\|\Theta\|_2 \leq E$, there exist $\Delta(m, n, W, D, E, \varepsilon)$ and $\Lambda(m, n, W, D, E, \varepsilon) \in \mathbb{R}_+$ such that*

- (1) $\|C_{\Theta}(x) - C_{\Theta+\alpha}(x)\|_2 \leq \Delta(m, n, W, D, E, \varepsilon)\|\alpha\|_2$, that is $C_{\Theta}(x)$ is Lipschitz on Θ .
- (2) $\|C_{\Theta}(x + \delta) - C_{\Theta}(x)\|_2 \leq \Lambda(m, n, W, D, E, \varepsilon)\|\delta\|$, that is $C_{\Theta}(x)$ is Lipschitz on x .

Thus \mathcal{C} is Lipschitz on Θ and x .

Proof. Without loss of generality, let \mathcal{C} be defined as in (7). Then $C_{\Theta}(x) = \Theta^D(\dots \sigma(\Theta^1 x) \dots)$ with Θ to be the set of all weight matrices, that is, $\Theta = \{\Theta^k | \forall k \in [D] = \{1, 2, \dots, D\}\}$ and σ is ReLU. The bias vectors are not considered, which can be included as parts of the weight matrices by extending the input space slightly, similar to [28]. We denote z_k and \hat{z}_k respectively to be the outputs of the k -th hidden layers of C_{Θ} and $C_{\Theta+\alpha}$, which are $z_k = \sigma(\Theta^k(\dots \sigma(\Theta^1 x) \dots))$ and $\hat{z}_k = \sigma(\hat{\Theta}^k(\dots \sigma(\hat{\Theta}^1 x) \dots))$ and $\hat{\Theta}^i$ is weight matrices of $C_{\Theta+\alpha}$, in particular $z_0 = \hat{z}_0 \in [-\varepsilon, 1 + \varepsilon]^n$ is the input. Since $\|\Theta^i - \hat{\Theta}^i\|_2 \leq \|\alpha\|_2$ for any $i \in [D]$ and $|\sigma(a) - \sigma(b)| \leq |a - b|$, we have

$$\begin{aligned} & \|C_{\Theta}(x) - C_{\Theta+\alpha}(x)\|_2 \\ &= \|(\Theta^D - \hat{\Theta}^D)z_{D-1} + \hat{\Theta}^D(z_{D-1} - \hat{z}_{D-1})\|_2 \\ &\leq \|\Theta^D - \hat{\Theta}^D\|_2 \|z_{D-1}\|_2 + \|\hat{\Theta}^D\|_2 \|z_{D-1} - \hat{z}_{D-1}\|_2 \\ &= \|\Theta^D - \hat{\Theta}^D\|_2 \|z_{D-1}\|_2 + \|\hat{\Theta}^D\|_2 \|\sigma(\Theta^{D-1}z_{D-2}) - \sigma(\hat{\Theta}^{D-1}\hat{z}_{D-2})\|_2 \\ &\leq \|\Theta^D - \hat{\Theta}^D\|_2 \|z_{D-1}\|_2 + \|\hat{\Theta}^D\|_2 \|\Theta^{D-1}z_{D-2} - \hat{\Theta}^{D-1}\hat{z}_{D-2}\|_2 \\ &\leq \|\Theta^D - \hat{\Theta}^D\|_2 \|z_{D-1}\|_2 + \|\hat{\Theta}^D\|_2 (\|\Theta^{D-1} - \hat{\Theta}^{D-1}\|_2 \|z_{D-2}\|_2 + \|\hat{\Theta}^{D-1}\|_2 \|z_{D-2} - \hat{z}_{D-2}\|_2) \\ &\leq \|\Theta^D - \hat{\Theta}^D\|_2 \|z_{D-1}\|_2 + \sum_{k=2}^D (\prod_{i=0}^{k-2} \|\hat{\Theta}^{D-i}\|_2) \|\Theta^{D-k+1} - \hat{\Theta}^{D-k+1}\|_2 \|z_{D-k}\|_2 \\ &\leq (\|z_{D-1}\|_2 + \sum_{k=2}^D (\prod_{i=0}^{k-2} \|\hat{\Theta}^{D-i}\|_2) \|z_{D-k}\|_2) \|\alpha\|_2. \end{aligned}$$

The coefficient $\Delta = (\|z_{D-1}\|_2 + \sum_{k=2}^D (\prod_{i=0}^{k-2} \|\widehat{\Theta}^{D-i}\|_2) \|z_{D-k}\|_2)$ is clearly bounded and depends $m, n, W, D, E, \varepsilon$. Thus $\mathcal{C}_\Theta(x)$ is Lipschitz on Θ . The Lipschitz continuity on x can be proved similarly:

$$\begin{aligned}
& \|\mathcal{C}_\Theta(x + \delta) - \mathcal{C}_\Theta(x)\|_2 \\
&= \|\Theta^D(\cdots \sigma \Theta^1(x + \delta) \cdots) - \Theta^D(\cdots \sigma \Theta^1(x) \cdots)\|_2 \\
&\leq \|\Theta^D\|_2 \|\sigma(\Theta^{D-1}(\cdots \sigma(\Theta^1(x + \delta)) \cdots)) - \sigma(\Theta^{D-1}(\cdots \sigma(\Theta^1(x)) \cdots))\|_2 \\
&\leq \|\Theta^D\|_2 \|\Theta^{D-1}(\cdots \sigma \Theta^1(x + \delta) \cdots) - \Theta^{D-1}(\cdots \sigma \Theta^1(x) \cdots)\|_2 \\
&\leq (\prod_{i=1}^D \|\Theta^i\|_2) \|\delta\|_2 \leq (\prod_{i=1}^D \|\Theta^i\|_2) \sqrt{n} \|\delta\|.
\end{aligned}$$

We denote the coefficient as $\Lambda(m, n, W, D, E, \varepsilon)$. The lemma is proved. We can also extend this result to convolutional neural networks. \square

2.3 Continuity of the loss function

Unless mentioned otherwise, we assume that the loss function $\mathbf{L}(z, y)$ is continuous on $z \in \mathbb{R}^m$ for a fixed $y \in \mathcal{Y}$. The mostly often used loss functions have much better properties. Consider the following loss functions: the mean square error, the crossentropy loss, and the margin loss introduced by Carlini-Wagner [6]:

$$\begin{aligned}
\mathbf{L}_{\text{mse}}(z, y) &= \|z - \mathbf{1}_y\|_2^2 \\
\mathbf{L}_{\text{ce}}(z, y) &= \ln(\sum_{i=1}^m \exp(z_i)) - z_y \\
\mathbf{L}_{\text{cw}}(z, y) &= \max_{l \in [m], l \neq y} z_l - z_y
\end{aligned} \tag{8}$$

where $\mathbf{1}_y \in \mathbb{R}^m$ is the vector whose y -th entry is 1 and all other entries are 0.

By Lemma 2.1, we can assume that the loss function is defined on a bounded cube:

$$\mathbf{L}(z, y) : [-B, B]^m \times \mathcal{Y} \rightarrow \mathbb{R} \tag{9}$$

where $B = \Omega(n, m, D, W, E, \varepsilon)$. Since $\mathcal{Y} = [m]$ is discrete, we need only consider the continuity of \mathbf{L} on z for a fixed y .

Lemma 2.3. *For a fixed y , all three loss functions in (8) are Lipschitz continuous on z over $[-B, B]^m$, with Lipschitz constants $2\sqrt{m} \max\{B, 1\}$, $\sqrt{2}$, $\sqrt{2}$, respectively.*

Proof. It suffices to show that $\|\nabla_z F(z)\|_2 \leq V$ is bounded over $[-B, B]^m$. For a fixed y , let $f(z) = \mathbf{L}(z, y)$. Then from $\|\nabla_z F(z)\|_2 \leq V$, by the mean value theorem and the Schwarz inequality, we have $\|F(z + \delta) - F(z)\|_2 = \|F'(z_1)\delta\|_2 \leq \|F'(z_1)\|_2 \|\delta\|_2 \leq V \|\delta\|_2$, where $z_1 \in (-B, B)^m$. Thus \mathbf{L} is Lipschitz with constant V .

For \mathbf{L}_{mse} , we have $\|\nabla_z \mathbf{L}_{\text{mse}}(z, y)\|_2 = 2\|(z - \mathbf{1}_y)\|_2 \leq 2\sqrt{m} \max\{B, 1\}$. For \mathbf{L}_{ce} , we have $\|\nabla_z \mathbf{L}_{\text{ce}}(z, y)\|_2 = \sqrt{\frac{\sum_{i=1}^m \exp(2z_i) + (\sum_{i=1}^m \exp(z_i))^2}{(\sum_{i=1}^m \exp(z_i))^2}} \leq \sqrt{2}$. For \mathbf{L}_{cw} , we have $\|\nabla_z \mathbf{L}_{\text{cw}}(z, y)\|_2 = \sqrt{2}$. The lemma is proved. \square

3 Adversarial training as a Stackelberg game

In this section, we formulate the adversarial deep learning as a Stackelberg game and prove the existence of the Stackelberg equilibria.

3.1 Stackelberg game

Consider a two-player zero-sum minmax sequential or Stackelberg game $\mathcal{G} = (\mathcal{S}_L, \mathcal{S}_F, \varphi)$, where \mathcal{S}_L and \mathcal{S}_F are respectively the strategy spaces for the leader and the follower of the game and $\varphi : \mathcal{S}_L \times \mathcal{S}_F \rightarrow \mathbb{R}$ is the payoff function.

In the Stackelberg game \mathcal{G} , the leader moves first by picking a strategy $s_l \in \mathcal{S}_L$ to minimize the payoff, knowing the existence of the follower. After knowing s_l , the follower picks $s_f \in \mathcal{S}_F$ to maximize the payoff. Formally, $(s_l^*, s_f^*) \in \mathcal{S}_L \times \mathcal{S}_F$ is called a *Stackelberg equilibrium* of \mathcal{G} if

$$\gamma(s_l) = \{\operatorname{argmax}_{s_f \in \mathcal{S}_F} \varphi(s_l, s_f)\} \subset \mathcal{S}_F \quad (10)$$

is not empty for any $s_l \in \mathcal{S}_L$, and

$$s_l^* \in \operatorname{argmin}_{s_l \in \mathcal{S}_L, S(s_l) \in \gamma(s_l)} \varphi(s_l, S(s_l)) \text{ and } s_f^* \in \operatorname{argmax}_{s_f \in \mathcal{S}_F} \varphi(s_l^*, s_f) = \gamma(s_l^*). \quad (11)$$

Let

$$\Gamma = \{(s_l, s_f) : s_l \in \mathcal{S}_L, s_f \in \gamma(s_l)\}. \quad (12)$$

Then, (11) is equivalent to $(s_l^*, s_f^*) \in \operatorname{argmin}_{(s_l, s_f) \in \Gamma} \varphi(s_l, s_f)$. We have the following result.

Theorem 3.1 ([39]). *If the strategy spaces are compact and the payoff function is continuous, then the sequential game \mathcal{G} has a Stackelberg equilibrium, which is also a subgame perfect Nash equilibrium of game \mathcal{G} as an extensive form game [14].*

3.2 Adversarial training as a Stackelberg game

We formulate adversarial deep learning as a two-player zero-sum minmax Stackelberg game \mathcal{G}_s , which is the best defence for adversarial deep learning in certain sense.

The leader of the game is the Classifier, whose goal is to train a robust DNN $\mathcal{C}_\Theta : \mathbb{I}^n \rightarrow \mathbb{R}^m$ in the hypothesis space \mathcal{H} in (1). Without loss of generality, we assume that the parameters of \mathcal{C} are in

$$\mathcal{S}_c = [-E, E]^K \quad (13)$$

for some $E \in \mathbb{R}_+$, that is, the *strategy space* for the Classifier is \mathcal{S}_c .

The follower of the game is the Adversary, whose goal is to create the best adversary within a given attack radius $\varepsilon \in \mathbb{R}_+$. The strategy space for the Adversary is

$$\mathcal{S}_a = \{A : \mathcal{X} \rightarrow \mathbb{B}_\varepsilon\} \quad (14)$$

where $\mathbb{B}_\varepsilon = \{\delta \in \mathbb{R}^n : \|\delta\| \leq \varepsilon\}$ is the ball with the origin point as the center and ε as the radius. By considering the L_∞ norm, \mathcal{S}_a becomes a metric space.

The payoff function. Given $\Theta \in \mathcal{S}_c$ and $A \in \mathcal{S}_a$, the payoff function is the expected loss

$$\varphi_s(\Theta, A) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y). \quad (15)$$

From (9), the composition of \mathbf{L} and $\mathcal{C}_\Theta(x + A(x))$ is well-defined, since $\|A(x)\| \leq \varepsilon$.

For game \mathcal{G}_s , γ and Γ defined in (10) and (12) are

$$\begin{aligned} \gamma_s(\Theta) &= \{\operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta, A)\} \text{ for } \Theta \in \mathcal{S}_c \\ \Gamma_s &= \{(\Theta, A) : \Theta \in \mathcal{S}_c, A \in \gamma_s(\Theta)\} \end{aligned} \quad (16)$$

and (Θ_s^*, A_s^*) is a Stackelberg equilibrium of \mathcal{G}_s if

$$\Theta_s^* \in \operatorname{argmin}_{\Theta \in \mathcal{S}_c, A(\Theta) \in \gamma_s(\Theta)} \varphi_s(\Theta, A(\Theta)) \text{ and } A_s^* \in \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta_s^*, A). \quad (17)$$

Lemma 3.2. $\varphi_s(\Theta, A) : \mathcal{S}_c \times \mathcal{S}_a \rightarrow \mathbb{R}$ defined in (15) is a continuous and bounded function.

Proof. It is clear that $\varphi_s(\Theta, A)$ is continuous on Θ , since \mathbf{L} is continuous on z and \mathcal{C}_Θ is continuous on Θ . Denote $\phi(x) = \mathbf{L}(\mathcal{C}_\Theta(x), y) : \mathbb{I}_\epsilon^n \rightarrow \mathbb{R}$ for fixed Θ and y . Then $\phi(x)$ is uniformly continuous by Lemmas 2.2 and 2.3. Given an $A_0 \in \mathcal{S}_a$ and $\epsilon > 0$, since $\phi(x)$ is uniformly continuous, there exists a $\delta > 0$ such that for $A(x) \in \mathcal{S}_a$ satisfying $\|A_0(x) - A(x)\|_\infty < \delta$, we have $|\phi(x + A_0(x)) - \phi(x + A(x))| < \epsilon$ for all $x \in \mathcal{X}$. Then

$$\begin{aligned} |\varphi_s(\Theta, A) - \varphi_s(\Theta, A_0)| &= |\mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y) - \mathbf{L}(\mathcal{C}_\Theta(x + A_0(x)), y)]| \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} |\mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y) - \mathbf{L}(\mathcal{C}_\Theta(x + A_0(x)), y)| \\ &\leq \epsilon. \end{aligned}$$

Hence $\varphi_s(\Theta, A)$ is continuous on \mathcal{S}_a . By Lemma 2.1, $\varphi_s(\Theta, A)$ is bounded, since $\|A(x)\| \leq \epsilon$. \square

Lemma 3.3. $\gamma_s(\Theta) \neq \emptyset$ and $A^* \in \gamma_s(\Theta)$ if and only if $A^*(x) \in \{\operatorname{argmax}_{A(x) \in \mathbb{B}_\epsilon} \mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y)\}$ for all $(x, y) \sim \mathcal{D}$.

Proof. We have

$$\begin{aligned} \max_{A \in \mathcal{S}_a} \varphi_s(\Theta, A) &= \max_{A \in \mathcal{S}_a} \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y) \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{A(x) \in \mathbb{B}_\epsilon} \mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y). \end{aligned}$$

Since $\mathbf{L}(\mathcal{C}(x), y)$ is continuous on x and \mathbb{B}_ϵ is compact, for every (x, y) , $\operatorname{argmax}_{A(x) \in \mathbb{B}_\epsilon} \mathbf{L}(\mathcal{C}_\Theta(x + A(x)), y)$ exists. Thus, by choosing these maximum values, we obtain an $A^* \in \mathcal{S}_a$, which achieves $\max_{A \in \mathcal{S}_a} \varphi_s(\Theta, A)$. The lemma is proved. \square

Lemma 3.4. Γ_s is a closed set in $\mathcal{S}_c \times \mathcal{S}_a$.

Proof. Let $(\Theta_i, A_i)_{i=1}^\infty \in \Gamma_s$ converge to (Θ_0, A_0) . Supposing $(\Theta_0, A_0) \notin \Gamma_s$, we will obtain a contradiction. By Lemma 3.3, there exists a $(\Theta_0, A^*) \in \Gamma_s$, and thus, $\varphi_s(\Theta_0, A^*) > \varphi_s(\Theta_0, A_0)$ by (16). Let $\eta = \varphi_s(\Theta_0, A^*) - \varphi_s(\Theta_0, A_0) > 0$. By Lemma 3.2, φ_s is continuous. Then there exists an i_0 such that $|\varphi_s(\Theta_{i_0}, A_{i_0}) - \varphi_s(\Theta_0, A_0)| < \eta/3$ and $|\varphi_s(\Theta_{i_0}, A^*) - \varphi_s(\Theta_0, A^*)| < \eta/3$. We thus have

$$\varphi_s(\Theta_{i_0}, A_{i_0}) < \varphi_s(\Theta_0, A_0) + \eta/3 = \varphi_s(\Theta_0, A^*) - \frac{2\eta}{3} < \varphi_s(\Theta_{i_0}, A^*) - \eta/3 < \varphi_s(\Theta_{i_0}, A^*)$$

which contradicts to $(\Theta_{i_0}, A_{i_0}) \in \Gamma_s$ meaning that $\varphi_s(\Theta_{i_0}, A_{i_0}) \geq \varphi_s(\Theta_{i_0}, A)$ for any $A \in \mathcal{S}_a$. The lemma is proved. \square

We have

Theorem 3.5. Game G_s has a Stackelberg equilibrium (Θ_s^*, A_s^*) . Furthermore, Θ_s^* is the solution to the adversarial training in (4).

Proof. By Lemma 3.2, $\varphi_s(\Theta, A)$ is bounded. Then $\alpha = \inf_{(\Theta, A) \in \Gamma_s} \varphi_s(\Theta, A)$ exists and is finite. There exist $(\Theta_i, A_i)_{i=1}^\infty \in \Gamma_s$ such that $\varphi_s(\Theta_i, A_i)$ converges to α . Since \mathcal{S}_c is compact, we can assume that Θ_i converges to Θ_0 . Then there exists an $A_0 \in \mathcal{S}_a$ such that $(\Theta_0, A_0) \in \Gamma_s$.

We claim that $\varphi_s(\Theta_i, A_i)$ converges to $\varphi_s(\Theta_0, A_0)$. Suppose the contrary, that is, $\varphi_s(\Theta_0, A_0) > \alpha$. Then there exists an $\eta > 0$ such that $\varphi_s(\Theta_0, A_0) > \alpha + \eta$. Since $\varphi_s(\Theta_i, A_i)$ converges to

α , $\exists K_1 \in \mathbb{N}_+$ such that $\varphi_s(\Theta_k, A_k) < \alpha + \frac{\eta}{3}$ for $\forall k > K_1$. Since $\varphi_s(\Theta, A)$ is continuous on Θ , $\exists K_2 \in \mathbb{N}_+$ such that $\varphi_s(\Theta_k, A_0) > \varphi_s(\Theta_0, A_0) - \frac{\eta}{3}$ for $\forall k > K_2$. Then for $k > \max\{K_1, K_2\}$, we have

$$\varphi_s(\Theta_k, A_0) > \varphi_s(\Theta_0, A_0) - \frac{\eta}{3} > \alpha + \frac{2\eta}{3} > \varphi_s(\Theta_k, A_k) + \frac{\eta}{3} > \varphi_s(\Theta_k, A_k)$$

which contradicts to $(\Theta_k, A_k) \in \Gamma_s$. Then (Θ_0, A_0) is a Stackelberg equilibrium of game \mathcal{G}_s .

Let (Θ_s^*, A_s^*) be a Stackelberg equilibria of game \mathcal{G}_s . By Lemma 3.3,

$$\begin{aligned} \Theta_s^* &\in \operatorname{argmin}_{\Theta \in \mathcal{S}_c, A(\Theta) \in \gamma(\Theta)} \varphi_s(\Theta, A(\Theta)) \\ &= \operatorname{argmin}_{\Theta \in \mathcal{S}_c, A_\Theta \in \gamma(\Theta)} \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{L}(\mathcal{C}_\Theta(x + A_\Theta(x)), y) \\ &\in \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{A_\Theta(x)} \mathbf{L}(\mathcal{C}_\Theta(x + A_\Theta(x)), y) \\ &= \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \mathbb{E}_{(x,y) \sim \mathcal{D}} \max_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}(\mathcal{C}_\Theta(\bar{x}), y). \end{aligned}$$

Briefly,

$$\begin{aligned} \Theta_s^* &= \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \varphi_s(\Theta, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta, A)) \\ &= \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \max_{A \in \mathcal{S}_a} \varphi_s(\Theta, A). \end{aligned} \tag{18}$$

That is, Θ_s^* is the solution to the adversarial training (4). \square

Remark 3.6. As a consequence of Theorem 3.5, the Stackelberg game \mathcal{G}_s gives the best defence in the hypothesis space \mathcal{H} for a given attack radius, if using $\operatorname{AR}_{\mathcal{D}}$ in (6) to measure the robustness. Precisely, let (Θ_s^*, A_s^*) be a Stackelberg equilibrium of game G_s . Then $\operatorname{AR}_{\mathcal{D}}(\mathcal{C}_{\Theta_s^*}, \varepsilon) = \operatorname{AR}_{\mathcal{D}}(\mathcal{H}, \varepsilon)$.

3.3 Refined properties of Γ_s

In the general case, $\gamma_s(\Theta)$ defined in (16) may have more than one elements. In this section, we will prove that if $\gamma_s(\Theta)$ contains a unique element, then Γ_s defined in (16) is compact, which will be used in section 6.

Assumption A_1 . For any $\Theta \in \mathcal{S}_c$, $\gamma_s(\Theta) = \{A^*(\Theta)\}$ defined in (16) has a unique element and the loss function \mathbf{L} is Lipschitz.

Remark 3.7. Assumption A_1 is true in the generic case. By Lemma 3.3, $A^* \in \gamma_s(\Theta)$ if and only if $A^*(x) \in \{\operatorname{argmax}_{A \in \mathbb{B}_\varepsilon} \mathbf{L}(\mathcal{C}_\Theta(x + A), y)\}$. Then Assumption A_1 is true if and only if $\operatorname{argmax}_{A \in \mathbb{B}_\varepsilon} \mathbf{L}(\mathcal{C}_\Theta(x + A), y)$ has a unique solution. Suppose the loss function is \mathbf{L}_{cw} . Then $\phi(A) = \mathbf{L}(\mathcal{C}_\Theta(x + A), y)$ is a piecewise linear function in A and its graph over \mathbb{B}_ε is a polyhedron as illustrated in Figure 1. Then its maximum can be achieved only at the vertex of the polyhedron or the intersection of the $(n-1)$ -dimensional sphere $\|x - x_0\| = \varepsilon$ and the one dimensional edges of the polyhedron. In the generic case, that is, when the parameters are sufficiently general (refer to Assumption 3.1 in [46] for more details), there exists only one maximum.

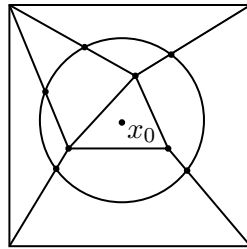


Figure 1: Illustration for the graph of $\mathbf{L}(\mathcal{C}(x + A), y)$ as a function of x and A .

We first introduce three notations which will be used in this section. By Lemma 2.3, $\mathbf{L}(z, y)$ is Lipschitz for z over $[-B, B]^m$ when the loss functions in (8) are used, and let Ψ be the Lipschitz constant. By Lemma 2.2, $\mathcal{C}_\Theta(x)$ is Lipschitz for Θ and x , and let Δ and Λ be the Lipschitz constants, respectively.

Lemma 3.8. *For any $\mathcal{C}_\Theta : \mathbb{I}_\epsilon^n \rightarrow \mathbb{R}^m$ and \mathcal{D} , $\varphi_s(\Theta, A)$ defined in (15) is Lipschitz on Θ and A when the loss function is Lipschitz.*

Proof. Firstly, consider $\varphi_s(\Theta, A)$ for any fixed A . For any $\epsilon > 0$, let $\delta = \frac{\epsilon}{\Psi\Delta}$. Then for any Θ_1, Θ_2 satisfying $\|\Theta_1 - \Theta_2\|_2 \leq \delta$, we have

$$\begin{aligned} |\varphi_s(\Theta_1, A) - \varphi_s(\Theta_2, A)| &= |\mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbf{L}(\mathcal{C}_{\Theta_1}(x + A(x)), y) - \mathbf{L}(\mathcal{C}_{\Theta_2}(x + A(x)), y)]| \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} \Psi \|\mathcal{C}_{\Theta_1}(x + A(x)) - \mathcal{C}_{\Theta_2}(x + A(x))\|_2 \\ &\leq \mathbb{E}_{(x,y) \sim \mathcal{D}} \Psi \Delta \|\Theta_1 - \Theta_2\|_2 \leq \epsilon \end{aligned}$$

that is, $\varphi_s(\Theta, A)$ is Lipschitz continuous on Θ . The proof for the Lipschitz continuity on A is similar. \square

Lemma 3.9. *For $\Theta_i \in \mathcal{S}_c$, if $\lim_{i \rightarrow \infty} \Theta_i = \Theta_0$ and $g_i \in \gamma_s(\Theta_i)$, then for any $(x, y) \sim \mathcal{D}$, the limit of any convergent subsequence of $\{g_i(x)\}_{i=1}^\infty$ belongs to $\operatorname{argmax}_{A \in \mathbb{B}_\epsilon} L(\mathcal{C}_{\Theta_0}(x + A), y)$.*

Proof. The result can be proved similar to that of Lemma 3.4. \square

Lemma 3.10. *Under Assumption A_1 , for any $\Theta \in \mathcal{S}_c$, $A^*(\Theta)(x)$ is continuous on x .*

Proof. Let $\{(x_i, y_i)\}_{i=1}^\infty \subset \mathcal{X} \times \mathcal{Y}$ converges to (x_0, y_0) . Since \mathcal{Y} is finite, we may assume $y_i = y_0$ for all i . Then for any Θ , we will prove $\lim_{i \rightarrow \infty} A^*(\Theta)(x_i) = A^*(\Theta)(x_0)$. Suppose the contrary. Then $\forall \eta > 0$, $\|A^*(\Theta)(x_i) - A^*(\Theta)(x_0)\| > \eta$ holds for infinitely many i . In the rest of the proof, we assume $\eta < \epsilon/2$.

Let $\zeta = \mathbf{L}(\mathcal{C}_\Theta(x_0 + A^*(\Theta)(x_0)), y_0) - \max_{\alpha \in \mathbb{B}_\epsilon, \|\alpha - A^*(\Theta)(x_0)\| > \eta} \mathbf{L}(\mathcal{C}_\Theta(x_0 + \alpha), y_0)$. Since $\eta < \epsilon/2$, $\{\alpha \in \mathbb{B}_\epsilon : \|\alpha - A^*(\Theta)(x_0)\| > \eta\} \neq \emptyset$. From the uniqueness of $A^*(\Theta)$, we have $\zeta > 0$. By the convergence of $\{x_i\}_{i=1}^\infty$, $\exists N$, such that when $i > N$, $\|x_0 - x_i\| < \frac{\epsilon}{3\Psi\Lambda}$. There exists a $k > N$ such that $\|A^*(\Theta)(x_k) - A^*(\Theta)(x_0)\| > \eta$. Then

$$\begin{aligned} \mathbf{L}(\mathcal{C}_\Theta(x_k + A^*(\Theta)(x_0)), y_0) &\geq \mathbf{L}(\mathcal{C}_\Theta(x_0 + A^*(\Theta)(x_0)), y_0) - \Psi\Lambda\|x_0 - x_k\| \\ &\geq \mathbf{L}(\mathcal{C}_\Theta(x_0 + A^*(\Theta)(x_k)), y_0) - \Psi\Lambda\|x_0 - x_k\| + \zeta \\ &\geq \mathbf{L}(\mathcal{C}_\Theta(x_k + A^*(\Theta)(x_k)), y_0) - 2\Psi\Lambda\|x_0 - x_k\| + \zeta \\ &> \mathbf{L}(\mathcal{C}_\Theta(x_k + A^*(\Theta)(x_k)), y_0) + \zeta/3 \\ &> \mathbf{L}(\mathcal{C}_\Theta(x_k + A^*(\Theta)(x_k)), y_0) \end{aligned}$$

which contradicts to the definition of $A^*(\Theta)(x_k)$. Hence $A^*(\Theta)(x)$ is continuous on x . \square

Lemma 3.11. *Under Assumption A_1 , $\psi(\Theta) = \varphi_s(\Theta, A^*(\Theta)) : \mathcal{S}_c \rightarrow \mathbb{R}$ is continuous on Θ .*

Proof. We will prove that for any $\zeta > 0$, $\exists \delta > 0$, such that if $\|\Theta_1 - \Theta_2\|_2 \leq \delta$ then $|\varphi_s(\Theta_1, A^*(\Theta_1)) - \varphi_s(\Theta_2, A^*(\Theta_2))| \leq \zeta$. Let $\delta = \frac{\zeta}{\Psi\Delta}$. Then for any x ,

$$\begin{aligned} \mathbf{L}(\mathcal{C}_{\Theta_1}(x + A^*(\Theta_1)(x))) &\leq \mathbf{L}(\mathcal{C}_{\Theta_2}(x + A^*(\Theta_1)(x))) + \Psi\Delta\delta \\ &\leq \mathbf{L}(\mathcal{C}_{\Theta_2}(x + A^*(\Theta_2)(x))) + \Psi\Delta\delta \\ &= \mathbf{L}(\mathcal{C}_{\Theta_2}(x + A^*(\Theta_2)(x))) + \zeta. \end{aligned}$$

By exchanging Θ_1 and Θ_2 , we have $|\mathbf{L}(\mathcal{C}_{\Theta_1}(x + A^*(\Theta_1)(x))) - \mathbf{L}(\mathcal{C}_{\Theta_2}(x + A^*(\Theta_2)(x)))| \leq \zeta$. Then $|\varphi_s(\Theta_1, A(\Theta_1)) - \varphi_s(\Theta_2, A(\Theta_2))| \leq \zeta$. Thus $\varphi_s(\Theta, A^*(\Theta))$ is continuous on Θ . \square

Lemma 3.12. *Under Assumption A_1 , $A^*(\Theta) : \mathcal{S}_c \rightarrow \mathcal{S}_a$ is continuous.*

Proof. It suffices to prove that when $\{\Theta_n\}_{n=1}^\infty$ converges to Θ_0 , $\lim_{n \rightarrow \infty} A^*(\Theta_n) = A^*(\Theta_0)$. Suppose the contrary. Then there exist $x \in \mathcal{X}$ and $\eta > 0$ such that $\|A^*(\Theta_n)(x) - A^*(\Theta_0)(x)\| > \eta$ holds for infinitely n . We assume $\eta < \varepsilon/2$.

Let $\zeta = \mathbf{L}(\mathcal{C}_{\Theta_0}(x + A^*(\Theta_0)(x)), y) - \max_{\alpha \in \mathbb{B}_\varepsilon, \|\alpha - A^*(\Theta_0)(x)\| > \eta} \mathbf{L}(\mathcal{C}_{\Theta_0}(x + \alpha), y)$. It is clear that $\zeta > 0$. There exists an $N \in \mathbb{N}_+$, such that for any $n > N$, we have $\|\Theta_n - \Theta_0\|_2 < \frac{\varepsilon}{2\Psi\Delta}$ and $|\mathbf{L}(\mathcal{C}_{\Theta_0}(x + A^*(\Theta_0)(x)), y) - \mathbf{L}(\mathcal{C}_{\Theta_n}(x + A^*(\Theta_n)(x)), y)| < \frac{\varepsilon}{2}$ by Lemma 3.11. There exists a $j > N$, $\|A^*(\Theta_j)(x) - A^*(\Theta_0)(x)\| > \eta$. Then

$$\begin{aligned} \mathbf{L}(\mathcal{C}_{\Theta_0}(x + A^*(\Theta_0)(x)), y) &\geq \mathbf{L}(\mathcal{C}_{\Theta_0}(x + A^*(\Theta_j)(x)), y) + \zeta \\ &\geq \mathbf{L}(\mathcal{C}_{\Theta_j}(x + A^*(\Theta_j)(x)), y) + \zeta - \Psi\Delta\|\Theta_j - \Theta_0\|_2 \\ &> \mathbf{L}(\mathcal{C}_{\Theta_j}(x + A^*(\Theta_j)(x)), y) + \frac{\zeta}{2} \end{aligned}$$

which contradicts to $|\mathbf{L}(\mathcal{C}_{\Theta_0}(x + A^*(\Theta_0)(x)), y) - \mathbf{L}(\mathcal{C}_{\Theta_j}(x + A^*(\Theta_j)(x)), y)| < \frac{\varepsilon}{2}$. Thus for any $x, \eta > 0$, there exists an N such that for $n > N$, $\|A^*(\Theta_n)(x) - A^*(\Theta_0)(x)\| \leq \eta$ holds, that is, $\lim_{n \rightarrow \infty} \|A^*(\Theta_n) - A^*(\Theta_0)\|_\infty = 0$, which means $A^*(\Theta)$ is continuous on Θ . \square

Proposition 3.13. *Under Assumption A_1 , Γ_s defined in (12) is a compact set in $\mathcal{S}_c \times \mathcal{S}_a$.*

Proof. Given a sequence $\{(\Theta_n, A^*(\Theta_n))\}_{n=1}^\infty$ in Γ_s , since \mathcal{S}_a is compact, there exists a subsequence $\{\Theta_{i_n}\}_{n=1}^\infty$ converges to Θ_0 , that is, $\lim_{n \rightarrow \infty} \Theta_{i_n} = \Theta_0$. By Lemma 3.12, $A^*(\Theta)$ is continuous on Θ , then $\lim_{n \rightarrow \infty} A^*(\Theta_{i_n}) = A^*(\Theta_0)$. Hence $\{(\Theta_{i_n}, A^*(\Theta_{i_n}))\}_{n=1}^\infty$ is subsequence converging to $(\Theta_0, A^*(\Theta_0))$. By Lemma 3.4, Γ_s is closed, thus $(\Theta_0, A^*(\Theta_0)) \in \Gamma_s$ and Γ_s is compact. \square

4 A Stackelberg game to achieve maximal adversarial accuracy

The *adversarial accuracy* of a DNN \mathcal{C} with respect to an attack radius ε is

$$\text{AA}_{\mathcal{D}}(\mathcal{C}, \varepsilon) = \mathbb{P}_{(x,y) \sim \mathcal{D}} (\forall \bar{x} \in \mathbb{B}(x, \varepsilon) (\widehat{\mathcal{C}}(\bar{x}) = y)) \quad (19)$$

which is the most widely used robustness measurement for DNNs. Comparing to the robustness measurement $\text{AR}_{\mathcal{D}}$ in (6), $\text{AA}_{\mathcal{D}}(\mathcal{C}, \varepsilon)$ does not depend on the loss function. In this section, we will show that adversarial training with the Carlini-Wagner loss function will give a DNN with the optimal adversarial accuracy.

We first introduce a new game. Denote \mathcal{G}_a to be the two person zero-sum minmax Stackelberg game with the Classifier as the leader, the Adversary as the follower, and

$$\varphi_a(\Theta, A) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{L}_A(\mathcal{C}_\Theta(x + A(x)), y). \quad (20)$$

as the payoff function, where the loss function is defined as

$$\mathbf{L}_A(\mathcal{C}(x), y) = \begin{cases} 0 & \mathbf{L}_{\text{cw}}(\mathcal{C}(x), y) \geq 0 \\ -1 & \mathbf{L}_{\text{cw}}(\mathcal{C}(x), y) < 0 \end{cases} \quad (21)$$

and \mathbf{L}_{cw} is the Carlini-Wagner loss function defined in (8).

For game \mathcal{G}_a , γ and Γ defined in (10) and (12) are

$$\begin{aligned} \gamma_a(\Theta) &= \{\operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_a(\Theta, A)\} \text{ for } \Theta \in \mathcal{S}_c \\ \Gamma_a &= \{(\Theta, A) : \Theta \in \mathcal{S}_c, A \in \gamma_a(\Theta)\}. \end{aligned} \quad (22)$$

Lemma 4.1. *Let $A_a \in \gamma_a(\Theta)$. Then $\varphi_a(\Theta, A_a) = -\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta}, \varepsilon)$.*

Proof. Note that $\mathbf{L}_A(\mathcal{C}, x, y) = -1$ if and only if $\widehat{\mathcal{C}}(x) = y$ and $\mathbf{L}_A(\mathcal{C}, x, y) = 0$ if and only if $\widehat{\mathcal{C}}(x) \neq y$ or there exists a $k \neq y$ such that $\mathcal{C}_k(x) = \mathcal{C}_y(x)$. From $A_a \in \gamma_a(\Theta)$, $\mathbf{L}_A(\mathcal{C}_{\Theta}(x + A_a(x)), y) = -1$ if and only if \mathcal{C}_{Θ} is robust over $\mathbb{B}(x, \varepsilon)$, or equivalently, $\widehat{\mathcal{C}}_{\Theta}(\bar{x}) = y$ for any $\bar{x} \in \mathbb{B}(x, \varepsilon)$. Then $\varphi_s(\Theta, A_a) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbf{L}_A(\mathcal{C}_{\Theta}(x + A_a(x)), y) = -\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta}, \varepsilon)$. \square

Lemma 4.2. *$\gamma_a(\Theta) \neq \emptyset$ and $A^* \in \gamma_a(\Theta)$ if and only if $A^*(x) \in \{\operatorname{argmax}_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}_A(\mathcal{C}_{\Theta}(\bar{x}), y)\}$ for all $(x, y) \sim \mathcal{D}$.*

Proof. We first show that $\gamma(\Theta, x) = \{\operatorname{argmax}_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}_A(\mathcal{C}_{\Theta}(\bar{x}), y)\} \neq \emptyset$ and lemma follows from this. Let $x^* \in \{\operatorname{argmax}_{\bar{x} \in \mathbb{B}(x, \varepsilon)} \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta}(\bar{x}), y)\}$. Then x^* exists, since \mathbf{L}_{cw} is continuous and $\mathbb{B}(x, \varepsilon)$ is compact. If $\mathbf{L}_{\text{cw}}(\mathcal{C}, x^*, y) \geq 0$, then $\mathbf{L}_A(\mathcal{C}, x^*, y) = 0$ and $x^* \in \gamma(\Theta, x)$. If $\mathbf{L}_{\text{cw}}(\mathcal{C}, x^*, y) < 0$, then $\mathbf{L}_A(\mathcal{C}, x^*, y) = -1$ for all $x^* \in \mathbb{B}(x, \varepsilon)$ and $\mathbb{B}(x, \varepsilon) = \gamma(\Theta, x)$. In either case, $\gamma(\Theta, x) \neq \emptyset$. \square

Lemma 4.3. *Let $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*)$ be a Stackelberg equilibrium of game \mathcal{G}_s when the loss function is \mathbf{L}_{cw} defined in (8). Then $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*)$ is a Stackelberg equilibrium of game \mathcal{G}_a .*

Proof. By Lemma 4.1, $\gamma_a(\Theta) \neq \emptyset$. So it suffices to show that $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*) \in \operatorname{argmin}_{(\Theta, A(\Theta)) \in \Gamma_a} \varphi_a(\Theta, A(\Theta))$. Denote $\gamma_{\text{cw}}, \Gamma_{\text{cw}}, \varphi_{\text{cw}}$ to be $\gamma_s, \Gamma_s, \varphi_s$, when the loss function $\mathbf{L}_C W$ is used.

We first prove $\gamma_{\text{cw}}(\Theta) \subset \gamma_a(\Theta)$. Hence $\Gamma_{\text{cw}} \subset \Gamma_a$. By Lemma 3.3, $A^* \in \gamma_{\text{cw}}(\Theta) = \{\operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_{\text{cw}}(\Theta, A)\}$ if and only if $A_{\text{cw}}^*(x) \in \gamma_{\text{cw}}(\Theta, x, y) = \{\operatorname{argmax}_{A(x) \in \mathbb{B}_{\varepsilon}} \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta}(x + A(x)), y)\}$. By Lemma 4.2, $A^* \in \gamma_a(\Theta)$ if and only if $A_a^*(x) \in \gamma_a(\Theta, x, y) = \{\operatorname{argmax}_{A(x) \in \mathbb{B}_{\varepsilon}} \mathbf{L}_a(\mathcal{C}_{\Theta}(x + A(x)), y)\}$. Since $\mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta}(x + A_1), y) \leq \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta}(x + A_2), y)$ implies $\mathbf{L}_a(\mathcal{C}_{\Theta}(x + A_1), y) \leq \mathbf{L}_a(\mathcal{C}_{\Theta}(x + A_2), y)$, we have $\gamma_{\text{cw}}(\Theta, x, y) \subset \gamma_a(\Theta, x, y)$. Then $A^* \in \gamma_{\text{cw}}(\Theta)$ implies $A^* \in \gamma_a(\Theta)$.

We next prove

$$\{\varphi_a(\Theta, A), \forall (\Theta, A) \in \Gamma_a\} = \{\varphi_a(\Theta, A), \forall (\Theta, A) \in \Gamma_{\text{cw}}\}. \quad (23)$$

Since $\Gamma_{\text{cw}} \subset \Gamma_a$, it suffices to show $\{\varphi_a(\Theta, A), \forall (\Theta, A) \in \Gamma_a\} \subset \{\varphi_a(\Theta, A), \forall (\Theta, A) \in \Gamma_{\text{cw}}\}$. For $(\Theta_a, A_a) \in \Gamma_a$, let $A_{\text{cw}}(x) \in \operatorname{argmax}_{A \in \mathbb{B}_{\varepsilon}} \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A), y)$. Then $(\Theta_a, A_{\text{cw}}) \in \Gamma_{\text{cw}}$. We will show that $\varphi_a(\Theta_a, A_a) = \varphi_a(\Theta_a, A_{\text{cw}})$. By Lemma 4.2, $A_a^* \in \gamma_a(\Theta_a)$ if and only if

$$A_a^*(x) \in \gamma_a(\Theta_a, x, y) = \{\operatorname{argmax}_{A(x) \in \mathbb{B}_{\varepsilon}} \mathbf{L}_a(\mathcal{C}_{\Theta_a}(x + A(x)), y)\}$$

for all $(x, y) \sim \mathcal{D}$. If $\mathbf{L}_a(\mathcal{C}_{\Theta_a}(x + A_a^*(x)), y) = -1$, then $\mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A), y) < 0$ for all $A \in \mathbb{B}_{\varepsilon}$. In this case, $\max_{A \in \mathbb{B}_{\varepsilon}} \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A), y) = \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A_{\text{cw}}^*(x)), y) < 0$ and hence $\mathbf{L}_a(\mathcal{C}_{\Theta_a}(x +$

$A_{\text{cw}}^*(x), y) = -1$. If $\mathbf{L}_a(\mathcal{C}_{\Theta_a}(x + A_a^*(x)), y) = 0$, then $\mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A_a^*(x)), y) \geq 0$. In this case, $\max_{A \in \mathbb{B}_\varepsilon} \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A), y) = \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a}(x + A_{\text{cw}}^*(x)), y) \geq 0$ and hence $\mathbf{L}_a(\mathcal{C}_{\Theta_a}(x + A_a^*(x)), y) = \mathbf{L}_a(\mathcal{C}_{\Theta_a}(x + A_{\text{cw}}^*(x)), y) = 0$. Then we have $\varphi_a(\Theta_a, A_a) = \varphi_a(\Theta_a, A_{\text{cw}})$.

By (23), $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*) \in \operatorname{argmin}_{(\Theta, A(\Theta)) \in \Gamma_{\text{cw}}} \varphi_a(\Theta, A(\Theta)) = \operatorname{argmin}_{(\Theta, A(\Theta)) \in \Gamma_a} \varphi_a(\Theta, A(\Theta))$. The lemma is proved. \square

Theorem 4.4. *Let $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*)$ be a Stackelberg equilibrium of game \mathcal{G}_s when the loss function is \mathbf{L}_{cw} in (8). Then $\mathcal{C}_{\Theta_{\text{cw}}^*}$ has the largest adversarial accuracy for all DNNs in \mathcal{H} defined in (1), that is $\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_{\text{cw}}^*}, \varepsilon) \geq \text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta}, \varepsilon)$ for any $\mathcal{C}_{\Theta} \in \mathcal{H}$.*

Proof. By Lemma 4.3, $(\Theta_{\text{cw}}^*, A_{\text{cw}}^*)$ be a Stackelberg equilibrium of game \mathcal{G}_a . By Lemma 4.1, $\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_{\text{cw}}^*}, \varepsilon) = -\varphi_a(\Theta_{\text{cw}}^*, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_a(\Theta_{\text{cw}}^*, A)) \geq -\varphi_a(\Theta, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_a(\Theta, A)) = \text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta}, \varepsilon)$. The theorem is proved. \square

Remark 4.5. *By Theorems 3.5 and 4.4, adversarial training using the loss function \mathbf{L}_{cw} gives a DNN which has the largest adversarial accuracy for all DNNs in the hypothesis space \mathcal{H} , which answers Question **Q1** positively for the hypothesis space \mathcal{H} .*

5 Trade-off between robustness and accuracy

In this section, we give trade-off results between the robustness and the accuracy in adversarial deep learning from game theoretical viewpoint.

5.1 Improve accuracy under maximal adversarial accuracy

By Remarks 3.6 and 4.5, adversarial training computes the DNNs with the best robustness measurement. A nature question is whether we can increase the accuracy of the DNN and still keep the maximal adversarial accuracy. That is, consider the bi-level optimization problem.

$$\begin{aligned} \Theta_o^* &= \operatorname{argmin}_{\Theta^*} \varphi_0(\Theta^*) \\ &\text{subject to} \\ \Theta_s^* &= \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \max_{A \in \mathcal{S}_a} \varphi_s(\Theta, A) \end{aligned} \quad (24)$$

where φ_0 and φ_s are defined in (2) and (15), respectively.

From Remark 3.7, if using the loss function \mathbf{L}_{cw} , $\gamma_s(\Theta)$ contains a unique solution and Θ_s^* is unique in the generic case. In this case, we cannot increase the accuracy of the DNN when keeping the maximal robust measure $\text{AR}_{\mathcal{D}}$.

A more interesting case is to consider game \mathcal{G}_a defined in section 4, which uses the loss function \mathbf{L}_A defined in (21).

We first introduce an assumption. We train \mathcal{C}_{Θ} with stochastic gradient descent starting from a randomly choosing initial point, and most probably will terminate at a random point in the neighborhood of a minimal point or a saddle point of the loss function. Therefore, the following assumption is valid for almost all trained DNNs [46].

Assumption A_2 . The parameters of a trained \mathcal{C}_{Θ} are *random values*.

We now estimate the possible values of Θ_s^* in (24). Suppose a finite data set $T = \{(x_i, y_i)\}_{i=1}^N$ is chosen iid from the distribution \mathcal{D} , which are used to train the network. Then it can be shown

that the game \mathcal{G}_a with payoff function (20) and trained with T has a Stackelberg equilibrium (Θ_a^*, A_a^*) (See section 6 for more details). With these notations, we have

Proposition 5.1. *Under Assumption A_2 , there exists a $\nu \in \mathbb{R}_+$ such that for all $\Theta_a^\circ \in \mathbb{R}^K$ satisfying $\|\Theta_a^\circ - \Theta_a^*\| < \nu$, game \mathcal{G}_a has a Stackelberg equilibrium $(\Theta_a^\circ, A_a^\circ)$.*

Proof. Denote $\phi(\Theta, x) = \mathbf{L}_{\text{cw}}(\mathcal{C}_\Theta(x), y)$ for a fixed y . Let $x^* \in \{\operatorname{argmax}_{\bar{x} \in \mathbb{B}(x_i, \varepsilon)} \phi(\Theta_a^*, \bar{x})\}$. If $\phi(\Theta_a^*, x^*) < 0$, then $\phi(\Theta_a^*, \bar{x}) < 0$ for all $\bar{x} \in \mathbb{B}(x_i, \varepsilon)$. Since $\mathbb{B}(x_i, \varepsilon)$ is compact and $\phi(\Theta, x)$ is continuous, there exists a $\nu_i \in \mathbb{R}_+$ such that $\phi(\Theta_a^* + \Delta, \bar{x}) < 0$ for all $\bar{x} \in \mathbb{B}(x_i, \varepsilon)$ and all $\Delta \in \mathbb{R}^K$ satisfying $\|\Delta\| < \nu_i$. Without loss of generality, we can assume $\Theta_a^* + \Delta \in \mathcal{S}_c$. It is easy to construct the best response of the Adversary in this case for $\Theta_a^\circ = \Theta_a^* + \Delta$: $A_a^\circ(x_i)$ can be any point in $\mathbb{B}(x_i, \varepsilon)$. If $\phi(\Theta_a^*, x^*) > 0$, then $S_i(\Theta_a^*) = \{x \in \mathbb{B}(x_i, \varepsilon) : \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a^*}(x), y) \leq 0\}$ is a compact set of dimension m , since $\mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a^*}(x), y)$ is piecewise linear in x . If ν_i is small enough, then $S_i(\Theta_a^* + \Delta)$ is also a compact set of dimension m for all $\Delta \in \mathbb{R}^K$ and $\|\Delta\| < \nu_i$. In this case, $A_a^\circ(x_i)$ can be any point in $\mathcal{S}_i(\Theta_a^* + \Delta)$.

By Assumption A_2 , the trained parameters of \mathcal{C} are random values. $\phi(\Theta_a^*, x^*) = \mathbf{L}_{\text{cw}}(\mathcal{C}_{\Theta_a^*}(x^*), y) = 0$ implies that $\mathcal{C}_{\Theta_a^*, i}(x^*) = \mathcal{C}_{\Theta_a^*, j}(x^*)$ for $i \neq j$, which gives an algebraic relation among the parameters of \mathcal{C}_Θ . This imposes an extra algebraic relation among the random parameters and thus will not happen under Assumption A_2 . So we have $\phi(\Theta, x^*) \neq 0$ under Assumption A_2 .

Let $\nu = \min_{i=1}^N \nu_i > 0$. Then for $\|\Theta_a^\circ - \Theta_a^*\| < \nu$, there exists an $A_a^\circ \in \mathcal{S}_c$ such that $\varphi_a(\Theta_a^\circ, A_a^\circ) = \varphi_a(\Theta_a^*, A_a^*)$, where φ_a is defined in (20). Since (Θ_a^*, A_a^*) is a Stackelberg equilibrium for game \mathcal{G}_a , so is $(\Theta_a^\circ, A_a^\circ)$. The proposition is proved. \square

By Proposition 5.1, Θ_s^* in (24) takes values in a K -dimensional set. As a consequence, there exist rooms for increase the accuracy under the maximal adversarial accuracy.

Example 5.2. *We use numerical experiments to show that it is possible to further increase the accuracy under the maximal adversarial accuracy. Two small CNNs with respectively 3 and 4 hidden layers are used, which have structures $(8 * 3 * 3)$, $(16 * 3 * 3)$, $(32 * 3 * 3)$ and $(32 * 3 * 3)$, $(64 * 3 * 3)$, $(128 * 3 * 3)$, $(128 * 3 * 3)$, respectively. We use loss function \mathbf{L}_{cw} to achieve maximal adversarial accuracy and the results are given in the columns 1-0 and 2-0 in Table 1. We then retrain the CNNs using the normal loss function in (2) to increase the accuracy. In order to keep the maximal adversarial accuracy fixed, the change of the parameters are limited to $i\%$ for $i = 1, 2, 3$ and the results are given in columns 1- i and 2- i , respectively. We can see that the adversarial accuracies are barely changed (up to 0.06% and 0.02% for networks 1 and 2), but the accuracies are increased evidently (up to 1.11% and 2.252% for networks 1 and 2).*

Table 1: Increase the accuracy (AC) under the condition of maximal adversarial accuracy (AA) for CIFRA-10. The attack radius is 8/255 and 50000 samples are used.

	Network 1				Network 2			
	1-0	1-1	1-2	1-3	2-0	2-1	2-2	2-3
AC (%)	45.718	46.762	46.814	46.828	72.156	75.284	75.344	75.408
AA (%)	29.018	28.996	28.98	28.958	40.08	40.076	40.036	40.06

5.2 An effective trade-off method

The bi-level optimization problem (24) is in general difficult to solve, especially when keeping the maximal adversarial accuracy as mentioned in the proof of Proposition 5.1. A natural way to

train a robust and more accurate DNN is to do adversarial training with the following objective function

$$\varphi_t(\Theta, A) = \varphi_s(\Theta, A) + \lambda\varphi_0(\Theta) \quad (25)$$

where $\lambda > 0$ is a small hyperparameter, φ_0 and φ_s are defined in (2) and (15), respectively. Problem (25) is also often used as an approximate way to solve (24). We will prove a trade-off result in this setting.

Similar to Theorem 3.5, adversarial training with loss function (25) can be considered as a Stackelberg game \mathcal{G}_t with φ_t as the payoff function. Then we have the following trade-off result.

Proposition 5.3. *Let (Θ_s^*, A_s^*) and (Θ_t^*, A_t^*) be the Stackelberg equilibria of the zero-sum sequential games with φ_s and φ_t as the payoff functions, respectively. Then*

$$\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_s^*}, \varepsilon) \geq \text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_t^*}, \varepsilon), \varphi_s(\Theta_s^*, A_s^*) \leq \varphi_s(\Theta_t^*, A_t^*) \text{ and } \varphi_0(\Theta_s^*) \geq \varphi_0(\Theta_t^*)$$

that is, the network $\mathcal{C}_{\Theta_s^*}$ is more robust but less accurate than $\mathcal{C}_{\Theta_t^*}$ measured by φ_0 .

Proof. $\text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_s^*}, \varepsilon) \geq \text{AA}_{\mathcal{D}}(\mathcal{C}_{\Theta_t^*}, \varepsilon)$ is a consequence of Theorem 4.4. Since (Θ_t^*, A_t^*) is a Stackelberg equilibrium of game \mathcal{G}_t , we have

$$\Theta_t^* \in \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \varphi_t(\Theta, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_t(\Theta, A)) \quad (26)$$

$$\begin{aligned} A_t^* &\in \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_t(\Theta_t^*, A) \\ &= \operatorname{argmax}_{A \in \mathcal{S}_a} (\varphi_s(\Theta_t^*, A) + \lambda\varphi_0(\Theta_t^*)) \\ &= \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta_t^*, A) \end{aligned} \quad (27)$$

where the last equality is due to the fact that $\varphi_0(\Theta_t^*)$ is free of A . Then, from (18),

$$\begin{aligned} \varphi_s(\Theta_s^*, A_s^*) &= \varphi_s(\Theta_s^*, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta_s^*, A)) \\ &\leq \varphi_s(\Theta_t^*, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_s(\Theta_t^*, A)) \\ &\leq \max_{A \in \mathcal{S}_a} \varphi_s(\Theta_t^*, A) = \varphi_s(\Theta_t^*, A_t^*). \end{aligned} \quad (28)$$

The last equality comes from (27). From (26),

$$\varphi_t(\Theta_t^*, A_t^*) \leq \varphi_t(\Theta_s^*, \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_t(\Theta_s^*, A)) \leq \max_{A \in \mathcal{S}_a} \varphi_t(\Theta_s^*, A) = \varphi_t(\Theta_s^*, A_s^*). \quad (29)$$

Adding inequalities (28) and (29), we obtain $\varphi_0(\Theta_s^*) \geq \varphi_0(\Theta_t^*)$. The proposition is proved. \square

Note that this trade-off result is quite different from the trade-off theorem in [42] in that, our result is for any data set, while the result in [42] is for a specifically designed data set.

6 Comparing three types of games for adversarial deep learning

In this section, we compare three types of games for adversarial deep learning when the data $T = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{I}^n \times \mathcal{Y}$ are a finite number of samples chosen iid from the distribution \mathcal{D} .

In this case, the strategy space for the Classifier is still \mathcal{S}_c in (13). The strategy space for the Adversary becomes much simpler:

$$\mathcal{S}_a = \prod_{i=1}^N \{(\bar{x}_i, y_i) : \|\bar{x}_i - x_i\| \leq \varepsilon\} \subset (\mathbb{I}_\varepsilon^n \times \mathcal{Y})^N \quad (30)$$

where $\mathbb{I}_\varepsilon = [-\varepsilon, 1 + \varepsilon]$. For $\Theta \in \mathcal{S}_c$ and $A = ((\bar{x}_i, y_i))_{i=1}^N \in \mathcal{S}_a$, the *empirical adversarial loss* is

$$\varphi_T(\Theta, A) = \frac{1}{N} \sum_{i=1}^N \mathbf{L}(\mathcal{C}_\Theta(\bar{x}_i), y_i). \quad (31)$$

We consider three games.

The adversarial training game \mathcal{G}_1 , which is the zero-sum minmax sequential game with the Classifier as the leader, the Adversary as the follower, and $\varphi_T(\Theta, A)$ as the payoff function, that is, to solve the following minmax problem

$$\Theta_1^* = \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \max_{A \in \mathcal{S}_a} \varphi_T(\Theta, A) \quad (32)$$

which is clearly equivalent to the adversarial training. By Theorem 3.1, game \mathcal{G}_1 has a Stackelberg equilibrium (Θ_1^*, A_1^*) , since \mathcal{S}_c and \mathcal{S}_a are compact and $\varphi_T(\Theta, A)$ is continuous. Similar to section 4, it can be shown that this game gives a DNN with the largest adversarial accuracy for the data set T , when the loss function is \mathbf{L}_{cw} .

The universal adversary game \mathcal{G}_2 , which is the zero-sum maxmin sequential game with the Adversary as the leader and the Classifier as the follower, that is, to solve the following maxmin problem

$$\mathcal{A}_2^* = \operatorname{argmax}_{A \in \mathcal{S}_a} \min_{\Theta \in \mathcal{S}_c} \varphi_T(\Theta, A) \quad (33)$$

By Theorem 3.1, game \mathcal{G}_2 has a Stackelberg equilibrium (Θ_2^*, A_2^*) . The solution (Θ_2^*, A_2^*) of this game is to compute the optimal *universal adversarial attack* for the given hypothesis space \mathcal{H} in (1), that is, $A_2^*(x)$ is the best adversary for any $(x, y) \sim \mathcal{D}$ and for all DNNs in \mathcal{H} . It is clear that \mathcal{A}_2^* is the optimal attack to the so-called *nobox model* proposed in [5], that is, nox model has an optimal solution for DNNs with a given structure. This gives a positive answer to question **Q₂** for the hypothesis space \mathcal{H} in (1).

The simultaneous adversary game \mathcal{G}_3 . We can also formulate the adversarial deep learning as a simultaneous game \mathcal{G}_3 . In this game, the two players and their strategy spaces are the same as that of game \mathcal{G}_1 . The difference is the way to play the game. In game \mathcal{G}_3 , the Classifier picks its action without knowing the action of the Adversary, and the Adversary chooses the attacking adversarial samples without knowing the action of the Classifier. But, both players know the payoff function. A point $(\Theta_3^*, A_3^*) \in \mathcal{S}_c \times \mathcal{S}_a$ is called a *pure strategy Nash equilibrium* of game \mathcal{G}_3 if

$$\Theta_3^* = \operatorname{argmin}_{\Theta \in \mathcal{S}_c} \varphi_T(\Theta, A_3^*) \text{ and } A_3^* = \operatorname{argmax}_{A \in \mathcal{S}_a} \varphi_T(\Theta_3^*, A). \quad (34)$$

In general, pure strategy Nash equilibria do not necessarily exist, and mixed strategy Nash equilibria are usually considered. *Mixed strategies* for the Classifier and the Adversary are two probability distributions

$$\tilde{\Theta} : \mathcal{S}_c \rightarrow \mathbb{I} \text{ and } \tilde{A} : \mathcal{S}_a \rightarrow \mathbb{I}$$

for Θ and A , respectively. For a mixed strategy $(\tilde{\Theta}, \tilde{A})$, the payoff function is

$$\varphi_T(\tilde{\Theta}, \tilde{A}) = \mathbb{E}_{\Theta \sim \tilde{\Theta}} \mathbb{E}_{A \sim \tilde{A}} \varphi_T(\Theta, A). \quad (35)$$

Denote $\tilde{\mathcal{S}}_c$ and $\tilde{\mathcal{S}}_a$ to be the sets of the mixed strategies for the Classifier and the Adversary, respectively. Then $(\tilde{\Theta}_3^*, \tilde{A}_3^*) \in \tilde{\mathcal{S}}_c \times \tilde{\mathcal{S}}_a$ is called a *mixed strategy Nash equilibrium* of game \mathcal{G}_3 if

$$\tilde{\Theta}_3^* = \operatorname{argmin}_{\tilde{\Theta} \in \tilde{\mathcal{S}}_c} \varphi_T(\tilde{\Theta}, \tilde{A}_3^*) \text{ and } \tilde{A}_3^* = \operatorname{argmax}_{\tilde{A} \in \tilde{\mathcal{S}}_a} \varphi_T(\tilde{\Theta}_3^*, \tilde{A}). \quad (36)$$

Since the strategy spaces of the two players are compact and the objective function is continuous, by Glicksberg's theorem [15], game \mathcal{G}_3 has a mixed strategy Nash equilibrium $(\tilde{\Theta}_3^*, \tilde{A}_3^*)$, and the minmax theorem holds for this equilibrium.

Remark 6.1. *By Proposition 3.13, we can show that, under Assumption A₁, game G_3 has a mixed strategy when the data set satisfies a general distribution \mathcal{D} .*

Proposition 6.2. *Let (Θ_i^*, A_i^*) be Nash equilibria of games \mathcal{G}_i for $i = 1, 2, 3$, respectively (mixed strategy for \mathcal{G}_3). Then*

$$\varphi_T(\Theta_1^*, A_1^*) \geq \varphi_T(\Theta_3^*, A_3^*) \geq \varphi_T(\Theta_2^*, A_2^*).$$

Proof. The mixed strategy (Θ_3^*, A_3^*) can be written as two distributions $\Delta_c : \mathcal{S}_c \rightarrow \mathbb{I}$ and $\Delta_a : \mathcal{S}_a \rightarrow \mathbb{I}$, respectively. To prove the first inequality, we have

$$\varphi_T(\Theta_1^*, A_1^*) = \mathbb{E}_{A \sim \Delta_a} \varphi_T(\Theta_1^*, A_1^*) \stackrel{(18)}{\geq} \mathbb{E}_{A \sim \Delta_a} \varphi_T(\Theta_1^*, A) = \varphi_T(\Theta_1^*, A_3^*) \stackrel{(36)}{\geq} \varphi_T(\Theta_3^*, A_3^*).$$

For the second inequality, we have

$$\varphi_T(\Theta_2^*, A_2^*) = \mathbb{E}_{\Theta \sim \Delta_c} \varphi_T(\Theta_2^*, A_2^*) \stackrel{(33)}{\leq} \mathbb{E}_{\Theta \sim \Delta_c} \varphi_T(\Theta, A_2^*) = \varphi_T(\Theta_3^*, A_2^*) \stackrel{(36)}{\leq} \varphi_T(\Theta_3^*, A_3^*).$$

The proposition is proved. □

The following example shows that the inequalities in Proposition 6.2 could be strict.

Example 6.3. *Consider a two-player zero-sum minmax game with payoff matrix*

$$\begin{pmatrix} 0 & -a \\ -1 & 0 \end{pmatrix}$$

where $0 < a < 1$. The strategy space for player one is the rows and its goal is minimize the payoff. Then, the Stackelberg game with player one as the leader is to solve the minmax problem and a Stackelberg equilibrium is (Row 1, Column 1) with payoff 0. The Stackelberg game with player two (column) as the leader is to solve the maxmin problem and a Stackelberg equilibrium is (Row 1, Column 2) with payoff $-a$. By the well known minmax theorem, the corresponding simultaneous game has no Nash equilibrium since $\min \max \neq \max \min$, and a mixed strategy Nash equilibrium exists: the first player plays $(\frac{1}{1+a}, \frac{a}{1+a})$ and the second player plays $(\frac{a}{1+a}, \frac{1}{1+a})$ with payoff $-\frac{a}{1+a}$. We summarize the above discussion as follows:

minmax	payoff = 0
maxmin	payoff = $-a$
Mixed strategy	payoff = $-\frac{a}{1+a} \in (-a, 0)$.

7 Conclusion

In this paper, we give a game theoretical analysis for adversarial deep learning from a more practical viewpoint. In previous work, the adversarial deep learning was formulated as a simultaneous game. In order for the Nash equilibrium to exist, the strategy spaces for the Classifier and the Adversary are assumed to be certain convex probability distributions, which are not

used in real applications. In this paper, the adversarial deep learning is formulated as a sequential game with the Classifier as the leader and the Adversary as the follower. In this case, we show that the game has Stackelberg equilibria when the strategy space for the classifier is DNNs with given width and depth, just like people do in practice.

We prove that Stackelberg equilibria for such a sequential game is the same as the DNNs obtained with adversarial training. Furthermore, if the margin loss introduced by Carlini-Wagner is used as the payoff function, the equilibrium DNN has the largest adversarial accuracy and is thus the provable optimal defence. Based on this approach, we also give theoretical analysis for other important issues such as the tradeoff between robustness and the accuracy, and the generation of optimal universal adversaries.

For future research, it is desirable to develop practical methods to use mixed strategy in deep learning, since it is proved that such strategy has more power than pure strategy when the depth and width of the DNNs are fixed. It is also interesting to analysis the properties of the Nash equilibria for adversarial deep learning, such as whether the equilibria are regular or essential [12, 43]? Finally, we can use game theory to analyze other adversarial problems in deep learning.

References

- [1] A. Athalye, N. Carlini, D. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *Proc. ICML*, PMLR, 274-283, 2018.
- [2] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok. Synthesizing Robust Adversarial Examples. ArXiv: 1707.07397, 2017.
- [3] A. Azulay and Y. Weiss. Why Do Deep Convolutional Networks Generalize so Poorly to Small Image Transformations? *Journal of Machine Learning Research*, 20, 1-25, 2019.
- [4] A. Bastounis, A.C. Hansen, V. Vlačić. The Mathematics of Adversarial Attacks in AI - Why Deep Learning is Unstable Despite the Existence of Stable Neural Networks. arXiv:2109.06098, 2021.
- [5] J. Bose, G. Gidel, H. Berard, A. Cianflone, P. Vincent, S. Lacoste-Julien, W. Hamilton. Adversarial Example Games. *Proc. NeurIPS*, 2020.
- [6] N. Carlini, D. Wagner. Towards Evaluating the Robustness of Neural Networks. *Proc. of IEEE Symposium on Security and Privacy*, IEEE Press, 39-57, 2017.
- [7] N. Carlini, D. Wagner. Adversarial Examples are not Easily Detected: Bypassing Ten Detection Methods. *Proc. 10th ACM Workshop on Artificial Intelligence and Security*, 3-14, 2017.
- [8] A.S. Chivukula, X. Yang, W. Liu, T. Zhu, W. Zhou. Game Theoretical Adversarial Deep Learning With Variational Adversaries. *IEEE Trans. Knowledge and Data Engineering*, 33(11), 3568-3581, 2021.
- [9] J. Cohen, E. Rosenfeld, Z. Kolter. Certified Adversarial Robustness via Randomized Smoothing. *Proc. ICML*, PMLR, 1310-1320, 2019.

- [10] M.J. Colbrook, V. Antun, A.C. Hansen. The Difficulty of Computing Stable And Accurate Neural Networks: On The Barriers of Deep Learning and Smale’s 18th Problem. 119 (12) e2107151119, 2022.
- [11] N. Dalvi, P. Domingos, S. Mausam, D. Verma. Adversarial Classification. *Proc. KDD’04*, 99-108, ACM Press, New York, 2004.
- [12] E. van Damme. Stability and Perfection of Nash Equilibria. Springer, 1987.
- [13] T. Fiez, B. Chasnov, L.J. Ratliff. Implicit Learning Dynamics in Stackelberg Games: Equilibria Characterization, Convergence Analysis, and Empirical Study. *Proc. ICML*, PMLR, 2020.
- [14] D. Fudenberg and J. Tirole. Game Theory. MIT Press, Cambridge, MA, 1991.
- [15] I.L. Glicksberg. A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points. *Proc. AMS*, 3(1), 1952.
- [16] G. Gidel, D. Balduzzi, W.M. Czarnecki, M. Garnelo, Y. Bachrach. Minimax Theorem for Latent Games or: How I Learned to Stop Worrying about Mixed-Nash and Love Neural Nets. arXiv:2002.05820v1, 2020.
- [17] P.W. Koh, P. Liang. Understanding Black-box Predictions via Influence Functions. *Proc. ICML*, PMLR, 1885-1894, 2017.
- [18] Y.P. Hsieh, C. Liu, V. Cevher. Finding Mixed Nash Equilibria of Generative Adversarial Networks. *Proc. ICML*, PMLR, 2019.
- [19] C. Jin, P. Netrapalli, M.I. Jordan. What is Local Optimality in Nonconvex-nonconcave Minimax Optimization? *Proc. ICML*, PMLR, 2020.
- [20] C.A. Kamhoua, C.D. Kiekintveld, F. Fang, Q. Zhu (eds). Game Theory and Machine Learning for Cyber Security. IEEE Press and Wiley, 2021.
- [21] A. Kurakin, I. Goodfellow, S. Bengio. Adversarial Examples in the Physical World. ArXiv: 1607.02533, 2016.
- [22] Y. LeCun, Y. Bengio, G. Hinton. Deep Learning. *Nature*, 521(7553), 436-444, 2015.
- [23] Y. Liu, L. Wei, B. Luo, Q. Xu. Fault Injection Attack on Deep Neural Network. *Proc. of the IEEE/ACM International Conference on Computer-Aided Design*, 131-138, 2017.
- [24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083, 2017.
- [25] L. Meunier, M. Scetbon, R. Pinot, J. Atif, Y. Chevaleyre. Mixed Nash Equilibria in the Adversarial Examples Game. *Proc. ICML*, PMLR 139, 2021.
- [26] G. Montúfar, R. Pascanu, K. Cho, Y. Bengio. On the Number of Linear Regions of Deep Neural Networks. *Proc. NIPS’2014*, 2014.
- [27] S.M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard. Universal Adversarial Perturbations. *Proc. CVPR*, 1765-1773, 2017.

- [28] B. Neyshabur, R. Tomioka, N. Srebro. Norm-based Capacity Control in Neural Networks. *Proc. COLT'15*, 1376-1401, 2015.
- [29] A. Pal and R. Vidal. A Game Theoretic Analysis of Additive Adversarial Attacks and Defenses. *Proc. NeurIPS*, 2020.
- [30] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami. The Limitations of Deep Learning in Adversarial Settings. *IEEE European Symposium on Security and Privacy*, IEEE Press, 2016, 372-387.
- [31] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, A. Swami. Practical Black-box Attacks Against Machine Learning. *Proc. ACM on Asia Conference on Computer and Communications Security*, ACM Press, 506-519, 2017.
- [32] R. Pinot, R. Ettetdgui, G. Rizk, Y. Chevaleyre, J. Atif. Randomization Matters: How to Defend Against Strong Adversarial Attacks. *Proc. ICML*, PMLR, 2020.
- [33] M.S. Pydi and V. Jog. Adversarial Risk via Optimal Transport and Optimal Couplings. *Proc. ICML*, PMLR, 2020.
- [34] F.A. Oliehoek, R. Savani, J. Gallego, E. van der Pol, R. Groß. Beyond Local Nash Equilibria for Adversarial Networks. arXiv:1806.07268, 2018.
- [35] J. Rena, D. Zhanga, Y. Wangb, L. Chen, Z. Zhou, Y. Chen, X. Cheng, X. Wang, M. Zhoua, J. Shi, Q. Zhang. A Unified Game-Theoretic Interpretation of Adversarial Robustness arXiv:2103.07364v2, 2021.
- [36] A. Shafahi, W.R. Huang, M. Najibi, O. Suciuc, C. Studer, T. Dumitras, T. Goldstein. Poison Frogs! Targeted Clean-label Poisoning Attacks on Neural Networks. *Proc. NeurIPS*, 6103-6113, 2018.
- [37] A. Shafahi, W.R. Huang, C. Studer, S. Feizi, T. Goldstein. Are Adversarial Examples Inevitable? arXiv:1809.02104, 2018.
- [38] Y. Shoham and K. Leyton-Brown. Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations. Cambridge University Press, 2008.
- [39] M. Simaan and J.B. Cruz Jr. On the Stackelberg Strategy in Nonzero-sum Games. *Journal of Optimization Theory and Applications*, 11, 533-555, 1973.
- [40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, R. Fergus. Intriguing Properties of Neural Networks. arXiv:1312.6199, 2013.
- [41] Y.L. Tsai, C.Y. Hsu, C.M. Yu, P.Y. Chen. Formalizing Generalization and Robustness of Neural Networks to Weight Perturbations. arXiv:2103.02200, 2021.
- [42] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry. Robustness May Be at Odds With Accuracy. *Proc. ICML*, PMLR, 2019.
- [43] W.T. Wu and J.H. Jiang. Essential Equilibrium Points of n -Person Noncooperative Games. *Scientia Sinica*, 11(10), 1307-1322, 1962.

- [44] H. Xu, Y. Ma, H.C. Liu, D. Deb, H. Liu J.L. Tang, A.K. Jain. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *International Journal of Automation and Computing*, 17(2), 151-178, 2020.
- [45] Y.Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, K. Chaudhuri. A Closer Look at Accuracy vs Robustness. arXiv:2003.02460v3, 2000.
- [46] L. Yu and X.S. Gao. Improve the Robustness and Accuracy of Deep Neural Network with $L_{2,\infty}$ Normalization. Accepted by *Journal of Systems Science and Complexity*, 2022. arXiv:2010.04912.
- [47] L. Yu, Y. Wang, X.S. Gao. Adversarial Parameter Attack on Deep Neural Networks. arXiv:2203.10502, 2022.
- [48] L. Yu and X.S. Gao. Robust and Information-theoretically Safe Bias Classifier against Adversarial Attacks. arXiv:2111.04404, 2021.
- [49] H. Zhang, Y. Yu, J. Jiao, E.P. Xing, L.E. Ghaoui, M.I. Jordan. Theoretically Principled Trade-off Between Robustness and Accuracy. *Proc. ICML*, PMLR, 2019.
- [50] Y. Zhou, M. Kantarcioglu, B. Xi. A survey of Game Theoretic Approach for Adversarial Machine Learning. *WIREs Data Mining Knowl Discov*, 1-9, 2019.