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
works have mostly focused on the effects of non-robust features introduced by Bose et al. [2020]. According to this, we aim to show that a smaller generalization gap not only empirically demonstrates that both explicit and implicit architectures. Our empirical results further support the existence of 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 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, Mohagheghi 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, Singh 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., 2019] perform VC-based generalization analysis 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 $\psi$:

$$\mathcal{F}_V = \left\{ f_v : f_v(\mathbf{x}) = V_L \psi(\cdots \psi(V_0 \mathbf{x}) \cdot), \ v \in V \right\} \quad (1)$$

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

Given a loss function $\ell$ 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 $f^* \in \mathcal{F}_V$ 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:

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

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 $(\mathbf{x}, y)$:

$$\max_{\delta : ||\delta|| \leq \epsilon} \ell(f(\mathbf{x} + \delta), y). \quad (3)$$

Here $\delta \in \mathbb{R}^d$ 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 $\epsilon \geq 0$.

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:

$$\min_{v \in V} \max_{\delta : ||\delta|| \leq \epsilon} \frac{1}{n} \sum_{i=1}^{n} \ell(f_v(\mathbf{x}_i + \delta), y_i) \quad (4)$$

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

$$\min_{v \in V} \mathbb{E}_{(\mathbf{x}, Y) \sim P}[\max_{\delta : ||\delta|| \leq \epsilon} \ell(f_v(\mathbf{X} + \delta), Y)]. \quad (5)$$

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

$$\min_{v \in V} \max_{\delta \in \Delta_c} \mathbb{E}_{(\mathbf{x}, Y) \sim P}[\ell(f_v(\mathbf{X} + \delta(\mathbf{x}, Y)), Y)]. \quad (6)$$

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 \( \Delta \) that achieves the maximum expected loss under the most robust classifier \( f_\v \in \mathcal{F}_\mathcal{V} \) 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:

\[
\max_{\delta \in \Delta} \min_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} \left[ \ell \left( f_\v(x_i + \delta(x_i, y_i)), y_i \right) \right] \quad (7)
\]

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 \( \delta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d \):

\[
\hat{\mathcal{L}}_{\text{transfer}}(\delta) := \min_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} \left[ \ell \left( f_\v(x_i + \delta(x_i, y_i)), y_i \right) \right] \quad (8)
\]

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

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

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 \( P_{X,Y} \):

\[
\max_{\delta \in \Delta} \mathcal{L}_{\text{transfer}}(\delta) \equiv \max_{\delta \in \Delta} \min_{v \in \mathcal{V}} \mathbb{E}_{(X,Y) \sim P} \left[ \ell \left( f_\mathcal{V}(X + \delta(X, Y)), Y \right) \right]. \quad (10)
\]

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:

\[
\epsilon_{\text{gen}}(\delta) := \hat{\mathcal{L}}_{\text{transfer}}(\delta) - \mathcal{L}_{\text{transfer}}(\delta) = \min_{v \in \mathcal{V}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ \ell \left( f_\v(x_i + \delta(x_i, y_i)), y_i \right) \right] \right\} - \min_{v \in \mathcal{V}} \mathbb{E} \left[ \ell \left( f_\mathcal{V}(X + \delta(X, Y)), Y \right) \right].
\]

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 \( H_\mathcal{W} \). 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 \( h_\mathcal{w} \in H_\mathcal{W} \), which revisits the distributionally robust optimization approach to the adversarial training problem [Sinha et al., 2017].

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

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

The above definition of a \( \lambda \)-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 \( \epsilon \) implicitly depends on coefficient \( \lambda \). Here, the definition of \( \lambda \)-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 \( \lambda \)-optimal attack schemes and...
\(\lambda\)-smooth DNN classifiers. The one-to-one correspondence property addresses the intractable nature of the analysis of an optimal \(\epsilon\)-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 \(\ell \circ h_w\) is a \(\lambda\)-smooth differentiable function of \(x\), i.e. for every \(x, x', y\) we have \(\|\nabla_x \ell(h_w(x), y) - \nabla_x \ell(h_w(x'), y)\|_2 \leq \lambda\|x - x'\|_2\). Then, there exists a unique \(\lambda\)-optimal attack scheme \(\delta^*(x, y)\) for \(h_w\) given by:

\[
\delta^*(x, y) = \left(\text{Id}_x - \frac{1}{\lambda} \nabla_x \ell \circ h_w\right)^{-1} (x, y) - x.
\]

In the above equation \(\text{Id}_x\) represents the identity function on feature vector \(x\), 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 \(\lambda\)-optimal attack schemes corresponding to \(\lambda\)-smooth DNN substitute classifiers.

In the following theorem, we show a generalization error bound for the class of \(\lambda\)-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 learning framework. In the theorem, we use the following set of assumptions on the loss function \(\ell\) 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 \(\|\ell(y_1, y')\| \leq c\), \(\|\ell(y_1, y') - \ell(y_2, y')\| \leq \|y_1 - y_2\|_2\), and \(\|\nabla_y \ell(y_1, y') - \nabla_y \ell(y_2, y')\|_2 \leq \|y_1 - y_2\|_2\).

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

**Assumption 3.** The class of target classifiers \(\mathcal{F}_Y = \{f_y : y \in \mathcal{Y}\}\) consists of \(K\)-layer neural network functions \(f_y(x) = V_K \psi_L(V_{K-1} \phi_{L-1} \cdots V_1 \psi_1(V_0 x))\) with activation function \(\psi_i\). We suppose that the dimensions of matrices \(V_0, \ldots, V_k\) is bounded by \(D\). Also, we assume every \(\psi_i\) satisfies \(\psi_i(0) = 0\) and is \(\xi_i\)-Lipschitz, i.e. \(\max\{|\psi'_i(z)|, |\psi''_i(z)|\} \leq \xi_i\). Also, we define the capacity \(R_Y\) as:

\[
R_Y := \sup_{v \in \mathcal{V}} \left\{ \left( \prod_{i=0}^{K} \xi_i \|V_i\|_2 \right) \left( \sum_{i=0}^{K} \|V_i^\top W_i\|_2^{2/3} \right)^{3/2} \right\}.
\]

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

\[
\epsilon_{\text{gen}}(\delta_w) \leq \mathcal{O}\left( \left( B + \frac{L_w \lambda}{\xi} \right) \left( R_Y + \frac{1}{\tau^2} L_w R_w \right) \frac{\log(n) \log(D)}{n} \right)
\]

where the Lipschitz and capacity terms \(L_w, R_w\) are defined as:

\[
L_w := \prod_{i=0}^{L} \gamma_i \|W_i\|_2,
\]

\[
R_w := \left( \sum_{i=0}^{L} \gamma_i \|W_i\|_2 \right) \left( \sum_{i=0}^{L} \|W_i^\top W_i\|_2^{2/3} \right)^{3/2}.
\]

**Proof.** We defer the proof to the Appendix.

The above theorem bounds the generalization error of the attack scheme \(\delta_w\) corresponding to the substitute DNN \(f_w\) 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
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 $L_\infty$-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 $\epsilon = \gamma \mathbb{E}_P[\|X\|_2]$ with $\gamma = 0.05$ unless stated otherwise. For $L_\infty$-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 $\alpha$. 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., 2015], 3) ResNet18, and 4) VGG-16 architectures were used as the target DNNs.

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.
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).

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-
The above operation will regularize the matrix’s operator norm. Then, the standard spectral normalization method modifies each weight matrix $W_i$ in (1) to $\tilde{W}_i$:

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

The above operation will regularize the matrix’s operator norm to be upper-bounded by $\beta$.

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 $\beta$’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 [2] shows our numerical results validating the connection between the substitute DNN’s generalization and $L_2$-norm-based designed adversarial examples’ transferability.
In this table, we report the performance of spectral regularization under the best $\beta$ 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 $L_\infty$-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


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.


