On the Role of Generalization in Transferability of Adversarial Examples

Yilin Wang[1] **Farzan Farnia**[1]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong,
Hong Kong SAR

Abstract

Black-box adversarial attacks designing adversarial examples for unseen deep neural networks (DNNs) have received great attention over the past years. However, the underlying factors driving the transferability of black-box adversarial examples still lack a thorough understanding. In this paper, we aim to demonstrate the role of the generalization behavior of the substitute classifier used for generating adversarial examples in the transferability of the attack scheme to unobserved DNN classifiers. To do this, we apply the max-min adversarial example game framework and show the importance of the generalization properties of the substitute DNN from training to test data in the success of the black-box attack scheme in application to different DNN classifiers. We prove theoretical generalization bounds on the difference between the attack transferability rates on training and test samples. Our bounds suggest that operator norm-based regularization methods could improve the transferability of the designed adversarial examples. We support our theoretical results by performing several numerical experiments showing the role of the substitute network's generalization in generating transferable adversarial examples. Our empirical results indicate the power of Lipschitz regularization and early stopping methods in improving the transferability of designed adversarial examples.

1 INTRODUCTION

Deep neural networks (DNNs) have attained impressive results in many machine learning problems from image recognition [Krizhevsky et al., 2009b], speech processing [Deng et al., 2013], and bioinformatics [Alipanahi et al., 2015]. The standard evaluation of a trained DNN machine is typically performed over test samples drawn from the same underlying distribution that has generated the empirical training data. The numerous successful applications of deep learning models reported in the literature demonstrate DNNs' surprising generalization power from training samples to unseen test data. Such promising results on unobserved data despite DNNs' enormous capacity for memorizing training examples have attracted a lot of attention in the machine learning community.

While DNNs usually achieve satisfactory generalization performance, they have been frequently observed to lack robustness against minor adversarial perturbations to their input data [Szegedy et al., 2013, Biggio et al., 2013, Goodfellow et al., 2014], widely known as adversarial attacks. According to these observations, an adversarial attack scheme can generate imperceptible perturbations that fools the DNN classifier to predict wrong labels with high confidence scores. Such adversarial perturbations are usually created through maximizing a target DNN's prediction loss over a small neighborhood around an input sample. While DNNs often show successful generalization behavior to test samples drawn from the underlying distribution of training data, the minor perturbations designed by adversarial attack schemes can completely undermine their prediction results.

Specifically, adversarial examples have been commonly reported to be capable of transferring to unseen DNN classifiers [Tramèr et al., 2017a, Ilyas et al., 2018, Cheng et al., 2018, Zhou et al., 2018]. Based on these reports, an adversarial example designed for a specific classifier could further alter the prediction of another DNN machine with a different architecture and training set. Such observations have inspired the development of several *black-box adversarial attack schemes* in which the adversarial examples are designed for a substitute classifier and then are evaluated on a different target DNN.

Several recent papers have attempted to theoretically study the transferability of black-box adversarial attacks. These

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:2259–2270.

works have mostly focused on the effects of non-robust features [Tramèr et al., 2017b, Ilyas et al., 2019, Inkawhich et al., 2019], causality [Zhang et al., 2021], and equilibrium [Bose et al., 2020, Meunier et al., 2021] in adversarial training problems on transferable adversarial examples. The mentioned studies reveal the dependency of adversarial examples on non-robust features that can be easily perturbed through minor adversarial noise, and also how the transferability of adversarial examples depends on the equilibrium in the game between the adversary and classifier players. On the other hand, the connection between the train-to-test generalization performance of the substitute network and the transferability of the designed examples has not been explored in the literature. Hence, it remains unclear whether a substitute DNN with a smaller generalization gap results in more transferable adversarial examples.

In this work, we attempt to understand the interconnections between the train-to-test generalization error and the attack transferability rate of DNNs in black-box adversarial attacks. We aim to show that a smaller generalization gap not only improves the classification accuracy on unseen test data, but further could result in higher transferability rates for the designed adversarial examples. To this end, we analyze the transferability of adversarial examples through the lens of the max-min framework of *Adversarial Example Game (AEG)* introduced by Bose et al. [2020]. According to this approach, the adversary player searches for the most transferable attack strategy that reaches the maximum prediction error under the most robust DNN classifier. We focus on the generalization performance of the AEG learner from training samples to test data, and demonstrate its importance in the transferability power of the generated adversarial perturbations.

Specifically, we focus on the standard class of norm-bounded adversarial attacks and define the train-to-test generalization error of a function class's minimum risk under norm-bounded adversarial perturbations. Subsequently, we prove theoretical bounds on the defined generalization error for multi-layer DNNs with spectrally-normalized weight matrices, which enables us to bound the generalization gap between the training and test transferability rates of norm-bounded attack schemes. Also, the shown generalization bound suggests the application of Lipschitz regularization methods in training a substitute DNN with improved transferability of generated adversarial examples.

Finally, we numerically evaluate our theoretical results on multiple standard image recognition datasets and DNN architectures. Our empirical results further support the existing connections between the generalization and transferability properties of black-box adversarial attacks. The numerical findings demonstrate that a better generalization score for the substitute DNN could significantly boost the transferability rate of designed adversarial examples. In addition, we empirically demonstrate that both explicit and implicit regularization techniques can help generate more transferable examples. We validate this result for explicit Lipschitz regularization and implicit early-stopping schemes. We can summarize the main contributions of our work as follows:

- Drawing connections between the generalization properties of the substitute DNN classifier and the transferability rate of designed adversarial examples
- Proving generalization error bounds on the difference between the transferability rates of DNN-based adversarial examples designed for training and test data
- Demonstrating the power of Lipschitz regularization and early stopping methods in generating more transferable adversarial examples
- Conducting numerical experiments on the generalization and transferability aspects of black-box adversarial attacks

2 RELATED WORK

Transferability of adversarial examples has been extensively studied in the deep learning literature. The related literature includes a large body of papers [Ilyas et al., 2018, Cheng et al., 2018, Bhagoji et al., 2018, Alzantot et al., 2019, Cheng et al., 2019, Moon et al., 2019, Guo et al., 2019, Mohaghegh Dolatabadi et al., 2020, Wang et al., 2020] proposing black-box adversarial attack schemes aiming to transfer from a source DNN to an unseen target DNN classifier and several related works [Levine and Feizi, 2020, Salman et al., 2020, Singla and Feizi, 2020, Li et al., 2020] on developing robust training mechanisms against black-box adversarial attacks. Regarding the relationship between accuracy and transferability, [Wu et al., 2018] observes a positive correlation between the clean accuracy and transferability of adversarial examples following the neural net. On the other hand, Gubri et al. [2022] report that the best clean test accuracy does not provide the highest transferability rate. [Qin et al., 2022, Gubri et al., 2022] also study the relationship between transferability rate and the loss function's sharpness.

In addition, several game theoretic frameworks have been proposed to analyze the transferability of adversarial examples. The related works [Bose et al., 2020, Meunier et al., 2021] study the adversarial example game between the classifier and adversary players. However, these works mostly focus on the equilibrium and convergence behavior in adversarial example games and do not discuss the generalization aspect of the game. In another related work, Pal and Vidal [2020] study the adversarial learning task through the lens of game theory. Unlike our work, the generalization analysis in [Pal and Vidal, 2020] focuses only on the generalization behavior of the robust classification rule and not on the generalization properties of the transferable adversary player.

Furthermore, the generalization properties of adversarially-learned models have been the topic of several related papers. References [Schmidt et al., 2018, Raghunathan et al., 2019] discuss numerical and theoretical results that generalization of adversarially-trained neural nets is inferior to that of standard ERM-learned models with the same number of training data. The related work by Rice et al. [2020] empirically studies the overfitting phenomenon in adversarial training problems and reveals the different generalization properties of standard and adversarial training schemes. In another study, Wu et al. [2020] show the connection between the generalization of adversarially-learned models and the flatness of the weight loss landscape. [Yin et al., 2019, Awasthi et al., 2020] develop Rademacher-complexity-based generalization bounds for adversarially-trained models which suggest the application of norm-based regularization techniques for improving the generalization behavior of adversarial training methods. Farnia et al. [2018] prove Pac-Bayes generalization bounds for adversarially-learned DNNs with bounded spectral norms for their weight matrices. Also, Attias et al. [2019] perform VC-based generalization analysis for adversarial training schemes and derives upper-bounds on their sample complexity. However, we note that all these papers focus on the generalization of adversarially-trained models and do not study the connection between generalization and transferability of black-box attacks.

3 PRELIMINARIES: ADVERSARIAL ATTACKS AND TRAINING

In this section, we give a brief review of standard norm-bounded adversarial attack and training schemes. Consider a supervised learning problem where the learner seeks a prediction rule f from function space $\mathcal{F}$ to predict a label variable $Y \in \mathcal{Y}$ from the observation of a d-dimensional feature vector $\mathbf{X} \in \mathcal{X}$. In this work, we focus on the following set of L-layer neural network functions with activation function ψ:

$$\mathcal{F}_{\mathcal{V}} = \Big\{ f_{\mathbf{v}} : f_{\mathbf{v}}(\mathbf{x}) = V_L \psi\big(\cdots \psi(V_0 \mathbf{x}) \cdot\big),\ \mathbf{v} \in \mathcal{V} \Big\} \tag{1}$$

In the above, we use vector $\mathbf{v}$ belonging to feasible set $\mathcal{V}$ to parameterize the L-layer neural net $f_{\mathbf{v}}$. According to this notation, $\mathbf{v}$ concatenates all the entries of the neural net's weight matrices $V_0, \ldots, V_L$.

Given a loss function ℓ and n training samples in dataset $S = \{(\mathbf{x}_i, y_i)_{i=1}^n\}$, the standard risk minimization approach aims to find the prediction rule $f^* \in \mathcal{F}_{\mathcal{V}}$ minimizing the expected loss (risk) $\mathbb{E}[\ell(f(\mathbf{X}), Y)]$ where the expectation is taken according to the underlying distribution of data $P_{\mathbf{X},Y}$. Since the supervised learner only observes the training samples and lacks any further knowledge of the underlying $P_{\mathbf{X},Y}$, the empirical risk minimization (ERM) framework sets out to minimize the empirical risk function estimated using the training examples:

$$\min_{\mathbf{v} \in \mathcal{V}} \ \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_{\mathbf{v}}(\mathbf{x}_i), y_i\big). \tag{2}$$

However, the ERM learner typically lacks robustness to norm-bounded adversarial perturbations. A standard approach to generate a norm-bounded adversarial perturbation is through maximizing the loss function over a norm ball around a given data point $(\mathbf{x}, y)$:

$$\max_{\boldsymbol{\delta}:\, \|\boldsymbol{\delta}\| \le \epsilon} \ \ell\big(f(\mathbf{x} + \boldsymbol{\delta}), y\big). \tag{3}$$

Here $\boldsymbol{\delta} \in \mathbb{R}^d$ represents the d-dimensional perturbation vector added to the feature vector $\mathbf{x}$, and $\| \cdot \|$ denotes a norm function used to measure the attack power that is bounded by parameter $\epsilon \ge 0$.

In order to gain robustness against norm-bounded perturbations, the adversarial training (AT) scheme [Madry et al., 2017] alters the ERM objective function to the expected worst-case loss function over norm-bounded adversarial perturbations and solves the following min-max optimization problem:

$$\begin{aligned} &\min_{\mathbf{v} \in \mathcal{V}} \ \frac{1}{n} \sum_{i=1}^{n} \Big[\max_{\boldsymbol{\delta}_i:\, \|\boldsymbol{\delta}_i\| \le \epsilon} \ \ell\big(f_{\mathbf{v}}(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big) \Big] \\ \equiv &\min_{\mathbf{v} \in \mathcal{V}} \ \max_{\substack{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n: \\ \forall i,\, \|\boldsymbol{\delta}_i\| \le \epsilon}} \ \frac{1}{n} \sum_{i=1}^{n} \big[\ell\big(f_{\mathbf{v}}(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big) \big] \end{aligned} \tag{4}$$

Note that the above minimax problem indeed estimates the solution to the following learning problem formulated over the true distribution of data $P_{\mathbf{X},Y}$:

$$\min_{\mathbf{v} \in \mathcal{V}} \ \mathbb{E}_{(\mathbf{X},Y) \sim P} \Big[\max_{\boldsymbol{\delta}:\, \|\boldsymbol{\delta}\| \le \epsilon} \ \ell\big(f_{\mathbf{v}}(\mathbf{X} + \boldsymbol{\delta}), Y\big) \Big]. \tag{5}$$

It can be seen that the above optimization problem is indeed equivalent to the following min-max problem where the maximization is performed over Δ_ϵ containing all mappings $\delta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ whose output is ϵ-norm-bounded, i.e. $\forall \mathbf{x}, y : \ \|\delta(\mathbf{x}, y)\| \le \epsilon$:

$$\min_{\mathbf{v} \in \mathcal{V}} \ \max_{\delta \in \Delta_\epsilon} \ \mathbb{E}_{\mathbf{X},Y \sim P} \big[\ell\big(f_{\mathbf{v}}(\mathbf{X} + \delta(\mathbf{X}, Y)), Y\big) \big]. \tag{6}$$

In next sections, we will discuss the association between the above min-max problem and the adversarial example game for generating transferable adversarial examples.

4 A MAX-MIN APPROACH TO TRANSFERABLE ADVERSARIAL EXAMPLES

The transferability of adversarial examples has been extensively studied in the literature. A useful framework to theoretically study transferable examples is the max-min framework of *adversarial example game (AEG)* proposed by Bose et al. [2020]. According to this approach, the adversary searches for the most transferable attack scheme $\delta \in \Delta$ from a set of attack strategies Δ that achieves the maximum expected loss under the most robust classifier $f_{\mathbf{v}} \in \mathcal{F}_{\mathcal{V}}$ from DNN function space $\mathcal{F}_{\mathcal{V}}$. Therefore, the AEG approach reduces the transferable adversary's task to solving the following max-min optimization problem:

$$\max_{\delta\in\Delta}\ \min_{\mathbf{v}\in\mathcal{V}}\ \frac{1}{n}\sum_{i=1}^{n}\Big[\ell\big(f_{\mathbf{v}}(\mathbf{x}_i+\delta(\mathbf{x}_i,y_i)),y_i\big)\Big] \tag{7}$$

The above bi-level optimization problem indeed swaps the maximization and minimization order of the AT optimization problem, and focuses on the max-min version of the min-max AT optimization task. Note that as shown by Meunier et al. [2021], the adversarial example game is in general not guaranteed to have a pure Nash equilibrium where each player's deterministic strategy is optimal when fixing the other player's strategy. Due to the lack of pure Nash equilibria, the AEG max-min and AT min-max optimization problems may not share any common solutions.

Note that the AEG framework introduces the following metric for evaluating the transferability of an attack scheme $\delta:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}^d$:

$$\widehat{\mathcal{L}}_{\text{transfer}}(\delta)\ :=\ \min_{\mathbf{v}\in\mathcal{V}}\ \frac{1}{n}\sum_{i=1}^{n}\big[\ell\big(f_{\mathbf{v}}(\mathbf{x}_i+\delta(\mathbf{x}_i,y_i)),y_i\big)\big] \tag{8}$$

The above transferability score indeed estimates the following score measuring transferability under the underlying distribution $P_{\mathbf{X},Y}$:

$$\mathcal{L}_{\text{transfer}}(\delta) := \min_{\mathbf{v}\in\mathcal{V}}\ \mathbb{E}_{P_{X,Y}}\Big[\ell\left(f_{\mathbf{v}}(\mathbf{X}+\delta(\mathbf{X},Y)),Y\right)\Big]. \tag{9}$$

Based on this discussion, the AEG optimization problem in (7) similarly estimates the solution to the following max-min AEG problem formed around the underlying distribution $P_{\mathbf{X},Y}$:

$$\begin{aligned}&\max_{\delta\in\Delta}\ \mathcal{L}_{\text{transfer}}(\delta)\ \equiv\\ &\max_{\delta\in\Delta}\ \min_{\mathbf{v}\in\mathcal{V}}\ \mathbb{E}_{(\mathbf{X},Y)\sim P}\Big[\ell\left(f_{\mathbf{v}}(\mathbf{X}+\delta(\mathbf{X},Y)),Y\right)\Big].\end{aligned} \tag{10}$$

Therefore, the primary goal of the transferable adversary is to solve the above problem targeting the distribution of test data instead of training examples. However, since the true distribution is unknown to the adversary, the AEG framework switches to the empirical max-min problem (7). This discussion motivates the following definition of the generalization error for adversarial examples' transferability performance:

Definition 1. *We define the generalization error of an attack scheme $\delta:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}^d$ over DNN classifier space $\mathcal{F}_{\mathcal{V}}$ as follows:*

$$\begin{aligned}\epsilon_{\text{gen}}(\delta) &:= \widehat{\mathcal{L}}_{\text{transfer}}(\delta) - \mathcal{L}_{\text{transfer}}(\delta) \\ &= \min_{\mathbf{v}\in\mathcal{V}}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big[\ell\big(f_{\mathbf{v}}(\mathbf{x}_i+\delta(\mathbf{x}_i,y_i)),y_i\big)\Big]\Big\} \\ &\quad - \min_{\mathbf{v}\in\mathcal{V}}\Big\{\mathbb{E}\Big[\ell\big(f_{\mathbf{v}}(\mathbf{X}+\delta(\mathbf{X},Y)),Y\big)\Big]\Big\}.\end{aligned} \tag{11}$$

Note that the above definition is consistent with the standard definition of generalization error in minimax learning frameworks such as generative adversarial network (GAN) and adversarial training approaches in the literature [Arora et al., 2017, Yin et al., 2019, Farnia and Ozdaglar, 2020, Xing et al., 2021, Farnia and Ozdaglar, 2021, Lei et al., 2021] where the generalization error of the min (or max) player is defined as the difference between the worst-case empirical and population objectives under the other player's optimal action. Therefore, in order for a black-box adversarial attack to be effective, we need the attack scheme to generalize well from training samples to test data, and based on the max-min AEG framework the generalization error is defined in the sense of Definition 1.

5 A GENERALIZATION BOUND FOR ADVERSARIAL EXAMPLE GAMES

In this section, we aim to analyze the generalization error of a black-box adversarial attack scheme based on the substitute classifier of a L-layer DNN $\mathcal{H}_{\mathcal{W}}$. To characterize a one-to-one correspondence between the choice of the DNN weights and the assigned attack scheme, we consider the following definition of an optimal attack scheme for a substitute neural net $h_{\mathbf{w}} \in \mathcal{H}_{\mathcal{W}}$, which revisits the distributionally robust optimization approach to the adversarial training problem [Sinha et al., 2017].

Definition 2. *Given a classifier $h_{\mathbf{w}}$, we call the attack scheme $\delta^*_{\mathbf{w}}:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}^d$ λ-optimal if it solves the following optimization problem:*

$$\max_{\delta:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}^d}\ \mathbb{E}\big[\ell\big(h_{\mathbf{w}}(\mathbf{X}+\delta(\mathbf{X},Y)),Y\big)\big]-\frac{\lambda}{2}\mathbb{E}\big[\|\delta(\mathbf{X},Y)\|^2\big].$$

The above definition of a λ-optimal attack revisits the notion of Wasserstein-based distributional adversarial attacks in the distributionally robust optimization literature [Sinha et al., 2017], where the attack norm bound parameterized by ϵ implicitly depends on coefficient λ. Here, the definition of λ-optimal attacks employs a regularization term to penalize the averaged norm-squared of perturbations. As shown in Proposition 1, this definition allows us to establish a one-to-one correspondence between λ-optimal attack schemes and

λ-smooth DNN classifiers. The one-to-one correspondence property addresses the intractable nature of the analysis of an optimal ϵ-norm bounded adversarial attack scheme which could be non-unique for non-convex neural nets.

Proposition 1. *Consider the L_2-norm function $\|\cdot\|_2$ for measuring the attack power. Suppose that the composition $\ell \circ h_{\mathbf{w}}$ is a λ-smooth differentiable function of $\mathbf{x}$, i.e. for every $\mathbf{x}, \mathbf{x}', y$ we have $\|\nabla_{\mathbf{x}}\ell(h_{\mathbf{w}}(\mathbf{x}), y) - \nabla_{\mathbf{x}}\ell(h_{\mathbf{w}}(\mathbf{x}'), y)\|_2 \le \lambda\|\mathbf{x} - \mathbf{x}'\|_2$. Then, there exists a unique λ-optimal attack scheme $\delta^*(\mathbf{x}, y)$ for $h_{\mathbf{w}}$ given by:*

$$\delta^*(\mathbf{x}, y) \,=\, \Big(\mathrm{Id}_{\mathbf{x}} - \frac{1}{\lambda}\nabla_{\mathbf{x}}\ell \circ h_{\mathbf{w}}\Big)^{-1}(\mathbf{x}, y) \,-\, \mathbf{x}.$$

In the above equation $\mathrm{Id}_{\mathbf{x}}$ represents the identity function on feature vector $\mathbf{x}$, and $(\cdot)^{-1}$ denotes the inverse of an invertible transformation.

Proof. We defer the proof to the Appendix. □

The above proposition reveals a bijection between smooth DNN classifiers and optimal attack schemes. Therefore, in our generalization analysis, we focus on bounding the generalization error for the resulting λ-optimal attack schemes corresponding to λ-smooth DNN substitute classifiers.

In the following theorem, we show a generalization error bound for the class of λ-optimal black-box attack schemes coming from spectrally-regularized DNN functions. This theorem extends the uniform convergence generalization bounds [Bartlett et al., 2017, Neyshabur et al., 2017] from standard deep supervised learning problems to the max-min adversarial example game learning framework. In the theorem, we use the following set of assumptions on the loss function ℓ and the target and substitute classes of neural networks. Also, note that $\|\cdot\|_2$ denotes the L_2-operator (spectral) norm in application to a matrix, i.e. the matrix's maximum singular value, and $\|\cdot\|_{2,1}$ denotes the $(2,1)$-norm of a matrix which is the summation of the L_2-norms of the matrix's rows.

Assumption 1. *Loss function $\ell(y, y')$ is a c-bounded, 1-Lipschitz, and 1-smooth function of the input y, i.e. for every $y_1, y_2, y' \in \mathcal{Y}$ we have $|\ell(y_1, y')| \le c$, $|\ell(y_1, y') - \ell(y_2, y')| \le \|y_1 - y_2\|_2$, and $\|\nabla_y \ell(y_1, y') - \nabla_y \ell(y_2, y')\|_2 \le \|y_1 - y_2\|_2$.*

Assumption 2. *The set of substitute DNNs in the black-box attack scheme $\mathcal{H}_{\mathcal{W}} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ contains L-layer neural networks $h_{\mathbf{w}}(\mathbf{x}) = W_L\phi_L\big(W_{L-1}\phi_{L-1}(\cdots W_1\phi_1(W_0\mathbf{x})\cdot\big)$. We suppose that the dimensions of matrices $W_0, \ldots, W_k$ is bounded by D, and assume every activation ϕ_i satisfies $\phi_i(0) = 0$ and is γ_i-Lipschitz and γ_i-smooth, i.e. $\max\{|\phi_i'(z)|, |\phi_i''(z)|\} \le \gamma_i$ holds for every $z \in \mathbb{R}$.*

Assumption 3. *The class of target classifiers $\mathcal{F}_{\mathcal{V}} = \{f_{\mathbf{v}} : \mathbf{v} \in \mathcal{V}\}$ consists of K-layer neural network functions $f_{\mathbf{v}}(\mathbf{x}) = V_K\psi_L\big(V_{L-1}\psi_{L-1}(\cdots V_1\psi_1(V_0\mathbf{x})\cdot\big)$ with activation function ψ_i's. We suppose that the dimensions of matrices $V_0, \ldots, V_k$ is bounded by D. Also, we assume every ψ_i satisfies $\psi_i(0) = 0$ and is ξ_i-Lipschitz, i.e. $\max_z\, |\psi_i'(z)| \le \xi_i$. Also, we define the capacity $R_{\mathcal{V}}$ as*

$$R_{\mathcal{V}} := \sup_{\mathbf{v}\in\mathcal{V}} \left\{ \big(\prod_{i=0}^{K} \xi_i \|V_i\|_2\big)\Big(\sum_{i=0}^{K} \frac{\|V_i^\top\|_{2,1}^{2/3}}{\|V_i\|_2^{2/3}}\Big)^{3/2} \right\}.$$

Theorem 1. *Suppose that the loss function, substitute DNN, and target DNN in a black-box adversarial attack satisfy Assumptions 1, 2 and 3. Assuming $\|\mathbf{X}\|_2 \le B$ for the $n \times d$ data matrix $\mathbf{X}$ and $\lambda(1-\tau) \ge (\prod_{i=0}^{L}\gamma_i\|W_i\|_2)\sum_{i=0}^{L}\prod_{j=0}^{L}\gamma_j\|W_j\|_2$ holds for constant $\tau > 0$ and every $\mathbf{w} \in \mathcal{W}$, then for every $\omega > 0$ with probability at least $1 - \omega$ the following bound will hold for every $\mathbf{w} \in \mathcal{W}$:*

$$\begin{aligned}\epsilon_{\text{gen}}(\delta^*_{\mathbf{w}}) \;\le \mathcal{O}\bigg(c\sqrt{\frac{\log(1/\omega)}{n}} \\ + \frac{(B + \frac{L_{\mathbf{w}}}{\lambda})\big(R_{\mathcal{V}} + \frac{1}{\tau^2}L_{\mathbf{w}}R_{\mathbf{w}}\big)\log(n)\log(D)}{n}\bigg)\end{aligned} \tag{12}$$

where the Lipschitz and capacity terms $L_{\mathbf{w}},\, R_{\mathbf{w}}$ are defined as:

$$\begin{aligned} L_{\mathbf{w}} &:= \prod_{i=0}^{L}\gamma_i\|W_i\|_2, \\ R_{\mathbf{w}} &:= \bigg(\sum_{i=0}^{L}\prod_{j=0}^{i}\gamma_j\|W_j\|_2\bigg)\bigg(\sum_{i=0}^{L}\frac{\|W_i^\top\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\bigg)^{3/2}. \end{aligned} \tag{13}$$

Proof. We defer the proof to the Appendix. □

The above theorem bounds the generalization error of the attack scheme $\delta^*_{\mathbf{w}}$ corresponding to the substitute DNN $f_{\mathbf{w}}$ in terms of the spectral capacity of the substitute network. As a result, this bound motivates norm-based spectral regularization [Yoshida and Miyato, 2017, Miyato et al., 2018, Farnia et al., 2018] for improving the generalization performance of black-box attack schemes.

6 NUMERICAL RESULTS

In this section, we provide the results of our numerical experiments for validating the connection between the generalization and transferability properties of black-box adversarial attacks. The numerical discussion focuses on the question of whether achieving a better generalization score for the substitute DNN can improve the success of the designed perturbations in application to a different DNN classifier. To answer this question, we tested an explicit norm-based

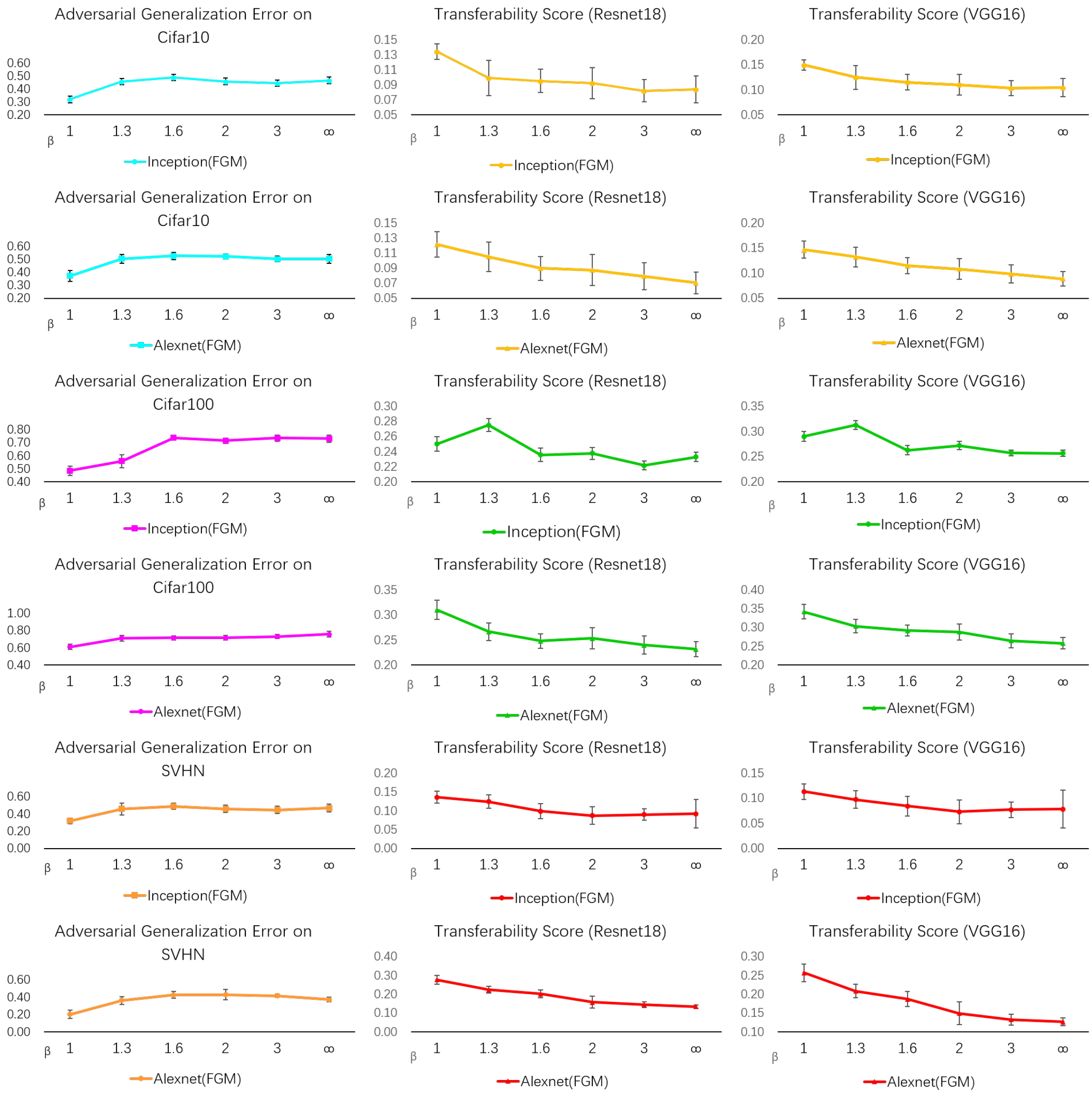

Figure 1: Generalization errors of substitute DNNs (the lower the better), and transferability rates of adversarial examples generated from the substitute model (the higher the better) for CIFAR-10 (rows 1-2), CIFAR-100 (rows 3-4) and SVHN (rows 5-6) datasets. ResNet18 and VGG-16 architectures were used as the target DNNs.

regularization method, spectral normalization [Yoshida and Miyato, 2017, Tsuzuku et al., 2018, Farnia et al., 2018], as well as an implicit regularization technique, early stopping [Yao et al., 2007, Rice et al., 2020], to evaluate the power of these regularization methods in attaining more transferable black-box attacks.

For generating norm-bounded perturbations, we used standard projected gradient descent (PGD) and fast gradient method (FGM) [Goodfellow et al., 2014] to design perturbations. We implemented the PGD and FGM algorithms by projecting the perturbations according to both standard L_2-norm and L_∞-norm, where the latter results in the widely-used fast gradient sign method (FGSM) attack scheme [Goodfellow et al., 2014] in the FGM case. For simulating L_2-norm-bounded perturbations, we chose the maximum L_2-norm (attack power) as $\epsilon = \gamma \mathbb{E}_{\hat{P}}[\|X\|_2]$ with $\gamma = 0.05$ unless stated otherwise. For L_∞-norm-bounded attacks, we chose $\epsilon = 8/255$ for the normalized samples. For optimizing PGD perturbations, we applied $r = 15$ PGD steps, where we used the standard rule $\alpha = 1.5\epsilon/r$ to choose the stepsize parameter α. We trained every DNN model for 100 epochs using the Adam optimizer [Kingma and Ba, 2014] with a batch-size of 128. The numerical experiments were implemented using the PyTorch platform and were run on one standard RTX-3090 GPU.

In our experiments, we used three standard image recognition datasets: 1) CIFAR-10, 2) CIFAR-100 [Krizhevsky et al., 2009a], 3) SVHN [Netzer et al., 2011], and the following four neural network architectures: 1) AlexNet [Krizhevsky et al., 2012], 2) Inception-Net [Szegedy et al.,

Dataset	Model	Method	β	Generalization Error	Transferability Rate (VGG-16)	Transferability Rate (ResNet-18)
Cifar10	AlexNet	PGD	∞	0.545± 0.031	0.105±0.011	0.087±0.009
			1.0	0.342±0.022	0.162±0.012	0.139±0.01
		I-FGM	∞	0.512±0.022	0.093±0.014	0.077±0.031
			1.0	0.414±0.018	0.149±0.022	0.123±0.009
		FGM	∞	0.505±0.028	0.089±0.007	0.070±0.007
			1.0	0.451±0.022	0.147±0.014	0.122±0.011
	Inception	PGD	∞	0.508±0.020	0.104±0.009	0.084±0.010
			1.0	0.258±0.015	0.150±0.008	0.134±0.007
		I-FGM	∞	0.487±0.011	0.093±0.010	0.081±0.010
			1.0	0.288±0.017	0.113±0.011	0.122±0.011
		FGM	∞	0.466±0.019	0.092±0.012	0.078±0.011
			1.0	0.320±0.031	0.136±0.018	0.113±0.015
Cifar100	AlexNet	PGD	∞	0.789±0.055	0.229±0.032	0.260±0.031
			1.0	0.601± 0.043	0.323±0.023	0.353±0.025
		I-FGM	∞	0.777±0.033	0.265±0.028	0.277±0.021
			1.0	0.655±0.030	0.313±0.021	0.321±0.026
		FGM	∞	0.758±0.041	0.258±0.019	0.232±0.023
			1.0	0.611±0.037	0.342±0.022	0.310±0.020
	Inception	PGD	∞	0.602±0.03	0.303±0.017	0.270±0.021
			1.0	0.494±0.028	0.330±0.017	0.301±0.020
		I-FGM	∞	0.700±0.040	0.288±0.054	0.255±0.019
			1.0	0.565±0.031	0.331±0.037	0.288±0.043
		FGM	∞	0.717±0.033	0.268±0.017	0.236±0.017
			1.0	0.558±0.030	0.313±0.020	0.275±0.019
SVHN	AlexNet	PGD	∞	0.298±0.020	0.211±0.018	0.225±0.017
			1.0	0.199±0.012	0.276±0.008	0.292±0.011
		I-FGM	∞	0.334±0.015	0.187±0.018	0.199±0.018
			1.0	0.211±0.015	0.279±0.013	0.287±0.014
		FGM	∞	0.373±0.021	0.134±0.013	0.126±0.013
			1.0	0.203±0.013	0.277±0.014	0.257±0.016
	Inception	PGD	∞	0.342±0.021	0.193±0.017	0.177±0.018
			1.0	0.115±0.010	0.339±0.021	0.313±0.019
		I-FGM	∞	0.366±0.011	0.156±0.011	0.166±0.015
			1.0	0.187±0.012	0.301±0.009	0.288±0.011
		FGM	∞	0.373±0.022	0.134±0.015	0.126±0.016
			1.0	0.203±0.014	0.277±0.018	0.257±0.016

Table 1: Generalization error (Gen. Err.) and L_2-norm-based adversarial examples' transferability rates on three image datasets, with and without spectral regularization ($\beta = \infty$ means no spectral regularization).

2015], 3) VGG-16 [Simonyan and Zisserman, 2015], 4) ResNet-18 [He et al., 2016]. In the reported results, we evaluate a prediction model's generalization performance using the accuracy gap between the training and test sets. For evaluating the transferability performance, we used the generated black-box adversarial examples and measured the transferability rate as the target network's averaged classification error over the designed adversarial examples on the test set. Therefore, a higher transferability rate implies more transferable adversarial examples, which implies that under a worse transferability score for training data, which is the case under a stronger norm-based regularization, the generalization of the attack scheme has improved.

In the transferability evaluation of the generated adversarial examples, we considered only the samples for which their clean data had been labeled correctly by the target network, because we expect the clean version of an adversarial example to be labeled correctly by the target network. Also, we used different training sets for the substitute and target classifiers to separate the generalization effects of the substitute and target DNNs. To do this, we split the training set in half and used each half for training one of the classifiers. Finally, consistent to our theoretical analysis, we used PGD adversarial training for training the substitute DNN and applied standard ERM training for training the target DNNs.

6.1 TRANSFERABILITY UNDER SPECTRAL REGULARIZATION

We evaluated the generalization and transferability performance of the discussed black-box attack schemes for Lipschitz-regularized neural nets. To apply spectral regular-

Dataset	Model	Method	Generalization Error	Transferability Rate (VGG-16)	Transferability Rate (ResNet-18)
Cifar10	Inception	PGD	0.517±0.027	0.127±0.013	0.104±0.012
		PGD-ES	0.073±0.018	0.198±0.014	0.172±0.011
		I-FGM	0.488±0.017	0.114±0.014	0.108±0.012
		I-FGM-ES	0.112±0.016	0.181±0.015	0.156±0.014
		FGM	0.467±0.017	0.100±0.012	0.089±0.012
		FGM-ES	0.126±0.031	0.170±0.010	0.147±0.014
	AlexNet	PGD	0.579±0.037	0.098±0.009	0.077±0.007
		PGD-ES	0.061±0.041	0.154±0.017	0.136±0.017
		I-FGM	0.533±0.054	0.102±0.023	0.098±0.016
		I-FGM-ES	0.077±0.054	0.149±0.031	0.132±0.014
		FGM	0.520±0.039	0.100±0.007	0.087±0.005
		FGM-ES	0.092±0.007	0.152±0.010	0.127±0.011
Cifar100	Inception	PGD	0.646±0.017	0.283±0.016	0.258±0.009
		PGD-ES	0.137±0.014	0.330±0.011	0.286±0.012
		I-FGM	0.688±0.016	0.284±0.022	0.254±0.021
		I-FGM-ES	0.165±0.010	0.333±0.024	0.289±0.019
		FGM	0.711±0.013	0.270±0.017	0.239±0.013
		FGM-ES	0.146±0.013	0.327±0.014	0.289±0.008
	AlexNet	PGD	0.764±0.012	0.252±0.011	0.227±0.016
		PGD-ES	0.091±0.010	0.294±0.015	0.266±0.017
		I-FGM	0.744±0.010	0.252±0.021	0.221±0.017
		I-FGM-ES	0.097±0.015	0.303±0.018	0.256±0.021
		FGM	0.756±0.022	0.261±0.025	0.232±0.017
		FGM-ES	0.122±0.019	0.291±0.017	0.259±0.032
SVHN	Inception	PGD	0.341±0.028	0.207±0.015	0.220±0.014
		PGD-ES	0.057±0.006	0.298±0.017	0.322±0.021
		I-FGM	0.336±0.041	0.188±0.026	0.176±0.036
		I-FGM-ES	0.066±0.040	0.233±0.035	0.220±0.028
		FGM	0.380±0.039	0.136±0.017	0.129±0.019
		FGM-ES	0.180±0.022	0.213±0.019	0.219±0.020
	AlexNet	PGD	0.307±0.041	0.211±0.011	0.228±0.012
		PGD-ES	0.030±0.004	0.256±0.011	0.278±0.013
		I-FGM	0.337±0.012	0.187±0.012	0.200±0.023
		I-FGM-ES	0.067±0.022	0.255±0.023	0.267±0.018
		FGM	0.373±0.029	0.157±0.021	0.170±0.019
		FGM-ES	0.064±0.011	0.241±0.015	0.260±0.014

Table 2: Generalization error and adversarial examples' transferability with and without early stopping (ES)

ization, we used the spectral normalization method [Miyato et al., 2018, Farnia et al., 2018] constraining the L_2-operator norm of the substitute DNN's weight matrices. We define hyper-parameter β as the maximum allowed L_2-operator norm. Then, the standard spectral normalization method modifies each weight matrix W_i in (1) to $\widetilde{W}_i$:

$$\widetilde{W}_i := \frac{W_i}{\max\{1, \frac{\|W_i\|_2}{\beta}\}} = \begin{cases} W_i & \text{if } \|W_i\|_2 \le \beta, \\ \frac{\beta}{\|W_i\|_2} W_i & \text{otherwise.} \end{cases}$$

The above operation will regularize the matrix's operator norm to be upper-bounded by β.

Figure 1 shows the generalization error of the model and attack transferability rates of the generated perturbations using the substitute classifier AlexNet and Inception-Net under different spectral-norm hyperparameter β's. The numerical results show that in all cases through applying the stronger regularization coefficients $\beta = 1.0, 1.3$, the AlexNet and Inception classifiers achieve the highest generalization performance and attack transferability rates to the target ResNet18 and VGG16. Therefore, spectral regularization not only helped the DNN classifier gain a better generalization score, which is an expected outcome, but further improved the transferability of the perturbations to unseen DNNs with different architectures. These numerical results suggest the impact of the substitute DNN's generalization on the transferability of the adversarial examples.

Table 1 shows our numerical results validating the connection between the substitute DNN's generalization and L_2-norm-based designed adversarial examples' transferability.

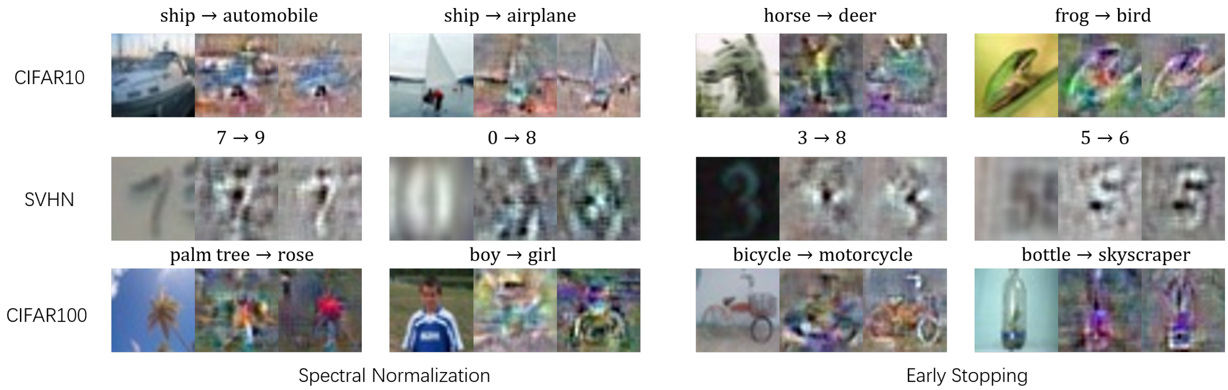

Figure 2: Visualization of adversarial perturbations. Each set of three pictures shows the original sample, the untransferable perturbation from the unregularized DNN, and the transferable perturbation generated by the regularized model (left to right). The perturbation is re-scaled to 0-255 for visualization. $A \rightarrow B$ indicates the groundtruth label A and the transferable example's predicted label B.

In this table, we report the performance of spectral regularization under the best β hyperparameter for validation samples. As can be seen in this table, spectral regularization manages to consistently improve the transferability rates of the adversarial examples, which confirms our hypothesis that better generalization will lead to more transferable adversarial examples. The numerical results for L_∞-norm-based adversarial examples can be found in the Appendix.

6.2 TRANSFERABILITY VIA EARLY STOPPING

Next, we used the implicit regularization mechanism of early stopping [Yao et al., 2007] to validate that better generalization achieved under early stopping can help to generate more transferable adversarial examples. To perform early stopping, we used 30% of the original test set as the validation set, and used the remaining 70% to measure the test accuracy. We stopped the DNN training when the trained model achieved its best performance on the validation samples.

We present the CIFAR-10 and SVHN numerical results in Table 2, and the complete set of obtained numerical results is in the Appendix. Our numerical results suggest that both the generalization and transferability scores considerably improve under early stopping regularization. The observation is consistent with the results reported in the literature [Benz et al., 2021] and our hypothesis on the impact of the generalization of the substitute network on the transferability of adversarial examples.

Finally, Figure 2 illustrates 12 uniformly-sampled transferable adversarial examples under spectral regularization and early stopping. We note that the adversarial examples designed by the unregularized DNN for these test samples failed to transfer to the target DNNs. We also observed that the transferable perturbations generated from a regularized DNN had sharper edges and less noise power in the background, and concentrated the power on the central part.

7 CONCLUSION

In this paper, we provided theoretical and numerical evidence on how the generalization properties of a substitute neural network can influence the transferability of the generated adversarial examples to other classifiers. While the transferability of black-box adversarial attacks and generalization power of the substitute classifier may seem two orthogonal factors, our results indicate existing interconnections between the two aspects. However, our bounds were based on uniform convergence analysis which cannot directly capture the interconnections between the generalization and optimization properties. An interesting future direction is to extend the generalization analysis to over-parameterized function spaces in order to understand the role of benign overfitting in the transferability of adversarial examples. Also, our experimental results motivate further studies of how other popular regularization methods in deep learning, such as batch normalization and dropout, can affect the transferability of adversarial perturbations.

ACKNOWLEDGMENTS

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and was partially supported by a CUHK Direct Research Grant. Also, the authors would like to thank the anonymous reviewers for their constructive feedback and suggestions.

References

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1111–1119, 2019.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR, 2017.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR, 2019.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7818–7827, 2021.

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and Will Hamilton. Adversarial example games. *Advances in neural information processing systems*, 33:8921–8934, 2020.

Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. Recent advances in deep learning for speech research at microsoft. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8604–8608. IEEE, 2013.

Farzan Farnia and Asuman Ozdaglar. Do gans always have nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR, 2020.

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

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

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 603–618. Springer, 2022.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7066–7074, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009a.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4585–4593, 2020.

Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Defending black-box adversarial attacks on deep neural networks. *arXiv preprint arXiv:2006.14042*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In *International Conference on Machine Learning*, pages 7677–7687. PMLR, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. *Advances in Neural Information Processing Systems*, 33:15871–15884, 2020.

Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *International Conference on Machine Learning*, pages 4636–4645. PMLR, 2019.

Y Netzer, T. Wang, A. Coates, A. Bissacco, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Ambar Pal and René Vidal. A game theoretic analysis of additive adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 33:1345–1355, 2020.

Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *arXiv preprint arXiv:2210.05968*, 2022.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

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

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations, ICLR*, 2015.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In *International conference on machine learning*, pages 8981–8991. PMLR, 2020.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–9, 2015.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017a.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017b.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Yang Liu. Amora: Black-box adversarial morphing attack. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1376–1385, 2020.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

Lei Wu, Zhanxing Zhu, Cheng Tai, et al. Understanding and enhancing the transferability of adversarial examples. *arXiv preprint arXiv:1802.09707*, 2018.

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

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

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

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.