Characterizing Adversarial Transferability via Gradient Orthogonality and Smoothness

Zhuolin Yang [*1] **Linyi Li** [*1] **Xiaojun Xu** [*1] **Shiliang Zuo** [1] **Qian Chen** [2] **Benjamin Rubinstein** [3] **Ce Zhang** [4] **Bo Li** [1]

Abstract

Transferability is an intriguing property of adversarial examples. In this work we aim to first understand the sufficient conditions of transferability. A theoretical analysis yields two lower bounds for transferability based on data distribution similarity and model gradient similarity. We then prove an upper bound on transferability for low risk classifiers based on gradient orthogonality and smoothness. We demonstrate that under the condition of gradient orthogonality, smoother classifiers guarantee lower transferability. Finally, based on our theoretical analysis of transferability, we propose a simple yet effective strategy to train a robust ensemble with low transferability by enforcing model smoothness and gradient orthogonality between base models.

1. Introduction

Machine learning systems, especially those based on deep neural networks (DNNs), have become widely applied in numerous settings, including image recognition [19], speech recognition [11], and natural language processing [29]. However, recently it has been shown that DNNs are vulnerable to adversarial examples, which are able to mislead DNNs by adding small magnitude perturbations to the original instances [30; 10]. There have also been a number of efforts to explore adversarial examples in general machine learning systems beyond those on DNNs [1; 2; 21; 22; 9].

Though most of the attack strategies mentioned above require access to the information of target machine learning models (whitebox attacks), it has been found that even without knowledge about the exact target model, adversarial examples generated against one model can *transferably* attack others, giving rise to blackbox attacks [26; 28]. Some work have been conducted to understand the properties of transferability [31; 23; 7]. However, a rigorous theoretical analysis or explanation for transferability is still lacking in the literature. *Can we deepen our theoretical understanding on transforability? Can we then take advantage of any new theoretical understanding to enable better robust ensemble models?*

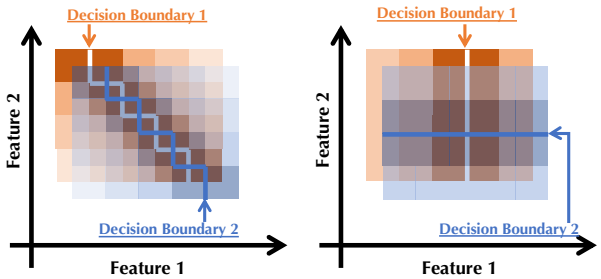

Figure 1: An illustration of the relationship between *transferability*, *gradient orthogonality*, and *smoothness*. (a) Gradient orthogonality alone cannot guarantee transferability as the decision boundaries between two classifiers can be arbitrarily close yet have orthogonal gradients almost everywhere; (b) Gradient orthogonality with smoothness provides a stronger guarantee on diversity, as our theorems will show.

In this paper, we focus on these two questions. From the theoretical side, we are interested in conditions under which the transferability can be *lower bounded* or *upper bounded*, both of which could lead to insights that could have profound empirical implications: *An upper bound on transferability could potentially deliver a new optimization objective when training robust ensemble models, while a lower bound on transferability could help avoid "doomed" scenarios*.

Intuition. Our theoretical arguments admit intuitive interpretation. As shown in Figure 1, gradient orthogonality between learning models cannot directly imply low transferability, which contrasts with common understanding; on the other hand only orthogonal and smoothed models would potentially limit transferability. Our analysis is inspired by this intuition, with a focus on understanding the impact of *model smoothness* and *gradient similarity* on transferability.

Contributions. We make a first attempt towards theoretical understanding of adversarial transferability, and providing

[*]Equal contribution [1]University of Illinois at Urbana-Champaign, USA [2]Tencent Inc., China [3]University of Melbourne, Australia [4]ETH Zurich, Switzerland. Correspondence to: Zhuolin Yang <zhuolin5@illinois.edu>, Linyi Li <linyi2@illinois.edu>, Xiaojun Xu <xiaojun3@illinois.edu>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

practical approaches to developing robust ensembles. Our contributions include: (1) We provide theoretical analysis for adversarial transferability. We prove lower bounds on transferability between low risk classifiers for both ℓ_p norm bounded and distribution enabled adversarial examples. (2) We prove an upper bound of transferability based on model similarity and smoothness, emphasizing the importance of model smoothness to decrease the transferability between models. We show that with smoother models, both the lower and upper bound are tighter. (3) We propose a simple yet effective approach to train a robust ensemble by enforcing model smoothness and reducing loss gradient similarity between models. (4) We conduct extensive experiments to evaluate the robustness of the proposed ensemble model, and show that it outperforms the state-of-the-art robust ensemble methods on multiple datasets against a range of attacks.

2. Transferability of Adversarial Perturbation

In this section, we connect transferability with different characteristics of models, which, in the next section, will allow us to explicitly minimize transferability by enforcing (or rewarding) certain properties of models. We first introduce necessary notation, then describe the attack model and theoretical analysis of adversarial transferability.

Let $\mathcal{X}$ be the *input space* on which predictions of *labels* $\mathcal{Y} \subset \mathbb{Z}$ are made. Denoting the product space as $\mathcal{M} = \mathcal{X} \times \mathcal{Y}$, we assume there exists a fixed but unknown probability measure Q on $\mathcal{M}$. A *classifier* $\mathcal{F}$ is a mapping from $\mathcal{X}$ into $\mathcal{Y}$. We consider the model-dependent loss function $\ell_{\mathcal{F}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Typically $\ell_{\mathcal{F}}(x, y)$ can be seen as the composition of the training loss and model output, i.e., $\ell_{\mathcal{F}}(x, y) = \ell(\mathcal{F}(x), y)$. Since $\ell_{\mathcal{F}}$ is used in training, we assume for simplicity of exposition that the loss function $\ell_{\mathcal{F}}$ is differentiable, i.e., $\nabla_x \ell_{\mathcal{F}}(x, y)$ exists.

We use P_x to represent the marginal distribution on $\mathcal{X}$, and $\Pr(E)$ to denote the probability of event E under P_x. A classifier's *risk* is defined as $\eta_{\mathcal{F}} = \Pr(\mathcal{F}(x) \neq y)$. A classifier's *empirical risk* is defined as $\xi_{\mathcal{F}} = \mathbb{E}_{(x,y)}[\ell_{\mathcal{F}}(x, y)]$, where $(x, y) \in \mathcal{M}$ is distributed according to Q.

We define an *attack strategy* as a function $\mathcal{A}(x) \in \mathcal{X}$ on *target point* $x \in \mathcal{X}$, which seeks an adversarial instance $\mathcal{A}(x)$, such that $\mathcal{F}(\mathcal{A}(x)) \neq \mathcal{F}(x)$. We use $P_{\mathcal{A}(x)}$ to represent the distribution of $\mathcal{A}(x) \in \mathcal{X}$ where $x \sim P_x$.

2.1. Attack Model

How should we define an adversarial attack? We first adopt a natural approach to defining an attack strategy — the attacker adds an ℓ_p norm bounded perturbation to data instance x. In practice, there are two types of attacks, *untargeted attacks* and *targeted attacks*. As previous work observed that the adversarial transferability is different under different attacks [23], we consider both in our analysis.

Definition 1 (Adversarial Attack)**.** *Consider an input x with existing label y, $\mathcal{F}(x) = y$.*

- *An untargeted attack satisfies that $\mathcal{A}_U(x) = x + \delta \in \underset{\delta:\|\delta\|_p \leq \epsilon}{\operatorname{argmax}}\ \ell_{\mathcal{F}}(x + \delta, y)$.*
- *A targeted attack with adversarial target y_t satisfies that $\mathcal{A}_T(x) = x + \delta \in \underset{\delta:\|\delta\|_p \leq \epsilon}{\operatorname{argmin}}\ \ell_{\mathcal{F}}(x + \delta, y_t)$.*

In this definition, $\|\delta\|_p$ represents the L_p norm of δ. In untargeted attacks, provided the prediction on an adversarial instance differs from the ground truth, the attack is considered successful. In targeted attacks, the attack succeeds only when the adversarial instance can be mis-recognized as the specific adversarial target y_t.

How do we formally define that an attack is effective? We define the attack effectiveness for both targeted and untargeted attacks based on the statistical confidence.

Definition 2 ($(\alpha, \mathcal{F})$-effective attack)**.** *Consider a input $x \in \mathcal{X}$ with true label y. An attack is $(\alpha, \mathcal{F})$- effective in both untargeted and targeted (with class target y_t) scenarios if:*

- *Untargeted:* $\Pr(\mathcal{F}(\mathcal{A}_U(x)) \neq y) \geq 1 - \alpha$.
- *Targeted:* $\Pr(\mathcal{F}(\mathcal{A}_T(x)) = y_t) \geq 1 - \alpha$.

2.2. Bounding Adversarial Transferability

Given two models $\mathcal{F}$ and $\mathcal{G}$, what are the characteristics of $\mathcal{F}$ and $\mathcal{G}$ that have impact on transferability under a given attack strategy? In this section, we obtain new theoretical insights for formalizing this intuition, which later inspires a practical algorithm for minimizing transferability.

2.2.1. Transferability and Preliminaries

Before we present our result, we formally define transferability and relevant characteristics of models. We then connect transferability with these properties.

Definition 3 (Transferability)**.** *Consider an adversarial instance $\mathcal{A}_U(x)$ or $\mathcal{A}_T(x)$ constructed against a surrogate model $\mathcal{F}$. The transferability T_r between $\mathcal{F}$ and a target model $\mathcal{G}$ is defined as follows (for adversarial target y_t):*

- *Untargeted:* $T_r(\mathcal{F}, \mathcal{G}, x) = \mathbb{1}[\mathcal{F}(x) = \mathcal{G}(x) = y \wedge \mathcal{F}(\mathcal{A}_U(x)) \neq y \ \wedge\ \mathcal{G}(\mathcal{A}_U(x)) \neq y]$.
- *Targeted:* $T_r(\mathcal{F}, \mathcal{G}, x, y_t) = \mathbb{1}[\mathcal{F}(x) = \mathcal{G}(x) = y \ \wedge\ \mathcal{F}(\mathcal{A}_T(x)) = \mathcal{G}(\mathcal{A}_T(x)) = y_t]$.

Definition 4 (Lower Loss Gradient Similarity)**.** *The lower loss gradient similarity $\underline{\mathcal{S}}$ between two differentiable loss functions $\ell_{\mathcal{F}}$ and $\ell_{\mathcal{G}}$ is defined as:*

$$\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}}) = \inf_{x' \in \mathcal{X}, y \in \mathcal{Y}} \frac{\nabla_x \ell_{\mathcal{F}}(x', y) \cdot \nabla_x \ell_{\mathcal{G}}(x', y)}{\|\nabla_x \ell_{\mathcal{F}}(x', y)\|_2 \cdot \|\nabla_x \ell_{\mathcal{G}}(x', y)\|_2}.$$

Definition 5 (Upper Loss Gradient Similarity)**.** *The upper loss gradient similarity $\overline{\mathcal{S}}$ between two differentiable loss*

functions $\ell_{\mathcal{F}}$ and $\ell_{\mathcal{G}}$ is defined as:

$$\overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}}) = \sup_{x' \in \mathcal{X}, y \in \mathcal{Y}} \frac{\nabla_x \ell_{\mathcal{F}}(x', y) \cdot \nabla_x \ell_{\mathcal{G}}(x', y)}{\|\nabla_x \ell_{\mathcal{F}}(x', y)\|_2 \cdot \|\nabla_x \ell_{\mathcal{G}}(x', y)\|_2}.$$

Definition 6 (Model Smoothness)**.** *The model $\mathcal{F}$ is said to be β-smooth if β satisfies:*

$$\beta \geq \sup_{x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}} \frac{\|\nabla_x \ell_{\mathcal{F}}(x_1, y) - \nabla_x \ell_{\mathcal{F}}(x_2, y)\|_2}{\|x_1 - x_2\|_2}.$$

2.2.2. Lower Bound of Transferability

Next we show that with both loss gradient similarity and model smoothness it is possible to lower bound the transferability between low risk classifiers.

Theorem 1 (Lower Bound on Untargeted Attack Transferability)**.** *Consider an instance $x \in \mathbb{R}^n$ with true label y and adversarial target y_t. An $(\alpha, \mathcal{F})$-effective (untargeted) adversarial attack $x^{\mathcal{A}} = \mathcal{A}_U(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with probability*

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x) = 1\right) \geq (1 - \alpha) - (\eta_{\mathcal{F}} + \eta_{\mathcal{G}}) - \frac{\epsilon(1+\alpha) - c_{\mathcal{F}}(1-\alpha)}{\epsilon - c_{\mathcal{G}}} - \frac{\epsilon(1-\alpha)}{\epsilon - c_{\mathcal{G}}} \sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})},$$

where

$$c_{\mathcal{F}} = \min_{(x,y) \in \mathcal{M}} \frac{\min_{y'; y' \neq y} \ell_{\mathcal{F}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{F}}(x, y) - \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{F}}(x, y)\|_2},$$

$$c_{\mathcal{G}} = \max_{(x,y) \in \mathcal{M}} \frac{\min_{y'; y' \neq y} \ell_{\mathcal{G}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{G}}(x, y) + \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{G}}(x, y)\|_2}.$$

Here $\eta_{\mathcal{F}}$, $\eta_{\mathcal{G}}$ are the risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively.*

Due to the page limit we omit the lower bound for target attack to the appendix. Though it seems pessimistic given the lower bound of transferability, we would ask another question: *Is it possible to upper bound the transferability given additional constraints?*

2.2.3. Upper Bound of Transferability For Smoothed Classifiers

Suppose we have two models $\mathcal{F}$ and $\mathcal{G}$, then for any targeted adversarial attack, we can upper bound the transferability of the two models by constraining their gradient similarity, model smoothness and respective risks.

Theorem 2 (Upper Bound on Untargeted Attack Transferability)**.** *Consider an instance $x \in \mathbb{R}^n$ with true label y and adversarial target y_t. Assume both model $\mathcal{F}$ and $\mathcal{G}$ are β-smooth with gradients bounded by B. An $(\alpha, \mathcal{F})$-effective (untargeted) adversarial attack $x^{\mathcal{A}} = \mathcal{A}_U(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with probability*

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) \leq \frac{\xi_{\mathcal{F}} + \xi_{\mathcal{G}}}{\ell_{\min} - \epsilon B \left(1 + \sqrt{\frac{1 + \overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}{2}}\right) - \beta\epsilon^2},$$

where $\ell_{\min} = \min_{x, y': (x,y) \in \mathcal{M}, y' \neq y} (\ell_{\mathcal{F}}(x, y'), \ell_{\mathcal{G}}(x, y'))$.

Here $\xi_{\mathcal{F}}$ and $\xi_{\mathcal{G}}$ are the empirical risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively, defined relative to a differentiable loss.*

Due to the page limit we omit the upper bound for target attack to the appendix. From both lower bound and upper bound, we observe that it is possible for us to restrict the loss gradient similarity between models to enforce model orthogonality, and at the same time regularize the model smoothness to reduce transferability whose intuition has been shown in Figure 1.

3. Improving Ensemble Robustness via Transferability Minimization

Built upon our theoretical analysis on transferability, we propose a simple yet effective robust ensemble training approach to reduce the transferability among base models by enforcing the *model smoothness* and *minimizing model loss gradient similarity*. we develop *Transferability Reduced Smooth* ensemble (TRS) to train a robust ensemble in which the transferability between base models is minimized.

The training algorithm of TRS-ensemble is shown in Algorithm 1 in Appendix E. The TRS loss is calculated as

$$L_{TRS}(x, y) = \frac{x \otimes y}{\sqrt{x \otimes x \cdot y \otimes y}} + \lambda_1 \|x\| + \lambda_2 \|y\|.$$

Here $\otimes$ denotes the operator that calculates the Frobenius norm [6] of the outer product of two vectors. Algorithm 1 describes training on one dataset for one epoch, and in practice we update by mini-batches.

We leverage the regularizer of the loss gradient to constrain model smoothness. From Theorems 2 we can see that when the models are forced to be smooth, the lower and upper bound of transferability would be tighter and therefore as long as the model loss gradient similarity is minimized, the transferability will be largely constrained. We validate this observation empirically in Section 4.

4. Experimental Evaluation

In this section we provide experimental evaluation of the proposed TRS-ensemble comparing with baselines to assess robustness against different attacks and transferability. We show that: (1) on benign instances, the trained TRS-ensemble achieves similar performance compared with vanilla models; (2) against different adversarial whitebox attacks, TRS-ensemble outperforms other baselines in terms of robustness; (3) we specifically analyze the transferability among base models within the TRS-ensemble and show that intra-ensemble transferability is indeed reduced significantly, across a range of attacks.

4.1. Experimental Setup

we conduct experiments on MNIST and CIFAR-10 [20; 18]. In our experiments, we employ Resnet20 network [13] as the ensembke base model and use the combined loss for TRS and Ensemble CrossEntropy (ECE) to train our TRS-ensemble. The weight for TRS loss is an adjusting constant

Table 1: Robustness of different approaches against various white-box attacks.

MNIST	para.	CosSim	CKAE	AdaBoost	GradientBoost	$ADP_{2,0.5}$	**TRS**
FGSM	$\epsilon = 0.1$	91.1	91.7	92.2	93.4	**96.3**	95.0
	$\epsilon = 0.2$	60.1	52.7	45.3	48.5	52.8	**65.7**
BIM	$\epsilon = 0.1$	86.2	87.2	82.0	84.3	88.5	**91.7**
	$\epsilon = 0.15$	73.4	70.1	69.5	70.2	73.6	**74.4**
PGD	$\epsilon = 0.1$	83.2	80.1	74.3	78.2	82.8	**93.3**
	$\epsilon = 0.15$	51.6	45.2	36.2	38.1	41.0	**74.0**
MIM	$\epsilon = 0.1$	87.2	90.4	85.1	85.3	92.0	**92.5**
	$\epsilon = 0.15$	73.6	73.2	72.2	71.9	**77.5**	75.4
CW	$c = 0.1$	93.2	90.1	92.8	94.1	**97.3**	95.2
	$c = 10$	34.2	40.1	30.2	31.3	23.8	**43.2**

c. We set $\lambda_1 = \lambda_2 = 0.1, c = 2.0$ for MNIST and $\lambda_1 = \lambda_2 = 0.01, c = 5.0$ for CIFAR-10. (We evaluated other parameters and observe similar results in Appendix.)

Adversarial attacks. We considered the following adversarial attacks. FGSM [10], which generates adversarial example $x^{\mathcal{A}} = x + \delta$ by assigning $\delta = \epsilon \cdot \text{sgn}(\nabla_x \ell(x, y))$. BIM [24] is an iterative FGSM by adding perturbations step by step: $x_i = \text{clip}(x_{i-1} + \epsilon/r \cdot \nabla_x \ell(x_{i-1}, y))$, while clip denotes back projection to the ϵ perturbation range. MIM [8] can be regarded as a variant of BIM by utilizing the momentum in the gradient backward as $g_{i+1} = \mu g_i + \frac{\ell(x_i, y)}{\|\nabla_x \ell(x, y)\|_1}$ and $x_i = \text{clip}(x_{i-1} + \epsilon/r \times g_i)$ in iteration i. PGD [24] starts by searching for a randomly perturbed instance within the ℓ_p ball and moves along the gradient direction until convergence. JSMA [27] is a greedy attack algorithm that perturbs the pixels with high values on the saliency map at each iteration. C&W [3] solves the optimization problem $x^{\mathcal{A}} := \min_{x'} \|x' - x\|_2^2 + c \cdot f(x', y)$, where c is a constant to balance the perturbation scale and attack success rate. EAD [5] follows the same optimization setting by solving $x^{\mathcal{A}} := \min_{x'} \|x' - x\|_2^2 + \beta \|x' - x\|_1 + c \cdot f(x', y)$. Other detailed settings will be omitted to Appendix F and we will also open source our code for further comparison.

4.2. Baseline Methods to Reduce Transferability

In this section, we consider other baseline approaches and show that TRS-ensemble outperforms others significantly in terms of improving the learning robustness. We briefly introduce them and put more details to Appendix. CosSim considers the ensemble which only minimizes the gradient similarity between base models. CKAE [17] develops ensemble based on CKA measurement proved to be effective to measure the orthogonality between representations. AdaBoost [12] the final prediction will be a weighted average of all the weak learners where higher weight is placed on stronger learners. GradientBoost [15] identifies weaker learners based on gradient information. ADP (Adaptive Diversity Promoting) [25] is proposed recently as a regularizer to reduce transferability between base models within an ensemble to improve robustness and become the current state-of-the-art ensemble method.

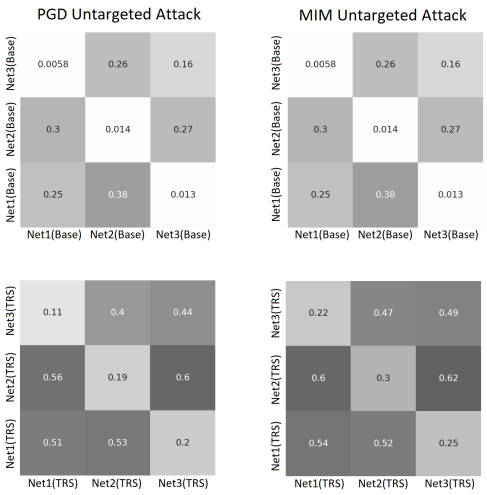

Figure 2: Transferability analysis for vanilla and base models from TRS-ensemble on MNIST ($\epsilon = 0.2$). Each cell (i, j) shows the *classification accuracy* of ith model on the adversarial examples generated against the jth model. The first row presents vanilla models, and second row the base models from TRS-ensemble. It shows that the base models from TRS-ensemble achieves higher accuracy indicating lower transferability.

4.3. Experimental Results

We first apply a whitebox attack against different ensembles (each contains 3 base models for fair comparison with state-of-the-art method ADP). Table 1 presents the robustness (accuracy) of different ensemble methods against a range of whitebox attacks. It is shown that the proposed TRS outperforms other baselines including the state-of-the-art ADP approach in most cases under different attacks with various perturbation budgets. (The benign accuracy of all methods are above 99.3%.)

To further understand the effects that TRS has on model transferability, we measure the inter model transferability given different attacks as shown in Figure 2. Due to space limitations we only show results based on two attacks, with more results confirming these observations in Appendix Figure 3, 4. We measure the transferability of TRS-ensemble under attacks as shown in Figure 2 and find that base models from TRS achieve higher accuracy (robustness), indicating lower transferability. In addition, we evaluate TRS-ensemble against an Intermediate Level Attack [14] which aims to improve transferability of adversarial examples, and find that its robustness still remains around 96.22% with $\epsilon = 0.1$ on MNIST and 78.82% with $\epsilon = 0.01$ on CIFAR-10. Detailed results are shown in Appendix Table 2, 6.

5. Conclusion

In this paper we deliver deep understanding of adversarial transferability theoretically. We propose an ensemble training approach which reduces transferability by promoting model smoothness and reducing loss gradient similarity.

References

[1] Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[2] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.

[3] Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

[4] Chambolle, A. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1):89–97, 2004.

[5] Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C.-J. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[6] Custódio, A. L., Rocha, H., and Vicente, L. N. Incorporating minimum frobenius norm models in direct search. *Computational Optimization and Applications*, 46(2):265–278, 2010.

[7] Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., and Roli, F. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 321–338, 2019.

[8] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

[9] Fawzi, A., Fawzi, O., and Frossard, P. Analysis of classifiers' robustness to adversarial perturbations. *arXiv preprint arXiv:1502.02590*, 2015.

[10] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[12] Hastie, T., Rosset, S., Zhu, J., and Zou, H. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[13] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., and Lim, S.-N. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4733–4742, 2019.

[15] Jiao, F., Xu, J., Yu, L., and Schuurmans, D. Protein fold recognition using the gradient boost algorithm. In *Computational Systems Bioinformatics*, pp. 43–53. World Scientific, 2006.

[16] Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019.

[18] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

[19] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[20] LeCun, Y. and Cortes, C. The MNIST database of handwritten digits. 1998.

[21] Li, B. and Vorobeychik, Y. Feature cross-substitution in adversarial classification. In *Advances in Neural Information Processing Systems*, pp. 2087–2095, 2014.

[22] Li, B. and Vorobeychik, Y. Scalable optimization of randomized operational decisions in adversarial classification settings. In *AISTATS*, 2015.

[23] Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[24] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846*, 2019.

[26] Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[27] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

[28] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z., and Swami, A. Practical black-box attacks against deep learning systems using adversarial examples. In *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, 2017.

[29] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[30] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[31] Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

A. Discussion: Beyond ℓ_p-Attack

Besides the widely used ℓ_p norm based adversarial examples, here we plan to extend our understanding to the distribution distance analysis.

Definition 7 (Total variation distance; [4])**.** *For two probability distributions P_x and $P_{\mathcal{A}(x)}$ on $\mathcal{X}$, the total variation distance between them is defined as*

$$\|P_x - P_{\mathcal{A}(x)}\|_{TV} = \sup_{C\subset\mathcal{X}} |P_x(C) - P_{\mathcal{A}(x)}(C)|.$$

Informally, the total variation distance measures the largest change in probability over all events. For discrete probability distributions, the total variation distance is just the ℓ_1 distance between the vectors in the probability simplex representing the two distributions.

Definition 8. *Given $\rho \in (0,1)$, an attack strategy $\mathcal{A}(\cdot)$ is called* ρ-conservative[1]*, if for $x \sim P_x$, $\|P_x - P_{\mathcal{A}(x)}\|_{TV} \leq \rho$.*

This definition formalizes the general objective of generating adversarial examples against deep neural networks: attack samples are likely to be observed, while they do not themselves arouse suspicion.

Lemma 3. *Let $f, g : \mathcal{X} \to \mathcal{Y}$ be classifiers, $\delta, \rho, \epsilon \in (0,1)$ be constants, and $\mathcal{A}(\cdot)$ be an attack strategy. Suppose that $\mathcal{A}(\cdot)$ is ρ-conservative and f, g have risk at most ϵ. Then $\Pr(\mathcal{F}(\mathcal{A}(x)) \neq \mathcal{G}(\mathcal{A}(x))) \leq 2\epsilon + \rho$ for a given random instance $x \sim P_x$.*

Remark. This result provides theoretical backing for the intuition that the boundaries of low risk classifiers under certain dense data distribution are close [31]. It considers two classifiers that have risk at most ϵ, which indicates their boundaries are close for benign data. It then shows that their boundaries are also close for the perturbed data as long as the attack strategy satisfies a conservative condition which constrains the drift in distribution between the benign and adversarial data.

Proof of Lemma 3. Given $\mathcal{A}(\cdot)$ is ρ-*covert*, by Definition 8 we know

$$\begin{aligned}
&|P_X[f(\mathcal{A}(x)) = g(\mathcal{A}(x))] - P_X[f(x) = g(x)]| \\
&= \left|P_{\mathcal{A}(X)}[f(x) = g(x)] - P_X[f(x) = g(x)]\right| \\
&\leq \rho.
\end{aligned}$$

Therefore, we have

$$\Pr(f(\mathcal{A}(x)) = g(\mathcal{A}(x))) \geq \Pr(f(x) = g(x)) - \rho.$$

From the low-risk conditions, the classifiers agree w.h.p.

$$\begin{aligned}
&\Pr(f(\mathcal{A}(x)) \neq g(\mathcal{A}(x))) \\
&\leq \Pr(f(x) \neq g(x)) + \rho \\
&\leq 1 - \Pr(f(x) = y, g(x) = y) + \rho \\
&\leq 1 - (1 - \Pr(f(x) \neq y) - \Pr(g(x) \neq y)) + \rho \\
&= \epsilon + \epsilon + \rho\ ,\ ^{2} \\
&\leq 2\epsilon + \rho\ ,
\end{aligned}$$

where the third inequality follows from the union bound.[3] □

Theorem 4. *Let $\mathcal{F}, \mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be classifiers ($\mathcal{Y} \in \{-1, 1\}$), $\delta, \rho, \epsilon \in (0,1)$ be constants, and $\mathcal{A}(\cdot)$ an attack strategy. Suppose that $\mathcal{A}(\cdot)$ is ρ-conservative and $\mathcal{F}, \mathcal{G}$ have risk at most ϵ. Given random instance $x \in \mathcal{X}$, if $\mathcal{A}(\cdot)$ is $(\delta, \mathcal{F})$-effective, then it is also $(\delta + 4\epsilon + \rho, \mathcal{G})$-effective.*

This result formalizes the intuition that low-risk classifiers possess close decision boundaries in high-probability regions. In such settings, an attack strategy that successfully attacks one classifier would have high probability to mislead the other. This theorem explains why we should expect successful transferability in practice: defenders will naturally prefer low-risk binary classifiers. This desirable quality of classifiers is a potential liability.

[1] We use the total variation distance as it is a natural way to measure distances between distributions. Other notions of distance may also be applied.

[2] Here we assume y is the ground truth label.

[3] Recall that for arbitrary events $A_1, \ldots, A_n$, the union bound implies $P\left(\bigcap_{i=1}^n A_i\right) \geq 1 - \sum_{i=1}^n P\left(\overline{A_i}\right)$.

Proof of Theorem 4. From Lemma 3, and the union bound we have

$$\begin{aligned}
&\Pr\left(g(x) \neq g(\mathcal{A}\left(x\right))\right) \\
&\geq \Pr\left(f(x) \neq f(\mathcal{A}\left(x\right)), g(x) = f(x), g(\mathcal{A}\left(x\right)) = f(\mathcal{A}\left(x\right))\right) \\
&\geq 1 - \Pr\left(f(x) = f(\mathcal{A}\left(X\right))\right) - \Pr\left(g(x) \neq f(x)\right) - \Pr\left(g(\mathcal{A}\left(x\right)) \neq f(\mathcal{A}\left(x\right))\right) \\
&\geq 1 - \delta - 4\epsilon - \rho \ ,
\end{aligned}$$

as claimed. □

B. Proof of Transferability Lower Bound (Theorem 1)

Here we present the proof of Theorem 1 first stated in Section 2.2.2.

The following lemma is used in the proof.

Lemma 5. *For arbitrary vector δ, x, y, when $\|\delta\|_2 \leq \epsilon$, x and y are unit vectors, i.e., $\|x\|_2 = \|y\|_2 = 1$, $\cos\langle x, y\rangle = \dfrac{x \cdot y}{\|x\|_2 \cdot \|y\|_2} = m$. Let c denote any real number. Then*

$$\delta \cdot y > c + \epsilon\sqrt{2 - 2m} \;\;\Rightarrow\;\; \delta \cdot x > c.$$

Proof. $\delta \cdot x = \delta \cdot y + \delta \cdot (x - y) > c + \epsilon\sqrt{2 - 2m} + \delta \cdot (x - y)$. By law of cosines, $\delta \cdot (x - y) \geq -\epsilon\sqrt{2 - 2\cos\langle x, y\rangle} = -\epsilon\sqrt{2 - 2m}$. Hence, $\delta \cdot x > c$. □

Theorem (Lower Bound on Targeted Attack Transferability)**.** *Consider an instance $x \in \mathbb{R}^n$ with true label y and adversarial target y_t. Assume both model $\mathcal{F}$ and $\mathcal{G}$ are β-smooth. An $(\alpha, \mathcal{F})$-effective (targeted) attack $x^{\mathcal{A}} = \mathcal{A}_T(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with bounded probability*

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) \geq (1 - \alpha) - (\eta_{\mathcal{F}} + \eta_{\mathcal{G}}) - \frac{\epsilon(1 + \alpha) + c_{\mathcal{F}}(1 - \alpha)}{c_{\mathcal{G}} + \epsilon} - \frac{\epsilon(1 - \alpha)}{c_{\mathcal{G}} + \epsilon}\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})},$$

$$\textit{where}\quad c_{\mathcal{F}} = \max_{x \in \mathcal{X}} \frac{\min_y \ell_{\mathcal{F}}(x^{\mathcal{A}}, y) - \ell_{\mathcal{F}}(x, y_t) + \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2},\; c_{\mathcal{G}} = \min_{x \in \mathcal{X}} \frac{\min_y \ell_{\mathcal{G}}(x^{\mathcal{A}}, y) - \ell_{\mathcal{G}}(x, y_t) - \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2},$$
$$\eta_{\mathcal{F}} = \Pr\left(\mathcal{F}(x) \neq y\right),\; \eta_{\mathcal{G}} = \Pr\left(\mathcal{G}(x) \neq y\right).$$

Here $\eta_{\mathcal{F}}, \eta_{\mathcal{G}}$ are the risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively.*

Proof. Define auxiliary function $f, g : \mathcal{X} \mapsto \mathbb{R}$ such that

$$f(x) = \frac{\min_{y \in \mathcal{Y}} \ell_{\mathcal{F}}(x^{\mathcal{A}}, y) - \ell_{\mathcal{F}}(x, y_t) + \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2},\; g(x) = \frac{\min_{y \in \mathcal{Y}} \ell_{\mathcal{G}}(x^{\mathcal{A}}, y) - \ell_{\mathcal{G}}(x, y_t) - \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2}.$$

Note that $c_{\mathcal{F}} = \max_{x \in \mathcal{X}} f(x)$ and $c_{\mathcal{G}} = \min_{x \in \mathcal{X}} g(x)$.

The transferability of concern satisfies:

$$\begin{aligned}
\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) &= \Pr\left(\mathcal{F}(x) = y \cap \mathcal{G}(x) = y \cap \mathcal{F}(x^{\mathcal{A}}) = y_t \cap \mathcal{G}(x^{\mathcal{A}}) = y_t\right) && \text{(B.1)}\\
&\geq 1 - \Pr\left(\mathcal{F}(x) \neq y\right) - \Pr\left(\mathcal{G}(x) \neq y\right) \\
&\quad - \Pr\left(\mathcal{F}(x^{\mathcal{A}}) \neq y_t\right) - \Pr\left(\mathcal{G}(x^{\mathcal{A}}) \neq y_t\right) && \text{(B.2)}\\
&\geq 1 - \eta_{\mathcal{F}} - \eta_{\mathcal{G}} - \alpha - \Pr\left(\mathcal{G}(x^{\mathcal{A}}) \neq y_t\right). && \text{(B.3)}
\end{aligned}$$

Eq. B.1 follows the definition (Definition 3). Eq. B.1 to Eq. B.2 follows from the union bound. From Eq. B.1 to Eq. B.2 definition of model risk and definition of adversarial effectiveness (Definition 2) are applied.

Now consider $\Pr\left(\mathcal{F}(x^{\mathcal{A}}) \neq y_t\right)$ and $\Pr\left(\mathcal{G}(x^{\mathcal{A}}) \neq y_t\right)$. Given that model predicts the label for which $\ell_{\mathcal{F}}$ is minimized, $\mathcal{F}(x^{\mathcal{A}}) \neq y_t \iff \ell_{\mathcal{F}}(x + \delta, y_t) > \min_y \ell_{\mathcal{F}}(x + \delta, y)$. Similarly, $\mathcal{G}(x^{\mathcal{A}}) \neq y_t \iff \ell_{\mathcal{G}}(x + \delta, y_t) > \min_y \ell_{\mathcal{G}}(x + \delta, y)$.

Following Taylor's Theorem with Lagrange remainder, we have

$$\ell_{\mathcal{F}}(x + \delta, y_t) = \ell_{\mathcal{F}}(x, y_t) + \delta\nabla_x \ell_{\mathcal{F}}(x, y_t) + \frac{1}{2}\xi^\top \mathbf{H}_{\mathcal{F}}\xi, \tag{B.4}$$

$$\ell_{\mathcal{G}}(x + \delta, y_t) = \ell_{\mathcal{G}}(x, y_t) + \delta\nabla_x \ell_{\mathcal{G}}(x, y_t) + \frac{1}{2}\xi^\top \mathbf{H}_{\mathcal{G}}\xi. \tag{B.5}$$

In Eq. B.4 and Eq. B.5, $\xi = k\delta$ for some $k \in [0, 1]$. $\mathbf{H}_{\mathcal{F}}$ and $\mathbf{H}_{\mathcal{G}}$ are Hessian matrices of $\ell_{\mathcal{F}}$ and $\ell_{\mathcal{G}}$ respectively. Since $\ell_{\mathcal{F}}(x + \delta, y_t)$ and $\ell_{\mathcal{G}}(x + \delta, y_t)$ are β-smooth, the maximum eigenvalues of $\mathbf{H}_{\mathcal{F}}$ and $\mathbf{H}_{\mathcal{G}}$ are bounded by β, As the result,

$|\xi^\top \mathbf{H}_{\mathcal{F}} \xi| \leq \beta \cdot \|\xi\|_2^2 \leq \beta\epsilon^2$. Applying them to Eq. B.4 and Eq. B.5, we thus have

$$\ell_{\mathcal{F}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{F}}(x, y_t) - \frac{1}{2}\beta\epsilon^2 \leq \quad \ell_{\mathcal{F}}(x+\delta, y_t) \leq \quad \ell_{\mathcal{F}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{F}}(x, y_t) + \frac{1}{2}\beta\epsilon^2, \tag{B.6}$$

$$\ell_{\mathcal{G}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{G}}(x, y_t) - \frac{1}{2}\beta\epsilon^2 \leq \quad \ell_{\mathcal{G}}(x+\delta, y_t) \leq \quad \ell_{\mathcal{G}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{G}}(x, y_t) + \frac{1}{2}\beta\epsilon^2. \tag{B.7}$$

Apply left hand side of Eq. B.6 to $\Pr\left(\mathcal{F}(x^{\mathcal{A}}) \neq y_t\right) \leq \alpha$ (from Definition 2):

$$\Pr\left(\mathcal{F}(x^{\mathcal{A}}) \neq y_t\right) \tag{B.8}$$

$$= \Pr\left(\ell_{\mathcal{F}}(x+\delta, y_t) > \min_y \ell_{\mathcal{F}}(x+\delta, y)\right) \tag{B.9}$$

$$\geq \Pr\left(\ell_{\mathcal{F}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{F}}(x, y_t) - \frac{1}{2}\beta\epsilon^2 > \min_y \ell_{\mathcal{F}}(x+\delta, y)\right) \tag{B.10}$$

$$= \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{F}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2} > f(x)\right), \tag{B.11}$$

$$\implies \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{F}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2} > f(x)\right) \leq \alpha. \tag{B.12}$$

Similarly, we apply right hand side of Eq. B.7 to $\Pr\left(\mathcal{G}(x^{\mathcal{A}}) = y_t\right)$:

$$\Pr\left(\mathcal{G}(x^{\mathcal{A}}) \neq y_t\right) \tag{B.13}$$

$$= \Pr\left(\ell_{\mathcal{G}}(x+\delta, y_t) > \min_y \ell_{\mathcal{G}}(x+\delta, y)\right) \tag{B.14}$$

$$\leq \Pr\left(\ell_{\mathcal{G}}(x, y_t) + \delta \nabla_x \ell_{\mathcal{G}}(x, y_t) + \frac{1}{2}\beta\epsilon^2 > \min_y \ell_{\mathcal{G}}(x+\delta, y)\right) \tag{B.15}$$

$$= \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > g(x)\right). \tag{B.16}$$

Knowing that $\|\delta\|_2 \leq \epsilon$, from Lemma 5 we have

$$\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > f(x) + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})} \tag{B.17}$$

$$\implies \delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > f(x) + \epsilon\sqrt{2 - 2\cos\langle \nabla_x \ell_{\mathcal{F}}(x, y_t), \nabla_x \ell_{\mathcal{G}}(x, y_t)\rangle} \tag{B.18}$$

$$\implies \delta \cdot \frac{\nabla_x \ell_{\mathcal{F}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2} > f(x). \tag{B.19}$$

From Eq. B.17 to Eq. B.18, the infimum in definition of $\underline{\mathcal{S}}$ (Definition 4) indicates that

$$\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}}) \leq \cos\langle \nabla_x \ell_{\mathcal{F}}(x, y_t), \nabla_x \ell_{\mathcal{G}}(x, y_t)\rangle.$$

Hence,

$$f(x) + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})} \geq f(x) + \epsilon\sqrt{2 - 2\cos\langle \nabla_x \ell_{\mathcal{F}}(x, y_t), \nabla_x \ell_{\mathcal{G}}(x, y_t)\rangle}.$$

Eq. B.18 to Eq. B.19 directly uses Lemma 5. As the result,

$$\Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > f(x) + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right) \leq \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{F}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2} > f(x)\right) \leq \alpha. \tag{B.20}$$

Note that $f(x) \leq c_{\mathcal{F}}$, we have

$$\Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right) \leq \alpha. \tag{B.21}$$

Now we consider the maximum expectation of $\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2}$. Its maximum is $\max\|\delta\|_2 = \epsilon$. Therefore, its expectation is bounded:

$$\mathbb{E}\left[\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2}\right] \leq \epsilon \cdot \alpha + \left(c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right)(1-\alpha).$$

Now applying Markov's inequality, we get

$$\Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > c_{\mathcal{G}}\right) \leq \frac{\epsilon \cdot \alpha + \left(c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right)(1-\alpha) + \epsilon}{c_{\mathcal{G}} + \epsilon} \tag{B.22}$$

$$= \frac{\epsilon(1+\alpha) + \left(c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right)(1-\alpha)}{c_{\mathcal{G}} + \epsilon}. \tag{B.23}$$

Since $g(x) \geq c_{\mathcal{G}}$,

$$\Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > g(x)\right) \leq \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > c_{\mathcal{G}}\right) \tag{B.24}$$

$$\leq \frac{\epsilon(1+\alpha) + \left(c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right)(1-\alpha)}{c_{\mathcal{G}} + \epsilon}. \tag{B.25}$$

Combine with Eq. B.19, finally,

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) \geq 1 - \eta_{\mathcal{F}} - \eta_{\mathcal{G}} - \alpha - \Pr\left(\mathcal{G}(x^{\mathcal{A}}) \neq y_t\right) \tag{B.26}$$

$$\geq 1 - \eta_{\mathcal{F}} - \eta_{\mathcal{G}} - \alpha - \Pr\left(\delta \cdot \frac{\nabla_x \ell_{\mathcal{G}}(x, y_t)}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2} > g(x)\right) \tag{B.27}$$

$$\geq 1 - \eta_{\mathcal{F}} - \eta_{\mathcal{G}} - \alpha - \frac{\epsilon(1+\alpha) + \left(c_{\mathcal{F}} + \epsilon\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}\right)(1-\alpha)}{c_{\mathcal{G}} + \epsilon} \tag{B.28}$$

$$= (1-\alpha) - (\eta_{\mathcal{F}} + \eta_{\mathcal{G}}) - \frac{\epsilon(1+\alpha) + c_{\mathcal{F}}(1-\alpha)}{c_{\mathcal{G}} + \epsilon} - \frac{\epsilon(1-\alpha)}{c_{\mathcal{G}} + \epsilon}\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}. \tag{B.29}$$

Eq. B.26 to Eq. B.27 follows Eq. B.16. □

Theorem (Lower Bound on Untargeted Attack Transferability)**.** *Under the same setting as Theorem B. An $(\alpha, \mathcal{F})$-effective (untargeted) adversarial attack $x^{\mathcal{A}} = \mathcal{A}_U(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with probability*

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x) = 1\right) \geq (1-\alpha) - (\eta_{\mathcal{F}} + \eta_{\mathcal{G}}) - \frac{\epsilon(1+\alpha) - c_{\mathcal{F}}(1-\alpha)}{\epsilon - c_{\mathcal{G}}} - \frac{\epsilon(1-\alpha)}{\epsilon - c_{\mathcal{G}}}\sqrt{2 - 2\underline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})},$$

where $c_{\mathcal{F}} = \min_{(x,y)\in\mathcal{M}} \frac{\min_{y';y'\neq y} \ell_{\mathcal{F}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{F}}(x, y) - \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{F}}(x, y)\|_2}$, $c_{\mathcal{G}} = \max_{(x,y)\in\mathcal{M}} \frac{\min_{y';y'\neq y} \ell_{\mathcal{G}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{G}}(x, y) + \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{G}}(x, y)\|_2}$,

$\eta_{\mathcal{F}} = \Pr\left(\mathcal{F}(x) \neq y\right)$, $\eta_{\mathcal{G}} = \Pr\left(\mathcal{G}(x) \neq y\right)$.

Here $\eta_{\mathcal{F}}$ and $\eta_{\mathcal{G}}$ are the risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively.*

Proof. Define auxiliary function $f, g : \mathcal{M} \to \mathbb{R}$ such that

$$f(x, y) = \frac{\min_{y';y'\neq y} \ell_{\mathcal{F}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{F}}(x, y) - \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{F}}(x, y_t)\|_2}, \; g(x, y) = \frac{\min_{y';y'\neq y} \ell_{\mathcal{G}}(x^{\mathcal{A}}, y') - \ell_{\mathcal{G}}(x, y) + \beta\epsilon^2/2}{\|\nabla_x \ell_{\mathcal{G}}(x, y_t)\|_2}.$$

Note that $c_{\mathcal{F}} = \min_{(x,y)\in\mathcal{M}} f(x, y)$ and $c_{\mathcal{G}} = \max_{(x,y)\in\mathcal{M}} g(x, y)$.

The proof is similar to that of Theorem B.

$$\Pr\left(T_r(\mathcal{F},\mathcal{G},x)=1\right)=\Pr\left(\mathcal{F}(x)=y\cap\mathcal{G}(x)=y\cap\mathcal{F}(x^{\mathcal{A}})\neq y\cap\mathcal{G}(x^{\mathcal{A}})\neq y\right) \tag{B.30}$$

$$\geq 1-\Pr\left(\mathcal{F}(x)\neq y\right)-\Pr\left(\mathcal{G}(x)\neq y\right)-\Pr\left(\mathcal{F}(x^{\mathcal{A}})=y\right)-\Pr\left(\mathcal{G}(x^{\mathcal{A}})=y\right) \tag{B.31}$$

$$=1-\eta_{\mathcal{F}}-\eta_{\mathcal{G}}-\alpha-\Pr\left(\mathcal{G}(x^{\mathcal{A}})=y\right). \tag{B.32}$$

From Taylor's Theorem and Lemma 5, we observe that

$$\Pr\left(\mathcal{G}(x^{\mathcal{A}})=y\right)\leq\Pr\left(\delta\cdot\frac{\nabla_x\ell_{\mathcal{G}}(x,y)}{\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2}<c_{\mathcal{G}}\right), \tag{B.33}$$

$$\Pr\left(\delta\cdot\frac{\nabla_x\ell_{\mathcal{G}}(x,y)}{\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2}<c_{\mathcal{F}}-\epsilon\sqrt{2-2\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})}\right)\leq\Pr\left(\mathcal{F}(x^{\mathcal{A}})=y\right)=\alpha. \tag{B.34}$$

According to Markov's inequality, Eq. B.34 implies that

$$\Pr\left(\delta\cdot\frac{\nabla_x\ell_{\mathcal{G}}(x,y)}{\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2}<c_{\mathcal{G}}\right)\leq\frac{\epsilon(1+\alpha)-\left(c_{\mathcal{F}}-\epsilon\sqrt{2-2\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})}\right)(1-\alpha)}{\epsilon-c_{\mathcal{G}}}. \tag{B.35}$$

Combine Eq. B.33 with Eq. B.35, then pump into Eq. B.32,

$$\Pr\left(T_r(\mathcal{F},\mathcal{G},x)=1\right)\geq 1-\eta_{\mathcal{F}}-\eta_{\mathcal{G}}-\alpha-\Pr\left(\mathcal{G}(x^{\mathcal{A}})=y\right) \tag{B.36}$$

$$\geq(1-\alpha)-(\eta_{\mathcal{F}}+\eta_{\mathcal{G}})-\frac{\epsilon(1+\alpha)-\left(c_{\mathcal{F}}-\epsilon\sqrt{2-2\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})}\right)(1-\alpha)}{\epsilon-c_{\mathcal{G}}} \tag{B.37}$$

$$=(1-\alpha)-(\eta_{\mathcal{F}}+\eta_{\mathcal{G}})-\frac{\epsilon(1+\alpha)-c_{\mathcal{F}}(1-\alpha)}{\epsilon-c_{\mathcal{G}}}-\frac{\epsilon(1-\alpha)}{\epsilon-c_{\mathcal{G}}}\sqrt{2-2\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})}. \tag{B.38}$$

This completes the proof. □

Implications. In both Theorem B and Theorem 1, the only term which correlates both $\mathcal{F}$ and $\mathcal{G}$ is $\mathcal{S}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})$, while all other terms are dependent on either model $\mathcal{F}$ or $\mathcal{G}$ individually. Thus, we can view all other terms as constant. Note that $c_{\mathcal{G}}$ is within the range $(-\epsilon,\epsilon)$ and usually very small, because β is usually very small compared with ϵ, and the attack is typically successful. Then both $\Pr\left(T_r(\mathcal{F},\mathcal{G},x,y_t)=1\right)$ and $\Pr\left(T_r(\mathcal{F},\mathcal{G},x)=1\right)$ have the form $C-k\sqrt{1-\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})}$, where C and $k>0$ are both constants. We can easily observe the positive correlation between $\underline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}})$—loss gradient similarity, and lower bound of $T_r(\mathcal{F},\mathcal{G},x,y_t)$ or $T_r(\mathcal{F},\mathcal{G},x)$—adversarial transferability.

Meanwhile, note that when β increases (i.e., model smoothness decreases), both lower bounds decreases, which implies that less model smoothness would reduce the tightness of the lower bound. In other words, when the model becomes smoother, the correlation between gradient similarity and transferability lower bound becomes stronger, which motivates us to improve the model smoothness to increase the effect of constraining gradient similarity.

C. Proof of Transferability Upper Bound (Theorem 2)

Here we present the proof of Theorem 2 as stated in Section 2.2.3.

The following lemma is used in the proof.

Lemma 6. *Suppose two unit vectors x, y satisfy $x \cdot y < S$, then for any δ, we have* $\min(\delta \cdot x, \delta \cdot y) < \|\delta\|_2 \sqrt{(1+S)/2}$.

Proof. For sake of contradiction, suppose $\delta \cdot x > \|\delta\|_2 \sqrt{(1+S)/2}$, $\delta \cdot y > \|\delta\|_2 \sqrt{(1+S)/2}$. Denote α to be the angel between x and y, then $\cos\alpha < S$, or $\alpha > \arccos S$. If α_x, α_y are the angles between δ and x and between δ and y respectively, then we have $\max(\alpha_x, \alpha_y) \geq \alpha/2 \geq \arccos S/2$. Thus $\min(\delta \cdot x, \delta \cdot y) \leq \|\delta\|_2 \cos(\alpha/2) = \|\delta\|_2 \sqrt{(1+S)/2}$. □

Theorem (Upper Bound on Targeted Attack Transferability)**.** *Consider an instance $x \in \mathbb{R}^n$ with true label y and adversarial target y_t. Assume both model $\mathcal{F}$ and $\mathcal{G}$ are β-smooth with gradients bounded by B. An $(\alpha, \mathcal{F})$-effective (targeted) attack $x^{\mathcal{A}} = \mathcal{A}_T(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with bounded probability*

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) \leq \frac{\xi_{\mathcal{F}} + \xi_{\mathcal{G}}}{\ell_{\min} - \epsilon B \left(1 + \sqrt{\frac{1 + \overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}{2}}\right) - \beta\epsilon^2}, \tag{C.1}$$

$$\textit{where} \quad \ell_{\min} = \min_{x \in \mathcal{X}}(\ell_{\mathcal{F}}(x, y_t), \ell_{\mathcal{G}}(x, y_t)),\ \xi_{\mathcal{F}} = \mathbb{E}_{x,y}\left[\ell_{\mathcal{F}}(x, y)\right],\ \xi_{\mathcal{G}} = \mathbb{E}_{x,y}\left[\ell_{\mathcal{G}}(x, y)\right].$$

Here $\xi_{\mathcal{F}}$ and $\xi_{\mathcal{G}}$ are the empirical risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively, defined relative to a differentiable loss.*

Proof. Since $\mathcal{F}(x)$ outputs label for which $\ell_{\mathcal{F}}$ is minimized, we have

$$\mathcal{F}(x) = y \Rightarrow \ell_{\mathcal{F}}(x, y_t) > \ell_{\mathcal{F}}(x, y) \tag{C.2}$$

and similarly

$$\mathcal{F}(x^{\mathcal{A}}) = y_t \Rightarrow \ell_{\mathcal{F}}(x^{\mathcal{A}}, y) > \ell_{\mathcal{F}}(x^{\mathcal{A}}, y_t), \tag{C.3}$$

$$\mathcal{G}(x) = y \Rightarrow \ell_{\mathcal{G}}(x, y_t) > \ell_{\mathcal{G}}(x, y), \tag{C.4}$$

$$\mathcal{G}(x^{\mathcal{A}}) = y_t \Rightarrow \ell_{\mathcal{G}}(x^{\mathcal{A}}, y) > \ell_{\mathcal{G}}(x^{\mathcal{A}}, y_t). \tag{C.5}$$

Since $\ell_{\mathcal{F}}(x, y)$ and $\ell_{\mathcal{G}}(x, y)$ are β-smooth,

$$\ell_{\mathcal{F}}(x, y) + \delta \cdot \nabla_x \ell_{\mathcal{F}}(x, y) + \frac{\beta}{2}\|\delta\|^2 \geq \ell_{\mathcal{F}}(x^{\mathcal{A}}, y), \tag{C.6}$$

which implies

$$\delta \cdot \nabla_x \ell_{\mathcal{F}}(x, y) \geq \ell_{\mathcal{F}}(x^{\mathcal{A}}, y) - \ell_{\mathcal{F}}(x, y) - \frac{\beta}{2}\|\delta\|^2 \tag{C.7}$$

$$\geq \ell_{\mathcal{F}}(x^{\mathcal{A}}, y_t) - \ell_{\mathcal{F}}(x, y) - \frac{\beta}{2}\|\delta\|^2 =: c'_{\mathcal{F}}. \tag{C.8}$$

Similarly for $\mathcal{G}$,

$$\delta \cdot \nabla_x \ell_{\mathcal{G}}(x, y) \geq \ell_{\mathcal{G}}(x^{\mathcal{A}}, y_t) - \ell_{\mathcal{G}}(x, y) - \frac{\beta}{2}\|\delta\|^2 =: c'_{\mathcal{G}}. \tag{C.9}$$

Thus,

$$\begin{aligned}
&\Pr\left(\mathcal{F}(x)=y,\mathcal{G}(x)=y,\mathcal{F}(x^{\mathcal{A}})=y_t,\mathcal{G}(x^{\mathcal{A}})=y_t\right) && \text{(C.10)}\\
\leq&\Pr\left(\ell_{\mathcal{F}}(x,y_t)>\ell_{\mathcal{F}}(x,y),\ell_{\mathcal{F}}(x^{\mathcal{A}},y)>\ell_{\mathcal{F}}(x^{\mathcal{A}},y_t),\ell_{\mathcal{G}}(x,y_t)>\ell_{\mathcal{G}}(x,y),\ell_{\mathcal{G}}(x^{\mathcal{A}},y)>\ell_{\mathcal{G}}(x^{\mathcal{A}},y_t)\right) && \text{(C.11)}\\
\leq&\Pr\left(\delta\cdot\nabla_x\ell_{\mathcal{F}}(x,y)\geq c'_{\mathcal{F}},\,\delta\cdot\nabla_x\ell_{\mathcal{G}}(x,y)\geq c'_{\mathcal{G}}\right) && \text{(C.12)}\\
\leq&\Pr\left(\left(c'_{\mathcal{F}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{F}}(x,y)\|_2\right)\bigcup\left(c'_{\mathcal{G}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2\right)\right) && \text{(C.13)}\\
\leq&\Pr\left(c'_{\mathcal{F}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{F}}(x,y)\|_2\right)+\Pr\left(c'_{\mathcal{G}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2\right), && \text{(C.14)}
\end{aligned}$$

where inequality Equation (C.11) comes from Equation (C.2) - Equation (C.5), inequality Equation (C.12) comes from Equation (C.8) and Equation (C.9). Equation (C.13) is a result of Lemma 6: either $\delta\cdot\frac{\nabla_x\ell_{\mathcal{F}}(x,y)}{\|\nabla_x\ell_{\mathcal{F}}(x,y)\|_2}\leq\|\delta\|_2\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}$ or $\delta\cdot\frac{\nabla_x\ell_{\mathcal{G}}(x,y)}{||\nabla_x\ell_{\mathcal{G}}(x,y)||}\leq\|\delta\|_2\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}$.

We observe that by β-smoothness condition of the loss function,

$$\begin{aligned}
c'_{\mathcal{F}}&=\ell_{\mathcal{F}}(x^{\mathcal{A}},y_t)-\ell_{\mathcal{F}}(x,y)-\frac{\beta}{2}\|\delta\|_2^2\\
&\geq\ell_{\mathcal{F}}(x,y_t)+\delta\cdot\nabla_x\ell_{\mathcal{F}}(x,y_t)-\frac{\beta}{2}\|\delta\|_2^2-\ell_{\mathcal{F}}(x,y)-\frac{\beta}{2}\|\delta\|_2^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr\left(c'_{\mathcal{F}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{F}}(x,y)\|_2\right) && \text{(C.15)}\\
&\leq\Pr\left(\ell_{\mathcal{F}}(x,y_t)-\ell_{\mathcal{F}}(x,y)\leq\epsilon B(1+\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2})+\beta\epsilon^2\right) && \text{(C.16)}\\
&\leq\Pr\left(\ell_{\mathcal{F}}(x,y)\geq\ell_{\mathcal{F}}(x,y_t)-\epsilon B(1+\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}-\beta\epsilon^2\right) && \text{(C.17)}\\
&\leq\frac{\xi_{\mathcal{F}}}{\min\limits_{x\in\mathcal{X}}\ell_{\mathcal{F}}(x,y_t)-\epsilon B\left(1+\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\right)-\beta\epsilon^2}. && \text{(C.18)}
\end{aligned}$$

Similarly for $\mathcal{G}$,

$$\begin{aligned}
&\Pr\left(c'_{\mathcal{G}}\leq\epsilon\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\|\nabla_x\ell_{\mathcal{G}}(x,y)\|_2\right) && \text{(C.19)}\\
&\leq\frac{\xi_{\mathcal{G}}}{\min\limits_{x\in\mathcal{X}}\ell_{\mathcal{G}}(x,y_t)-\epsilon B\left(1+\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2}\right)-\beta\epsilon^2}. && \text{(C.20)}
\end{aligned}$$

Combining the two and inject them into Equation (C.14), we get

$$\Pr\left(T_r(\mathcal{F},\mathcal{G},x,y_t)=1\right)\leq\frac{\xi_{\mathcal{F}}+\xi_{\mathcal{G}}}{\ell_{\min}-\epsilon B(1+\sqrt{(1+\overline{\mathcal{S}}(\ell_{\mathcal{F}},\ell_{\mathcal{G}}))/2})-\beta\epsilon^2}.$$

□

Theorem (Upper Bound on Untargeted Attack Transferability)**.** *Consider an instance $x\in\mathbb{R}^n$ with true label y and adversarial target y_t. Assume both model $\mathcal{F}$ and $\mathcal{G}$ are β-smooth with gradients bounded by B. An $(\alpha,\mathcal{F})$-effective*

(untargeted) adversarial attack $x^{\mathcal{A}} = \mathcal{A}_U(x)$ with perturbation ball $\|\delta\|_2 \leq \epsilon$ is transferable to $\mathcal{G}$ with probability

$$\Pr\left(T_r(\mathcal{F}, \mathcal{G}, x, y_t) = 1\right) \leq \frac{\xi_{\mathcal{F}} + \xi_{\mathcal{G}}}{\ell_{\min} - \epsilon B\left(1 + \sqrt{\frac{1 + \overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})}{2}}\right) - \beta\epsilon^2}, \tag{C.21}$$

where $\ell_{\min} = \min_{x,y':(x,y)\in\mathcal{M},y'\neq y}(\ell_{\mathcal{F}}(x,y'), \ell_{\mathcal{G}}(x,y'))$, $\xi_{\mathcal{F}} = \mathbb{E}_{x,y}\left[\ell_{\mathcal{F}}(x,y)\right]$, $\xi_{\mathcal{G}} = \mathbb{E}_{x,y}\left[\ell_{\mathcal{G}}(x,y)\right]$.

Here $\xi_{\mathcal{F}}$ and $\xi_{\mathcal{G}}$ are the empirical risks *of models $\mathcal{F}$ and $\mathcal{G}$ respectively, defined relative to a differentiable loss.*

Proof. The proof follows the proof for the targeted attack case, but instead of $\min_{(x,y_t)} \ell_{\mathcal{F}/\mathcal{G}}(x, y_t)$ we use $\min_{x,y':(x,y)\in\mathcal{M},y'\neq y} \ell_{\mathcal{F}/\mathcal{G}}(x, y')$ in C.18 and henceforth. □

Implications In both Theorem C and Theorem 2, we can observe that along with the rise of $\overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})$, the denominator decreases and henceforth the upper bound increases. Therefore, $\overline{\mathcal{S}}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}})$—upper loss gradient similarity and the upper bound of $T_r(\mathcal{F}, \mathcal{G}, x, y_t)$ or $T_r(\mathcal{F}, \mathcal{G}, x)$ are positively correlated.

Meanwhile, when β increases, T_r also increases, which implies that when the model becomes smoother (i.e., β decreases), the transferability upper bound decreases and becomes tighter, which motivates us to improve the model smoothness to increase the effect of constraining gradient similarity.

D. Baseline model details

CosSim considers the ensemble which only minimizes the gradient similarity between base models. It serves as a baseline to empirically verify our theoretical analysis that based on similar model loss gradient similarity, the smoother the models are, the less transferable they are. Such intuition is also illustrated in Figure 1. **CKAE** develops ensemble based on CKA measurement, which is recently shown to be effective to measure the orthogonality between representations [17]. In this paper we consider leveraging such representation orthogonality measurements as objective function to evaluate its effectiveness on reducing transferability.

AdaBoost. To generate a robust ensemble, it is natural to consider different variants of boosting algorithms, which build different weak learners in a sequential manner improving diversity in handling different task partitions. We consider AdaBoost [12], for which the final prediction will be a weighted average of all the weak learners where higher weight is placed on stronger learners.

GradientBoost. To further explore the "weakness" of model gradient, we also include GradientBoost [15] as another baseline to identify weaker learners based on gradient information and therefore generate an ensemble for comparison.

ADP (Adaptive Diversity Promoting) is proposed recently as a regularizer to reduce transferability between base models within an ensemble to improve robustness [25]. So far ADP has achieved state-of-the-art ensemble robustness performance. We will follow the same setting as the ADP method and compare the performance of our robust TRS-ensemble with ADP and other baselines.

E. TRS-ensemble Algorithm

Here we present the Robust TRS-ensemble training procedure. For every mini-batch samples, we obtain the gradients of each submodel to these samples and calculate our TRS loss. We optimize our model by minimizing the combination of the ensemble cross-entropy loss and the TRS loss until we reach the model's convergence.

Algorithm 1 Robust TRS-ensemble training.

Input: A training dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots\}$, models $\{F_{\theta_1}(x), \ldots, F_{\theta_N}(x)\}$.
for all $(x_i, y_i) \in \mathcal{D}$ **do**
  / Calculate classification loss and gradient. */*
  for $k = 1, \ldots, N$ **do**
    $\ell_k = \text{cross\_entropy}(F_{\theta_k}(x_i), y_i)$
    $g_k = \nabla_x \ell_k$
  end for
  / Calculate L_{TRS}. */*
  $\ell^{TRS} = \sum_{j<k} L_{TRS}(g_j, g_k)$
  / Update each model using the aggregated loss. */*
  for $k = 1, \ldots, N$ **do**
    $\theta_k = \theta_k - \eta \cdot \nabla_{\theta_k} \left(\ell_k + \lambda_{TRS} \ell^{TRS}\right)$
  end for
end for
Return: the trained models $\{F_{\theta_1}(x), \ldots, F_{\theta_N}(x)\}$.

F. Experimental Results and Details

Experiment details: We use ResNet-20 [13] as our ensemble models in both MNIST and CIFAR-10 Dataset. During training we use the Adam optimizer [16] with initial learning rate $\alpha = 0.001$. We run 40 epochs on MNIST and 180 epochs on CIFAR-10 to make sure the loss has converged well.

Here we show the details of our experiment results on CIFAR-10 as in Table 2. We see that our TRS approach outperforms any other methods including previous state-of-the-art ADP method[25] on most tasks except on FGSM. On difficult cases such as CW attack with $c = 0.1$, the performance improvement is much more significant.

G. Adversarial Transferability Analysis

We show the full transferability table of the untargeted attack on both MNIST and CIFAR10 model in Figure 3, 4, compared to the baseline model without TRS training. We can see that both individual model's robustness increased and the attack transferability decreased significantly.

Table 2: Robustness of different models against various whitebox attacks on CIFAR-10.

CIFAR-10	para.	CosSim	CKAE	AdaBoost	GradientBoost	$ADP_{2,0.5}$	**TRS**
FGSM	$\epsilon = 0.02$	31.0	38.2	32.1	33.2	**61.7**	40.9
	$\epsilon = 0.04$	20.1	18.5	20.5	23.1	**46.2**	15.2
BIM	$\epsilon = 0.01$	18.1	46.8	19.5	50.5	46.6	**65.2**
	$\epsilon = 0.02$	9.3	32.3	15.9	31.5	31.0	**33.8**
PGD	$\epsilon = 0.01$	21.5	46.1	35.2	52.1	48.4	**73.1**
	$\epsilon = 0.02$	9.1	28.2	15.1	31.2	30.4	**47.3**
MIM	$\epsilon = 0.01$	22.5	45.6	41.2	49.7	52.1	**66.0**
	$\epsilon = 0.02$	10.1	24.1	15.2	28.5	35.9	**36.0**
CW	$c = 0.001$	65.1	81.2	70.3	75.4	80.6	**83.7**
	$c = 0.1$	13.6	23.0	18.4	26.2	25.6	**62.6**

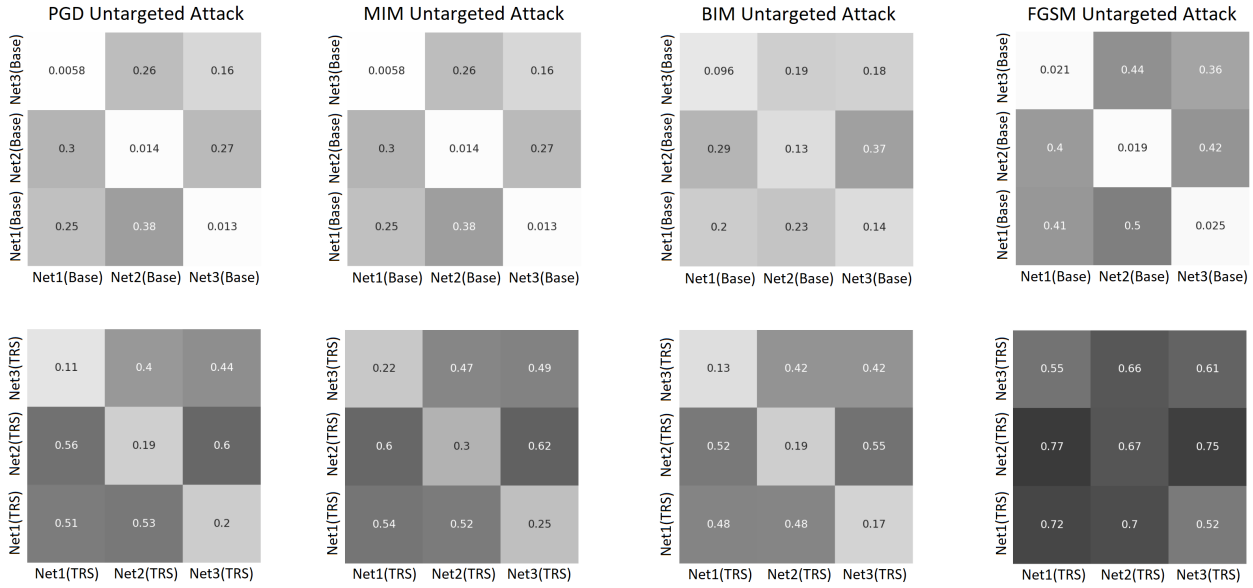

Figure 3: Transferability analysis for baseline and TRS models on MNIST with $\epsilon = 0.2$. Each cell shows the *classification accuracy* on the generated adversarial examples against different models. The first row presents vanilla models, and second row the base models within TRS-ensemble. It is clear that for the base models within TRS-ensemble the accuracy is higher indicating lower transferability.

H. Comparison of Different TRS-Based Models

We evaluate the robustness of several TRS-based models under MNIST and CIFAR-10 dataset and use the current state-of-art ensemble method ADP[25] as our baseline. In order to show our model's robustness under stronger attack, we evaluate PGD attack with more iterations (110 iterations in total). Results are shown in Table 3, 4. We show how the models robustness related to the different number of our sub-models ($k = 2, 3$), and how adversarial training will further improve our TRS-based model. As we can see, the robust accuracy increased with the increasing of the number of sub-models in all scenario of MNIST dataset and most scenario of CIFAR-10 dataset. And when we utilize the PGD Adversarial Training [24] to our ensemble model, the robust accuracy reaches the best in all settings.

In order to show that the adjusting parameter c of TRS loss will not affect the model's robustness too much so we don't need to tune it carefully, we show the comparison of TRS-loss based ensemble model with different parameter c. Results are shown in Table 5.

I. Robustness Against Intermediate Level Attack (ILA)

We also evaluate the robustness of our model against a novel Intermediate Level Attack (ILA) which aims at enhancing adversarial example transferbility. In particular, after we trained the ensemble of three sub-models, we will perform ILA on

Table 3: Robustness of different TRS-based models against various whitebox attacks on MNIST, k means the number of submodels. AT means adding adversarial training loss.

MNIST	para.	TRS($k=2$)	TRS($k=3$)	TRS($k=3$) + AT	$\text{ADP}_{2,0.5}$
FGSM	$\epsilon=0.1$	94.8	95.0	97.4	96.3
	$\epsilon=0.2$	61.6	65.7	91.2	52.8
BIM	$\epsilon=0.1$	91.4	91.7	97.2	88.5
	$\epsilon=0.15$	67.2	74.4	94.2	73.6
PGD - 10 iter	$\epsilon=0.1$	92.8	93.3	97.6	82.8
	$\epsilon=0.15$	71.5	74.0	95.7	41.0
PGD - 110 iter	$\epsilon=0.1$	88.6	89.7	97.1	28.1
	$\epsilon=0.15$	48.6	53.2	93.3	1.18
MIM	$\epsilon=0.1$	92.1	92.5	97.3	92.0
	$\epsilon=0.15$	73.1	75.4	94.6	77.5
CW	$c=0.1$	94.8	95.2	96.0	97.3
	$c=10$	39.4	43.2	57.2	23.8

Table 4: Robustness of different TRS-based models against various whitebox attacks on CIFAR-10, k means the number of submodels. AT means adding adversarial training loss.

CIFAR-10	para.	TRS($k=2$)	TRS($k=3$)	TRS($k=3$) + AT	$\text{ADP}_{2,0.5}$
FGSM	$\epsilon=0.02$	42.0	40.9	67.1	61.7
	$\epsilon=0.04$	21.8	15.2	44.9	46.2
BIM	$\epsilon=0.01$	59.2	65.2	78.4	46.6
	$\epsilon=0.02$	36.7	33.8	64.9	31.0
PGD - 10 iter	$\epsilon=0.01$	65.4	73.1	81.0	48.4
	$\epsilon=0.02$	46.6	47.3	71.9	30.4
PGD - 110 iter	$\epsilon=0.01$	57.7	63.1	78.0	15.1
	$\epsilon=0.02$	32.1	27.9	62.8	5.3
MIM	$\epsilon=0.01$	59.8	66.0	78.5	52.1
	$\epsilon=0.02$	38.5	36.0	65.8	35.9
CW	$c=0.001$	73.8	83.7	85.5	80.6
	$c=0.1$	61.6	62.6	82.4	25.6

Table 5: Robustness of TRS-based models with different adjusting parameter c against various whitebox attacks on MNIST. The number of ensemble models k equals to 3 and $\lambda_1=\lambda_2=0.1$ for all models.

MNIST	para.	TRS($c=0.5$)	TRS($c=1.0$)	TRS($c=2.0$)
FGSM	$\epsilon=0.1$	90.3	92.1	95.0
	$\epsilon=0.2$	62.4	64.1	65.7
BIM	$\epsilon=0.1$	89.5	91.2	91.7
	$\epsilon=0.15$	68.2	70.0	74.4
PGD	$\epsilon=0.1$	87.3	90.7	93.3
	$\epsilon=0.15$	72.8	74.2	74.0
MIM	$\epsilon=0.1$	89.0	91.7	92.5
	$\epsilon=0.15$	70.7	73.4	75.4
CW	$c=0.1$	94.1	94.8	95.2
	$c=10$	38.1	40.5	43.2

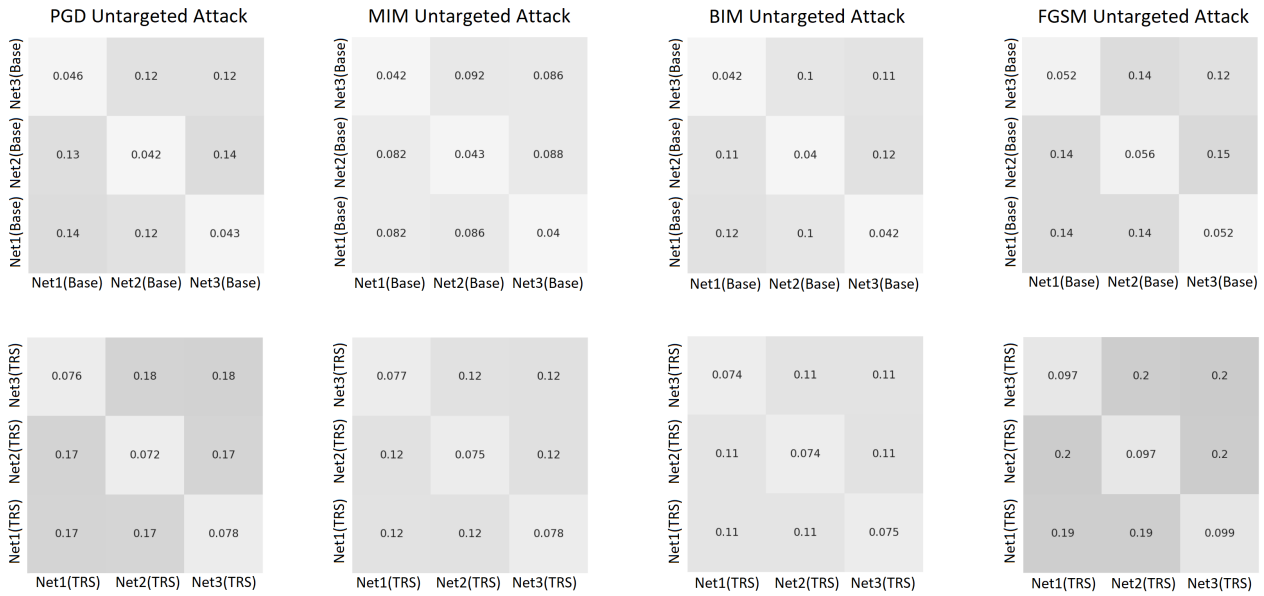

Figure 4: Transferability analysis for baseline and TRS models on CIFAR10 with $\epsilon = 0.05$. Each cell shows the *classification accuracy* on the generated adversarial examples against different models. The first row presents vanilla models, and second row the base models within TRS-ensemble.

sub-model 1 and evaluate the adversarial examples on the ensemble model and each sub-model. We tried $\epsilon = 0.1/0.15$ on MNIST and $\epsilon = 0.01/0.02$ on CIFAR and perform the attack for 10 iterations. The result is shown in Table 6. We see that the attack does decrease the performance on sub-model 1 (on which the attack is performed). But it does not affect the performance of other sub-models and the ensemble model. This shows that our TRS approach does reduce the transferbility of the adversarial examples.

Table 6: Robustness of TRS-based models against ILA attack. The attack is performed on sub-model 1 and evaluated on other models.

Task	para.	Sub-model 1	Sub-model 2	Sub-model 3	Ensemble Model
MNIST	$\epsilon = 0.1$	92.66	96.43	96.94	96.22
	$\epsilon = 0.15$	69.56	88.03	89.19	86.26
CIFAR-10	$\epsilon = 0.01$	71.18	79.75	80.17	78.82
	$\epsilon = 0.02$	37.13	58.28	57.77	51.93