UNDERSTANDING MODEL ENSEMBLE IN TRANSFERABLE ADVERSARIAL ATTACK

Anonymous authors
Paper under double-blind review

ABSTRACT

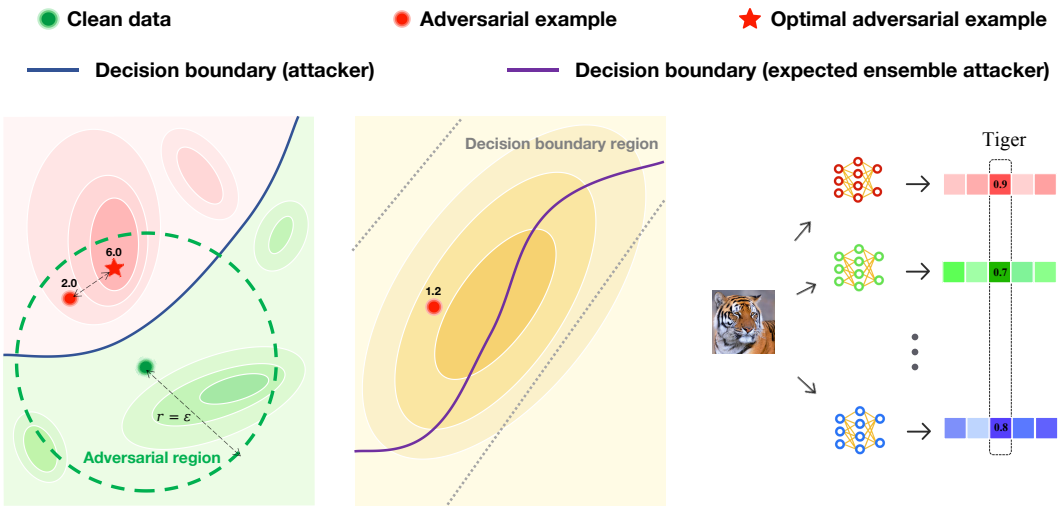
Model ensemble adversarial attack has become a powerful method for generating transferable adversarial examples that can target even unknown models, but its theoretical foundation remains underexplored. To address this gap, we provide early theoretical insights that serve as a roadmap for advancing model ensemble adversarial attack. We first define transferability error to measure the error in adversarial transferability, alongside concepts of diversity and empirical model ensemble Rademacher complexity. We then decompose the transferability error into vulnerability, diversity, and a constant, which rigidly explains the origin of transferability error in model ensemble attack: the vulnerability of an adversarial example to ensemble components, and the diversity of ensemble components. Furthermore, we apply the latest mathematical tools in information theory to bound the transferability error using complexity and generalization terms, contributing to three practical guidelines for reducing transferability error: (1) incorporating more surrogate models, (2) increasing their diversity, and (3) reducing their complexity in cases of overfitting. Finally, extensive experiments with 54 models validate our theoretical framework, representing a significant step forward in understanding transferable model ensemble adversarial attacks.

1 INTRODUCTION

Neural networks are highly vulnerable to adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014)—perturbations that closely resemble the original data but can severely compromise safety-critical applications (Zhang & Li, 2019; Kong et al., 2020; Bortsova et al., 2021). Even more concerning is the phenomenon of adversarial transferability (Papernot et al., 2016; Liu et al., 2017): adversarial examples crafted to deceive one model often succeed in attacking others. This property enables attacks without requiring any knowledge of the target model, significantly complicating efforts to ensure the robustness of neural networks (Dong et al., 2019; Silva & Najafirad, 2020).

To enhance adversarial transferability, researchers have proposed a range of algorithms that fall into three main categories: input transformation (Xie et al., 2019; Wang et al., 2021), gradient-based optimization (Gao et al., 2020; Xiong et al., 2022), and model ensemble attacks (Li et al., 2020; Chen et al., 2024b). Among these, model ensemble attacks have proven especially powerful, as they leverage multiple models to simultaneously generate adversarial examples that exploit the strengths of each individual model (Dong et al., 2018). Moreover, these attacks can be combined with input transformation and gradient-based optimization methods to further improve their effectiveness (Tang et al., 2024). However, despite the success of such attacks, their theoretical foundation remains poorly understood. This prompts an important question: *Can we establish a theoretical framework for transferable model ensemble adversarial attacks to shape the evolution of future algorithms?*

To conduct a preliminary exploration of this profound question, we propose three novel definitions as a prerequisite of our theoretical framework. Firstly, we define *transferability error* as the gap in expected loss between an adversarial example and the one with the highest loss within a feasible region of the input space. It captures the ability of an adversarial example to generalize across unseen models, representing its transferability. Secondly, we introduce *prediction variance* across the ensemble classifiers. It address an open problem in model ensemble attack about how to to quantify diversity and assists in selecting ensemble components. Finally, we also introduce the *empirical*



(a) Transferability Error (b) Vulnerability (c) Diversity

Figure 1: Vulnerability-diversity decomposition of transferability error. (a) The transferability error of a given adversarial example is defined as the difference in expected loss value between that example and the most transferable one. (b) Vulnerability is the loss value of the expected ensemble classifier on the adversarial example. (c) Diversity represents the variance in model ensemble predictions that correspond to the correct class.

model ensemble Rademacher complexity, inspired by Rademacher complexity (Bartlett & Mendelson, 2002), as a measure of the flexibility of ensemble components in an attack.

With these three definitions, we offer two key theoretical insights. First, we show the ***vulnerability-diversity decomposition*** of transferability error (Figure 1), highlighting the preference for ensemble components that are powerful attackers and induce greater prediction variance among themselves. However, this also uncovers a *fundamental trade-off between vulnerability and diversity*, making it challenging to maximize both simultaneously. To mitigate this issue and provide more practical guidelines, we present an upper bound for transferability error, incorporating empirical model ensemble Rademacher complexity and a generalization term. The primary challenge in proof lies in the application of cutting-edge mathematical tools from information theory (Esposito & Mondelli, 2024), which are crucial for addressing the complex issue of relaxing the independence assumption among surrogate classifiers. Our theoretical analysis leads to a crucial takeaway for practitioners: Including ***more and diverse surrogate models*** with ***reduced model complexity in cases of overfitting*** helps tighten the transferability error bound, thereby improving adversarial transferability. Finally, the experimental results support the soundness of our theoretical framework, highlighting a key step forward in the deeper understanding of transferable model ensemble adversarial attacks.

2 Related Work

2.1 Transferable Adversarial Attack

Researchers have developed various algorithms to enhance adversial transferability. Most of them fall into three categories: input transformation, gradient-based optimization, and model ensemble attack. **Input transformation** techniques apply data augmentation strategies to prevent overfitting to the surrogate model. For instance, random resizing and padding (Xie et al., 2019), downscaling (Lin et al., 2019), and mixing (Wang et al., 2021). **Gradient-based optimization** optimizes the generation of adversarial examples to achieve better transferability. Some popular ideas include applying momentum (Dong et al., 2018), Nesterov accelerated gradient (Lin et al., 2019), scheduled step size (Gao et al., 2020) and gradient variance reduction (Xiong et al., 2022). **Model ensemble attack** combine outputs from surrogate models to create an ensemble loss, increasing the likelihood to deceive various models simultaneously. It can be applied collectively with both input transformation and gradient-based optimization algorithms (Tang et al., 2024). Some popular ensemble paradigms

include loss-based ensemble (Dong et al., 2018), prediction-based (Liu et al., 2017), logit-based ensemble (Dong et al., 2018), and longitudinal strategy (Li et al., 2020). Moreover, advanced ensemble algorithms have been created to ensure better adversarial transferability (Li et al., 2023; Wu et al., 2024; Chen et al., 2024b). An extended and detailed summary of related work is in Appendix C.

Within the extensive body of research on model ensemble attacks, two notable observations stand out. First, increasing the number of models in an ensemble improves adversarial transferability (Liu et al., 2017; Dong et al., 2018; Lin et al., 2019; Gubri et al., 2022b). Second, using more diverse surrogate models with varying architectures and back-propagated gradients (Tang et al., 2024) further enhances transferability. However, to our best knowledge, these intriguing phenomena have yet to be fully understood from a theoretical perspective. In this paper, we are the first to provide a theoretical explanation for them, offering insights that can guide future algorithm design.

2.2 Theoretical Understanding of Adversarial Transferability

In contrast to the wealth of empirical and intuitive studies, research on the theoretical understanding of adversarial transferability remains limited. Recent efforts have primarily focused on aspects such as data (Tramèr et al., 2017), surrogate model (Wang & Farnia, 2023), optimization (Yang et al., 2021; Zhang et al., 2024a; Chen et al., 2024b) and target model (Zhao et al., 2023). Tramèr et al. (2017) investigates the space of transferable adversarial examples and establishes conditions on the data distribution that suggest transferability for some basic models. In terms of the surrogate model generalization, Wang & Farnia (2023) builds the generalization gap to show that a surrogate model with a smaller generalization error leads to more transferable adversarial examples. From an optimization perspective, Yang et al. (2021); Zhang et al. (2024a) establish upper and lower bounds on adversarial transferability, linking it to model smoothness and gradient similarity, while Chen et al. (2024b) provides theoretical evidence connecting transferability to loss landscape flatness and closeness to local optima. Regarding the target model, Zhao et al. (2023) theoretically reveals that reducing the discrepancy between the surrogate and target models can limit adversarial transferability.

Despite these theoretical advances, to the best of our knowledge, *transferable model ensemble adversarial attacks remain unexplored*. To address this gap, we take a pioneering step by presenting the first theoretical analysis of such attacks. Our work not only offers theoretical insights into these attacks but also incorporates recent advancements in learning theory, laying the groundwork for future theoretical investigations into adversarial transferability.

3 Key Definitions: Transferability Error, Diversity, and Ensemble Complexity

In this section, we first highlight the fundamental goal of model ensemble adversarial attack (Section 3.1). Then we define the transferability error (Section 3.2), diversity in transferable model ensemble attack (Section 3.3) and empirical model ensemble Rademacher complexity (Section 3.4).

3.1 Model Ensemble Adversarial Attack

Given the input space $\mathcal{X} \subset \mathbb{R}^d$ and the output space $\mathcal{Y} \subset \mathbb{R}$, we have a joint distribution $\mathcal{P}_\mathcal{Z}$ over the input space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The training set $Z_{\text{train}} = \{z_i | z_i = (x_i, y_i) \in \mathcal{Z}, y_i \in \{-1, 1\}, i = 1, \cdots, K\}$, which consists of K examples drawn independently from $\mathcal{P}_\mathcal{Z}$. We denote the hypothesis space by $\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$ and the parameter space by Θ. Let $f(\theta; \cdot) \in \mathcal{H}$ be a classifier parameterized by $\theta \in \Theta$, trained for a classification task using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_0^+$. Let $\mathcal{P}_\Theta$ represent the distribution over the parameter space Θ. Define $\mathcal{P}_{\Theta^N}$ as the joint distribution over the product space Θ^N, which denotes the space of N such sets of parameters. We use Z_{train} to train N surrogate models $f(\theta_1; \cdot), \cdots, f(\theta_N; \cdot)$ for model ensemble. The training process of these N classifiers can be viewed as sampling the parameter sets $(\theta_1, \ldots, \theta_N)$ from the distribution $\mathcal{P}_{\Theta^N}$. For a clean data $\hat{z} = (\hat{x}, y) \in \mathcal{Z}$, an adversarial example $z = (x, y) \in \mathcal{Z}$, and N classifiers for model ensemble attack, define the population risk $L_P(z)$ and the empirical risk $L_E(z)$ of the adversarial example z as

$$L_P(z) = \mathbb{E}_{\theta \sim \mathcal{P}_\Theta}[\ell(f(\theta; x), y)], \tag{1}$$

$$\text{and} \quad L_E(z) = \frac{1}{N}\sum_{i=1}^{N} \ell(f(\theta_i; x), y). \tag{2}$$

Intuitively, a transferable adversarial example leads to a large $L_P(z)$ because it can attack many classifiers with parameter $\theta \in \Theta$. Therefore, the most transferable adversarial example $z^* = (x^*, y)$ around z is defined as

$$x^* = \arg \max_{x \in \mathcal{B}_\epsilon(\hat{x})} L_P(z), \tag{3}$$

where $\mathcal{B}_\epsilon(\hat{x}) = \{x : \|x - \hat{x}\|_2 \leq \epsilon\}$ is an adversarial region centered at $\hat{x}$ with radius $\epsilon > 0$. However, the expectation in $L_P(z)$ cannot be computed directly. Thus, when generating adversarial examples, the empirical version Eq. (2) is used in practice, such as loss-based ensemble attack (Dong et al., 2018). So the adversarial example $z = (x, y)$ is obtained from

$$x = \arg \max_{x \in \mathcal{B}_\epsilon(x)} L_E(z). \tag{4}$$

There is a gap between the adversarial example z we find and the most transferable one z^*. It is due to the fact that the ensemble classifiers cannot cover the whole parameter space of the classifier, i.e., there is a difference between $L_P(z)$ and $L_E(z)$. Accordingly, the core objective of transferable model ensemble attack is to design approaches that approximate $L_E(z)$ to $L_P(z)$, thereby increasing the transferability of adversarial examples.

3.2 Transferability Error

Considering the difference between z and z^*, the transferability of an adversarial example z can be characterized as the difference in population risk between it and the optimal one.

Definition 1 (Transferability Error)**.** *The transferability error of z with radius ϵ is defined as:*

$$TE(z, \epsilon) = L_P(z^*) - L_P(z). \tag{5}$$

There always holds $TE(z, \epsilon) \geq 0$ as $L_P(z^*) \geq L_P(z)$. The closer $TE(z, \epsilon)$ is to 0, the better the transferability of z. Therefore, in principle, the essential goal of various model ensemble attack algorithms is to make transferability error $TE(z, \epsilon)$ as small as possible. Moreover, if the distribution over the parameter space $\mathcal{P}_\Theta$, adversarial region $\mathcal{B}_\epsilon(x)$ and loss function ℓ are fixed, then $L_P(z^*)$ becomes a constant, which means that the goal of minimizing $TE(z, \epsilon)$ becomes maximizing $L_P(z)$.

In the following lemma, we will show how the difference between empirical risk and population risk affects the transferability error of z. The proof is in Appendix B.1.

Lemma 1. *The transferability error defined by Eq. (5) is bounded by the largest absolute difference between $L_P(z)$ and $L_E(z)$, i.e.,*

$$TE(z, \epsilon) \leq 2 \sup_{z \in \mathcal{Z}} |L_P(z) - L_E(z)|. \tag{6}$$

The lemma strictly states that if we can bound the difference between $L_P(z)$ and $L_E(z)$, the transferability error can be constrained to a small value, thereby enhancing adversarial transferability. This indicates that we can develop strategies to make $L_E(z)$ closely approximate $L_P(z)$, ultimately improving the transferability of adversarial examples.

3.3 Quantifying Diversity in Model Ensemble Attack

Before the advent of model ensemble attacks, the formal definition of diversity in ensemble learning had remained a long-standing challenge for decades (Wood et al., 2023). Unfortunately, this challenge continues to persist in the context of transferable model ensemble attacks. Despite various intuitive approaches (Li et al., 2020; Tang et al., 2024), there is still no widely accepted method for rigorously quantifying diversity. In the following definition, we propose measuring diversity among ensemble attack classifiers through prediction variance, building on recent advances in ensemble learning theory (Ortega et al., 2022; Wood et al., 2023).

Definition 2 (Diversity of Model Ensemble Attack)**.** *The diversity of model ensemble attack across $\theta \sim \mathcal{P}_\Theta$ for a specific adversarial example $z = (x, y)$ is defined as the variance of model prediction:*

$$Var_{\theta\sim\mathcal{P}_\Theta}\left(f(\theta;x)\right) = \mathbb{E}_{\theta\sim\mathcal{P}_\Theta}\left[f(\theta;x) - \mathbb{E}_{\theta\sim\mathcal{P}_\Theta} f(\theta;x)\right]^2. \tag{7}$$

It indicates the degree of dispersion in the predictions of different ensemble classifiers for the same adversarial example. The diversity of model ensemble attack is a measure of ensemble member disagreement, independent of the label. From an intuitive perspective, the disagreement among the ensemble components helps prevent the adversarial example from overfitting to the classifiers in the ensemble, thereby enhancing adversarial transferability to some extent.

To calculate the diversity explicitly as a metric, we consider a dataset of adversarial examples $Z_{\text{attack}} = \{z_i | z_i = (x_i, y_i), i = 1, \cdots, M\}$ and N classifiers in the ensemble. The diversity is computed as the average sample variance of predictions for all adversarial examples in the dataset:

$$\text{Diversity} = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{N}\sum_{j=1}^{N}\left(f(\theta_j;x_i) - \frac{1}{N}\sum_{j=1}^{N} f(\theta_j;x_i)\right)^2\right]. \tag{8}$$

Remark. *For multi-class classification problems, $f(\theta;x)$ is replaced with the logit corresponding to the correct class prediction made by the classifier.*

3.4 Empirical Model Ensemble Rademacher Complexity

We define the empirical Rademacher complexity for model ensemble by analogy to the original empirical Rademacher complexity (Koltchinskii & Panchenko, 2000; Bartlett & Mendelson, 2002).

Definition 3 (Empirical Model Ensemble Rademacher Complexity)**.** *Given the input space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and N classifiers $f(\theta_1;\cdot), \cdots, f(\theta_N;\cdot)$. Let $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in[N]}$ be a collection of independent Rademacher variables, which are random variables taking values uniformly in $\{+1, -1\}$. We define the empirical model ensemble Rademacher complexity $\mathcal{R}_N(\mathcal{Z})$ as follows:*

$$\mathcal{R}_N(\mathcal{Z}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f(\theta_i;x),y)\right]. \tag{9}$$

In conventional settings of machine learning, the empirical Rademacher complexity captures how well models from a function class can fit a dataset with random noisy labels (Shalev-Shwartz & Ben-David, 2014). A sufficiently complex function class includes functions that can effectively fit arbitrary label assignments, thereby maximizing the complexity term (Mohri et al., 2018). Likewise, in model ensemble attack, Eq. (9) is expected to measure *the complexity of the input space $\mathcal{Z}$ relative to the N classifiers*. Some extreme cases are analyzed in Appendix D.1.

4 Theoretically Reduce Transferability Error

4.1 Vulnerability-diversity Decomposition of Transferability Error

Inspired by the bias-variance decomposition (Geman et al., 1992; Domingos, 2000) in learning theory, we provide the corresponding theoretical support for prediction variance by decomposing the transferability error into vulnerability, diversity and constants.

Theorem 1 (Vulnerability-diversity Decomposition)**.** *Consider the squared error loss $l(f(\theta;x),y) = [f(\theta;x) - y]^2$ for a data point $z = (x, y)$. Let $\tilde{f}(\theta;x) = \mathbb{E}_{\theta\sim\mathcal{P}_\Theta} f(\theta;x)$ be the expectation of prediction over the distribution on the parameter space. Then there holds*

$$TE(z,\epsilon) = L_P(z^*) - \underbrace{l(\tilde{f}(\theta;x),y)}_{\textit{Vulnerability}} - \underbrace{Var_{\theta\sim\mathcal{P}_\Theta} f(\theta;x)}_{\textit{Diversity}}. \tag{10}$$

The proof and the empirical version of it is in Appendix B.2. The "Vulnerability" term measures the risk of a data point z being compromised by the model ensemble. If the model ensemble is sufficiently strong to fit the direction opposite to the target label, the resulting high loss theoretically

reduces the transferability error. This insight suggests that ***selecting strong attackers*** as ensemble components leads to lower transferability error. The "Diversity" term implies that ***selecting diverse attackers*** in a model ensemble attack theoretically contributing to a reduction in transferability error. In conclusion, Theorem 1 provides the following guideline for reducing transferability error in model ensemble attack: we are supposed to choose ensemble components that are both strong and diverse.

Remark 1. Theorem 1 connects the existing body of work and clarifies how each algorithm strengthens adversarial transferability. For instance, some approaches tend to optimizing the attack process (Xiong et al., 2022; Chen et al., 2023) to improve "Vulnerability", while others aim to diversify surrogate models (Li et al., 2020; 2023; Wang et al., 2024) to enhance "Diversity". Also, there are other definitions of diversity based on gradient in previous literature (Yang et al., 2021; Kariyappa & Qureshi, 2019). A more detailed discussion of prior insights is presented in Appendix D.2.

Remark 2. Theorem 1 and Lemma 5 in Yang et al. (2021) offer complementary perspectives in the analysis of transferable adversarial attack. And the detailed discussion is in Appendix D.3.

However, due to the mathematical nature of Eq. (10), there remains a *vulnerability-diversity trade-off* in model ensemble attacks, similar to the well-known bias-variance trade-off (Geman et al., 1992). This means that, in practice, it is not feasible to maximize both "Vulnerability" and "Diversity" simultaneously. Recognizing this limitation, we proceed with further theoretical analysis to propose more guidelines for practitioners in the following section.

4.2 Upper Bound of Transferability Error

We develop an upper bound of transferability error in this section. We begin by taking Multi-Layer Perceptron (MLP) as an example of deep neural network and derive the upper bound of $\mathcal{R}_N(\mathcal{Z})$. The proof is in Appendix A.4.

Lemma 2 (Ensemble Complexity of MLP). *Let $\mathcal{H} = \{x \mapsto W_l\phi_{l-1}\left(W_{l-1}\phi_{l-2}\left(\dots\phi_1\left(W_1x\right)\right)\right)\}$ be the class of real-valued networks of depth l, where $x \in \mathbb{R}^{d_1}$, $W_i \in \mathbb{R}^{d_{i+1}\times d_i}$. Given N classifiers from $\mathcal{H}$, where the parameter matrix is $W_{ij}, i \in \{1,\cdots,n\}, j \in \{1,\cdots,l\}$ and $T = \prod_{j=1}^{l}\sup_{i\in[n]}\|W_{i,j}\|_F$. Let $\|x\|_F \le B$. With 1-Lipschitz activation functions $\phi_1,\cdots,\phi_{l-1}$ and 1-Lipschitz loss function $\ell(yf(x))$, there holds:*

$$\mathcal{R}_N(\mathcal{Z}) \le \frac{\left(\sqrt{(2\log 2)l}+1\right)BT}{\sqrt{N}}. \tag{11}$$

Remark. *We also derive the upper bound of $\mathcal{R}_N(\mathcal{Z})$ for the cases of linear model (Appendix A.2) and two-layer neural network (Appendix A.3). These results are special cases of the above theorem.*

In particular, a larger N and smaller T will give $\mathcal{R}_N(\mathcal{Z})$ a tighter bound. Notice that T contains the norm of weight matrices, which is related to model complexity (Bartlett et al., 2017; Neyshabur et al., 2018). And a smaller model complexity corresponds to a smaller T (Loshchilov & Hutter, 2019). In summary, Lemma 2 mathematically shows that *increasing the number of surrogate models* and *reducing the model complexity* of them can limit the value of $\mathcal{R}_N(\mathcal{Z})$.

We now provide the upper bound of transferability error, and the proof is in Appendix B.3.

Theorem 2 (Upper bound of Transferability Error). *Given the transferability error defined by Eq. (5) and general rademacher complexity defined by Eq. (9). Let $\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}$ be the joint measure induced by the product of the marginals. If the loss function ℓ is bounded by $\beta \in R_+$ and $\mathcal{P}_{\Theta^N}$ is absolutely continuous with respect to $\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}$ for any function f_i, then for $\alpha > 1$ and $\gamma = \frac{\alpha}{\alpha-1}$, with probability at least $1-\delta$, there holds*

$$TE(z,\epsilon) \le 4\mathcal{R}_N(\mathcal{Z}) + \sqrt{\frac{2\gamma\beta^2}{N}\ln\frac{2^{\frac{1}{\gamma}}H_\alpha^{\frac{1}{\alpha}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}\right)}{\delta}}, \tag{12}$$

where $H_\alpha\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}\right)$ is the Hellinger integrals (Hellinger, 1909) with parameter α, which measures the divergence between two probability distributions if $\alpha > 1$ (Liese & Vajda, 2006).

Remark 1. Theorem 2 can be naturally extended to scenarios where the distributions of the surrogate model and the target model differ. This extension is discussed in Appendix D.4 from two perspectives: domain adaptation theory (Blitzer et al., 2007) and a redefinition of the model space.

Remark 2. Theorem 2 is grounded in the empirical model ensemble Rademacher complexity defined in Eq. (9). However, it can be further extended to information-theoretic analysis (Xu & Raginsky, 2017), as demonstrated in Appendix D.5.

The first term in Eq. (12) suggests that *incorporating* ***more surrogate models*** *with* ***less model complexity*** in ensemble attack will constrain $\mathcal{R}_N(\mathcal{Z})$ and enhances adversarial transferability. Intuitively, incorporating more models helps prevent any single model from overfitting to a specific adversarial example. Such theoretical heuristic is also supported by experimental results (Liu et al., 2017; Dong et al., 2018; Lin et al., 2019; Li et al., 2020; Gubri et al., 2022b; Chen et al., 2023), which also stress the advantage of more surrogate models to obtain transferable attack. Additionally, when there is an overfitting issue, models with reduced complexity will mitigate it.

The second term also suggests that a large N (using more models) can lead to a tighter bound. Furthermore, it motivates the idea that reducing the interdependence among the parameters in ensemble components (i.e., increasing their diversity) results in a tighter upper bound for $TE(z, \epsilon)$. Recall that $H_\alpha(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta})$ represents the divergence between the joint distribution $\mathcal{P}_{\Theta^N}$ and the product of marginals $\mathcal{P}_{\bigotimes_{i=1}^N \Theta}$. The joint distribution captures dependencies, while the product of marginals does not. Therefore, $H_\alpha(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta})$ measures the degree of dependency among the parameters from N classifiers. As a result, ***increasing the diversity of parameters in surrogate models*** and reducing their interdependence enhances adversarial transferability. This theoretical conclusion is also supported by empirical algorithms (Li et al., 2020; Tang et al., 2024), which also advocate for generating adversarial examples from diverse models.

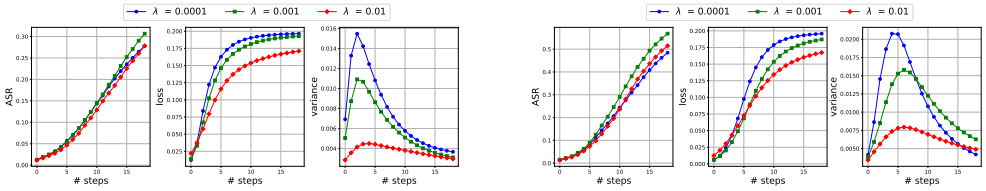
The trade-off between complexity and diversity. Reducing model complexity may conflict with increasing diversity. We discuss this issue from two angles. On one hand, when generating adversarial examples from simpler models to attack more complex ones, the overall model complexity is lower, but diversity may also be limited due to the simpler structure of the ensemble attackers. On the other hand, attacking simpler models with a stronger, more diverse ensemble may increase diversity but also raise model complexity. In this scenario, reducing complexity can help prevent overfitting and lead to a tighter transferability error bound, albeit with a slight reduction in ensemble diversity. In summary, striking a balance between model complexity and diversity is crucial in practice.

From generalization error to transferability error. The mathematical form of Eq. (12) is in line with the generalization error bound (Bartlett & Mendelson, 2002). However, we note that a key distinction between transferability error and generalization error lies in the *independence assumption*. Conventional generalization error analysis relies on an assumption: each data point from the dataset is independently sampled (Zou & Liu, 2023; Hu et al., 2023). By contrast, the surrogate models for ensemble attack are usually trained on the datasets with similar tasks, e.g., image classification. In this case, *we cannot assume these surrogate models behave independently for a solid theoretical analysis*, which increases the difficulty of proof in our paper. To build the gap between generalization error and transferability error, our proof introduces the latest techniques in information theory (Esposito & Mondelli, 2024). And Refer to Appendix D.6 for a detailed discussion.

4.3 Generalization and Adversarial Transferability

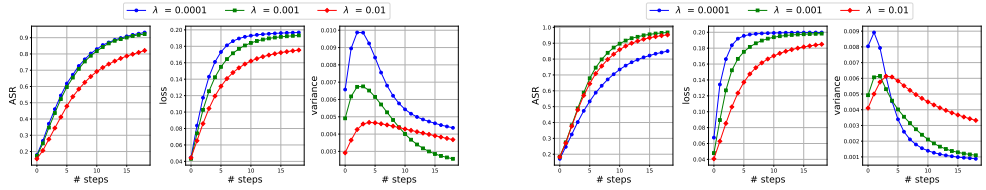
We offer new insights into a foundational and popular analogy in the literature: The transferability of an adversarial example is an analogue to the generalizability of the model (Dong et al., 2018; Lin et al., 2019; Wang et al., 2021; Wang & He, 2021; Xiong et al., 2022; Chen et al., 2024b). Interestingly, our theory also sheds light on this insight in several ways.

First, the mathematical formulations in Lemma 1 is similar to generalization error (Vapnik, 1998; Bousquet & Elisseeff, 2002). Also, Lemma 2 is similar to the bound of the original Rademacher complexity (Golowich et al., 2018). More importantly, recall that in the conventional framework of learning theory: (1) increasing the size of training set typically leads to a better generalization of the model (Bousquet & Elisseeff, 2002); (2) improving the diversity among ensemble classifiers makes it more advantageous for better generalization (Ortega et al., 2022); and (3) reducing the model complexity (Cherkassky, 2002) may mitigate overfitting and benefit the generalization ability.



(a) MLP (b) CNN

Figure 2: Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the MNIST dataset.



(a) MLP (b) CNN

Figure 3: Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the Fashion-MNIST dataset.

These ideas correspond to each of our theoretical understanding in Section 4. Overall, we support the analogy from a theoretical perspective. And further detailed discussion is in Appendix D.7.

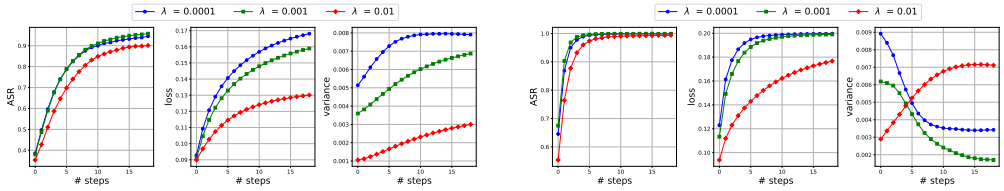
5 Experiments

We conduct our experiments on four datasets, including the MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), and ImageNet-1K (Russakovsky et al., 2015) datasets. We use the first three datasets to empirically validate our theoretical contribution, and leave the experiments on ImageNet to build a powerful ensemble adversarial attack in practice.

We build six deep neural networks for image classification, including three MLPs with one to three hidden layers followed by a linear classification layer, and three convolutional neural networks (CNNs) with one to three convolutional layers followed by a linear classification layer. To ensure diversity among the models, we apply three different types of transformations during training. Additionally, we set the weight decay under the L_2 norm to $10^{-4}, 10^{-3}, 10^{-2}$, respectively. This results in a total of $6 \times 3 \times 3 = 54$ models. To establish a gold standard for adversarial transferability evaluation, we additionally train a ResNet-18 (He et al., 2016) from scratch on three datasets (MNIST, Fashion-MNIST, and CIFAR-10), respectively. We will leverage the models at hand to attack this ResNet-18 for a reliable evaluation. For models trained on MNIST, Fashion-MNIST, we set the number of epochs as 10. For models trained on CIFAR-10, we set the number of epochs as 30. We use the Adam optimizer with setting the learning rate as 10^{-3}. We set the batch size as 64.

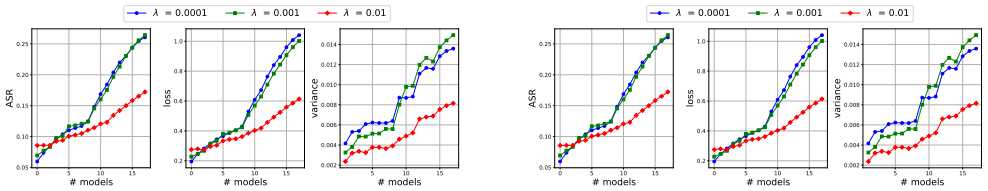
5.1 Evaluation on the Attack Dynamics

For each dataset (MNIST & Fashion-MNIST & CIFAR-10), we record the attack success rate (ASR), loss value, and the variance of model predictions with increasing the number of steps for attack. We use MI-FGSM (Dong et al., 2018) to craft the adversarial example and use the cross-entropy as the loss function to optimize the adversarial perturbation. Generally, the number of steps for the transferable adversarial attack is set as 10 (Zhang et al., 2024b), but to study the attack dynamics more comprehensively, we perform 20-step attack. In our plots, we use the mean-squared-error to validate our theory, which indicates the vulnerability from the theory perspective better. The first metric exhibits an inverse relationship with transferability error. And the latter two metrics correspond to the vulnerability and diversity components in the decomposition in Section 4.1. The number of steps for

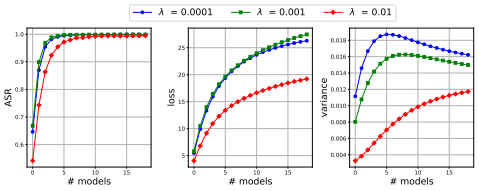


(a) MLP (b) CNN

Figure 4: Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the CIFAR-10 dataset.



(a) MNIST (b) Fashion-MNIST



(c) CIFAR-10

Figure 5: Evaluation of ensemble attacks with increasing the number of models using MLPs and CNNs on the three datasets.

attack is indicated by the x-axis. And we denote λ as the weight decay. We respectively report the results on three datasets in fig. 2, fig. 3, and fig. 4.

Validation of vulnerability-diversity decomposition. Across all three datasets, we observe a consistent pattern: as the number of steps increases, both ASR and loss values improve steadily, meaning that transferability error decreases while vulnerability increases. Notably, the magnitude of variance is approximately ten times smaller than that of the loss value, indicating a much smaller impact on transferability error. Thus, "vulnerability" predominantly drives the vulnerability-diversity decomposition, and the upward trend in vulnerability aligns with the reduction in transferability error.

The trend of variance. On the MNIST and Fashion-MNIST datasets, diversity initially increases but later declines. In contrast, on the CIFAR dataset, the variance for MLP consistently increases, whereas for CNNs, it decreases with a small regularization term but increases with a larger one. This intriguing phenomenon is tied not only to the trade-off between complexity and diversity discussed in Section 4.2, but also to the complex behavior of variance. In the bias-variance trade-off literature, different trends in variance have been observed. For example, Yang et al. (2020) suggests that variance follows a bell-shaped curve, rising initially and then falling as network width expands. Similarly, Lin & Dobriban (2021) provides a detailed decomposition of variance, illustrating the influence of factors like initialization, label noise, and training data. While a full investigation of variance behavior is beyond the scope of this work, we provide additional discussion in Appendix D.8.

The potential trade-off between diversity and complexity. Our experimental results (specifically the "variance" sub-figure), indicate the potential trade-off between diversity and complexity. Consider two distinct phases in the attack dynamics: 1) Initial phase of the attack (first few steps): During this phase, the adversarial example struggles to attack the model ensemble effectively (a low loss). Consequently, both the loss and variance increase, aligning with the vulnerability-diversity decomposition. 2) Potential "over-fitting" phase of the attack (subsequent steps): In this phase, the

adversarial example can effectively attack the model ensemble, achieving a high loss. Here, the trade-off between diversity and complexity becomes evident, particularly at the final step of the attack. As the regularization term λ increases (i.e., lower model complexity), the variance of the model ensemble may increase. For instance, in the variance sub-figure, the red curve may exceed one of the other curves, indicating this potential trade-off.

5.2 Evaluation on the Ensemble Framework

We further validate the effectiveness of the vulnerability-diversity decomposition within the ensemble framework. Specifically, instead of focusing solely on the training dynamics, we progressively increase the number of models in the ensemble attack to evaluate the decomposition's impact. We begin by incorporating MLPs with different architectures and regularization terms, followed by CNNs. In total, up to 18 models are included in a single attack. We depicted the results in fig. 5.

We can consistently observe that increasing the number of ensemble models improves the attack success rate, i.e., reduces the transferability error. On the MNIST and Fashion-MNIST datasets, both vulnerability and diversity also increase as the number of models grows. Although the diversity sometimes shows a decreasing trend on the CIFAR-10 dataset, its magnitude is approximately 100 times smaller than vulnerability, thus having a minimal impact on the attack success rate.

6 Conclusion

In this paper, we address the underdeveloped theoretical foundation of transferable model ensemble adversarial attacks. We introduce three key definitions: transferability error, prediction variance, and empirical model ensemble Rademacher complexity. Through the vulnerability-diversity decomposition of transferability error, we identify a crucial trade-off between vulnerability and diversity in ensemble components, presenting the challenge of optimizing both simultaneously. To overcome this, we introduce recent mathematical tools and derive an upper bound on transferability error, offering practical guidelines for improving adversarial transferability. Our extensive experiments validate these insights, marking a significant advancement in the understanding and development of transferable model ensemble adversarial attacks.

Ethics Statement

Our research adheres to the ICLR Code of Ethics, which promotes responsible stewardship of research and its applications. We recognize the potential societal impact of our work on transferable adversarial attacks and emphasize its contribution to improving the robustness and security of machine learning models. We have been careful to ensure that our findings advance the public good by providing insights that can enhance defenses against malicious use of adversarial attacks, rather than contributing to harmful applications. Our study upholds high standards of scientific excellence through transparency, rigor, and reproducibility. No human subjects were involved, and no privacy or confidentiality concerns arise from the data used. We have also ensured that our work does not introduce discriminatory biases, and we are committed to the fair and inclusive participation of all individuals in the research community. While our research focuses on theoretical advancements, we are aware of the potential risks associated with adversarial attack techniques. We encourage responsible use of these insights to build more secure AI systems and minimize any unintended harm. Finally, all authors have read and adhered to the ICLR Code of Ethics, and any potential conflicts of interest have been disclosed.

References

Taiga Abe, E Kelly Buchanan, Geoff Pleiss, and John P Cunningham. Pathologies of predictive diversity in deep ensembles. *arXiv preprint arXiv:2302.00704*, 2023.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Gerda Bortsova, Cristina González-Gonzalo, Suzanne C Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien PW Pluim, Mitko Veta, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4489–4498, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024a.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, 2024b.

Vladimir Cherkassky. Model complexity control and statistical learning theory. *Natural computing*, 1:109–133, 2002.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Yian Deng and Tingting Mu. Understanding and improving ensemble adversarial defense. *Advances in Neural Information Processing Systems*, 36, 2023.

Alexis Derumigny and Johannes Schmidt-Hieber. On lower bounds for the bias-variance trade-off. *The Annals of Statistics*, 51(4):1510–1533, 2023.

Pedro Domingos. A unified bias-variance decomposition for zero-one and squared loss. *AAAI/IAAI*, 2000:564–569, 2000.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4312–4321, 2019.

Ke-Lin Du and Madisetti NS Swamy. *Neural networks and statistical learning*. Springer Science & Business Media, 2013.

Amedeo Roberto Esposito and Marco Mondelli. Concentration without independence via information measures. *IEEE Transactions on Information Theory*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision*, pp. 307–322, 2020.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, 2018.

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

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *Transactions on Machine Learning Research*, 2024.

Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Efficient and transferable adversarial examples from bayesian neural networks. In *Uncertainty in Artificial Intelligence*, pp. 738–748, 2022a.

Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *European Conference on Computer Vision*, pp. 603–618, 2022b.

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

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. (in german). *Journal für die reine und angewandte Mathematik*, pp. 210–271, 1909.

Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *International Conference on Learning Representations*, 2023.

Hong Jun Jeon and Benjamin Van Roy. An information-theoretic framework for deep learning. *Advances in Neural Information Processing Systems*, 35:3279–3291, 2022.

Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*, 2019.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning*, pp. 16049–16096. PMLR, 2023.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pp. 443–457. Springer, 2000.

Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14254–14263, 2020.

Leonid Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 2008.

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

Christoph H Lampert, Liva Ralaivola, and Alexander Zimin. Dependency-dependent bounds for sums of dependent random variables. *arXiv preprint arXiv:1811.01404*, 2018.

HO Lancaster. Correlation and complete dependence of random variables. *The Annals of Mathematical Statistics*, 34(4):1315–1321, 1963.

Yann LeCun. The mnist database of handwritten digits, 1998. URL `http://yann.lecun.com/exdb/mnist/`.

Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Making substitute models more bayesian can enhance transferability of adversarial examples. In *International Conference on Learning Representations*, 2023.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. *Advances in Neural Information Processing Systems*, 36, 2024.

Shaojie Li and Yong Liu. Towards sharper generalization bounds for structured prediction. *Advances in Neural Information Processing Systems*, 34:26844–26857, 2021.

Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 11458–11465, 2020.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Stephan Sloth Lorenzen, Christian Igel, and Mads Nielsen. Information bottleneck: Exact analysis of (quantized) neural networks. *arXiv preprint arXiv:2106.12912*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wenshuo Ma, Yidong Li, Xiaofeng Jia, and Wei Xu. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4630–4639, 2023.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Brady Neal. On the bias-variance tradeoff: Textbooks need an update. *arXiv preprint arXiv:1912.08286*, 2019.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Luis A Ortega, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11720–11743. PMLR, 2022.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pp. 8093–8104, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Albert N Shiryaev. *Probability-1*, volume 95. Springer, 2016.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

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

Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, and Heng Tao Shen. Ensemble diversity facilitates adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24377–24386, 2024.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

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

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Vladimir N. Vapnik. Statistical learning theory. *Wiley-Interscience*, 1998.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24336–24346, 2024.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1924–1933, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16158–16167, 2021.

Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *European conference on computer vision*, pp. 156–174. Springer, 2022.

Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4607–4619, 2023a.

Yilin Wang and Farzan Farnia. On the role of generalization in transferability of adversarial examples. In *Uncertainty in Artificial Intelligence*, pp. 2259–2270, 2023.

Zhiyuan Wang, Zeliang Zhang, Siyuan Liang, and Xiaosen Wang. Diversifying the high-level features for better adversarial transferability. *arXiv preprint arXiv:2304.10136*, 2023b.

Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. *arXiv preprint arXiv:2210.00706*, 2022.

Ziqiao Wang and Yongyi Mao. On f-divergence principled domain adaptation: An improved framework. *arXiv preprint arXiv:2402.01887*, 2024.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24(359): 1–49, 2023.

Han Wu, Guanyan Ou, Weibin Wu, and Zibin Zheng. Improving transferable targeted adversarial attacks with model self-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24615–24624, 2024.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Wang Xiaosen, Kangheng Tong, and Kun He. Rethinking the backward propagation for adversarial transferability. *Advances in Neural Information Processing Systems*, 36:1905–1922, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2730–2739, 2019.

Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14983–14992, 2022.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Chiu Wai Yan, Tsz-Him Cheung, and Dit-Yan Yeung. Ila-da: Improving transferability of intermediate level attack with data augmentation. In *International Conference on Learning Representations*, 2023.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655, 2021.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777, 2020.

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

Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610, 2022.

Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593, 2019.

Rui-Ray Zhang and Massih-Reza Amini. Generalization bounds for learning under graph-dependence: A survey. *Machine Learning*, 113(7):3929–3959, 2024.

Rui Ray Zhang, Xingwu Liu, Yuyi Wang, and Liwei Wang. Mcdiarmid-type inequalities for graph-dependent variables and stability bounds. *Advances in Neural Information Processing Systems*, 32, 2019a.

Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, and Hai Jin. Why does little robustness help? a further step towards understanding adversarial transferability. In *Proceedings of the 45th IEEE Symposium on Security and Privacy (S&P'24)*, volume 2, 2024a.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, pp. 7404–7413. PMLR, 2019b.

Zeliang Zhang, Rongyi Zhu, Wei Yao, Xiaosen Wang, and Chenliang Xu. Bag of tricks to boost adversarial transferability. *arXiv preprint arXiv:2401.08734*, 2024b.

Anqi Zhao, Tong Chu, Yahao Liu, Wen Li, Jingjing Li, and Lixin Duan. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8162, 2023.

Rongyi Zhu, Zeliang Zhang, Susan Liang, Zhuo Liu, and Chenliang Xu. Learning to transform dynamically for better adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24273–24283, 2024.

Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *European Conference on Computer Vision*, pp. 563–579, 2020.

Xin Zou and Weiwei Liu. Generalization bounds for adversarial contrastive learning. *Journal of Machine Learning Research*, 24(114):1–54, 2023.

CONTENTS

APPENDIX

A PROOF OF GENERALIZED RADEMACHER COMPLEXITY

A.1 PRELIMINARY

For simplicity, denote $f(\theta_i; x)$ as $f_i(x)$. For 1-Lipschitz loss function $\ell(yf(x))$ (for example, hinge loss $\ell(f(x), y) = \max(0, 1 - yf(x))$, there holds:

$$\begin{aligned}\mathcal{R}_N(\mathcal{Z}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f_i(x), y)\right] \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i y f_i(x)\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(x)\right] := \Re_N(\mathcal{Z}).\end{aligned}$$

So we can bound $\Re_N(\mathcal{Z})$ instead of $\mathcal{R}_N(\mathcal{Z})$.

A.2 LINEAR MODEL

Given Section A.1, we provide the bound below.

Lemma 3 (Linear Model). *Let $\mathcal{H} = \left\{x \mapsto w^T x\right\}$, where $x, w \in \mathbb{R}^d$. Given N classifiers from $\mathcal{H}$, assume that $\|x\|_2 \leq B$ and $\|w\|_2 \leq C$. Then*

$$\Re_N(\mathcal{Z}) \leq \frac{BC}{\sqrt{N}}.$$

Proof. We have

$$\begin{aligned}\Re_N(\mathcal{Z}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2\leq B}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_i(x)\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2\leq B}\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_i^T x\right] && (f_i(x) = w_i^T x) \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2\leq B} x^T\left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_i\right)\right] && (a^T b = b^T a) \\ &= \frac{B}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^{N}\sigma_i w_i\right\|_2 && (a^T b \leq \|a\|_2\|b\|_2) \\ &\leq \frac{B}{N}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^{N}\sigma_i w_i\right\|_2^2\right)^{\frac{1}{2}} && (\text{Jensen inequality: } \mathbb{E}x \leq \sqrt{\mathbb{E}x^2}) \\ &= \frac{B}{N}\left\{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{i=1}^{N}\sigma_i w_i^T\right)\left(\sum_{i=1}^{N}\sigma_i w_i\right)\right]\right\}^{\frac{1}{2}} \\ &= \frac{B}{N}\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\sum_{i=1}^{N}\underbrace{\sigma_i^2}_{1} w_i^T w_i + \underbrace{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sigma_i\sigma_j w_i^T w_j}_{0}\right)\right]^{\frac{1}{2}}\end{aligned}$$

$$
\begin{aligned}
&= \frac{B}{N}\left(\sum_{i=1}^{N} w_i^T w_i\right)^{\frac{1}{2}} \\
&\leq \frac{B}{N}\left(N \max \|w\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{BC}{\sqrt{N}}. && (\|w\|_2 \leq C)
\end{aligned}
$$

The proof is complete.

□

A.3 Two-layer Neural Network

Given Section A.1, we provide the bound below.

Lemma 4 (Two-layer Neural Network)**.** *Let $\mathcal{H} = \{x \mapsto w^T \phi(Ux)\}$, where $x \in \mathbb{R}^d$, $U \in \mathbb{R}^{m \times d}$, $w \in \mathbb{R}^m$, m is the number of the hidden layer, and $\phi(x) = \max(0, x)$ is the element-wise ReLU function. Given N classifiers from $\mathcal{H}$, assume that $\|x\|_2 \leq B$, $\|w\|_2 \leq B'$, and $\|U_i\|_2 \leq C$, where U_j is the j-th row of U. Then*

$$
\mathfrak{R}_N(\mathcal{Z}) \leq \frac{\sqrt{m} B B' C}{\sqrt{N}}.
$$

Proof. We have

$$
\begin{aligned}
\mathfrak{R}_N(\mathcal{Z}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f_i(x)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \frac{1}{N}\sum_{i=1}^{N} \sigma_i w_i^T \phi(U_i x)\right] && (f_i(x) = w_i^T \phi(U_i x)) \\
&= \frac{B'}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \left\|\sum_{i=1}^{N} \sigma_i \phi(U_i x)\right\|_2\right] && (\|w\|_2 \leq B') \\
&= \frac{B'}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \left\|\sum_{i=1}^{N} \sigma_i V_i\right\|_2\right] && \left(\text{Denote } V_i = \begin{bmatrix} \phi(U_{1i}x) \\ \vdots \\ \phi(U_{mi}x) \end{bmatrix} \in \mathbb{R}^m\right) \\
&= \frac{B'}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \sqrt{\left(\sum_{i=1}^{N} \sigma_i V_i^T\right)\left(\sum_{i=1}^{N} \sigma_i V_i\right)}\right] \\
&= \frac{B'}{N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|x\|_2 \leq B} \left(\sum_{i=1}^{N} \underbrace{\sigma_i^2}_{1} V_i^T V_i + \underbrace{\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \sigma_i \sigma_j V_i^T V_j}_{0}\right)^{\frac{1}{2}}\right] \\
&= \frac{B'}{N}\sup_{\|x\|_2 \leq B}\left(\sum_{i=1}^{N} V_i^T V_i\right)^{\frac{1}{2}} \\
&\leq \frac{B'}{N}\sup_{\|x\|_2 \leq B}\left(N \max_i \|V_i\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{B'}{\sqrt{N}}\sup_{\|x\|_2 \leq B}\left(\max_i \|V_i\|_2\right)
\end{aligned}
$$

For $V_i = \begin{bmatrix} \phi(U_{1i}x) \\ \vdots \\ \phi(U_{mi}x) \end{bmatrix} \in \mathbb{R}^m$, we have

$$\begin{aligned}
\sup_{\|x\|_2 \le B} \left(\max_i \|V_i\|_2 \right) &= \sup_{\|x\|_2 \le B} \left(\max_i \left\| \begin{bmatrix} \phi(U_{1i}x) \\ \vdots \\ \phi(U_{mi}x) \end{bmatrix} \right\|_2 \right) \\
&\le \sup_{\|x\|_2 \le B} \left(\max_i \left\| \begin{bmatrix} U_{1i}x \\ \vdots \\ U_{mi}x \end{bmatrix} \right\|_2 \right) && (|\phi(x)| \le |x|) \\
&\le \sqrt{m} \sup_{\|x\|_2 \le B} \left(\max_i \max_j \|U_{ji}x\|_2 \right) \\
&\le \sqrt{m} \sup_{\|x\|_2 \le B} \left(\max_i \max_j \|U_{ji}\|_2 \|x\|_2 \right) \\
&= \sqrt{m}BC && (\|x\|_2 \le B \text{ and } \|U_{ji}\|_2 \le C)
\end{aligned}$$

Finally,

$$\Re_N(\mathcal{Z}) \le \frac{B'}{\sqrt{N}} \sup_{\|x\|_2 \le B} \left(\max_i \|V_i\|_2 \right) \le \frac{\sqrt{m}BB'C}{\sqrt{N}}$$

The proof is complete.

□

A.4 Proof of Lemma 2

For simplicity, denote $f(\theta_i; x)$ as $f_i(x)$ and $i \in \{1, \cdots, N\}$ as $i \in [N]$.

First, we begin with a lemma, which is a similar version of Lemma 1 from (Golowich et al., 2018).

Lemma 5. *Let ϕ be a 1-Lipschitz, positive-homogeneous activation function which is applied element-wise (such as the ReLU). Then for any class of vector-valued functions $\mathcal{F}$, any convex and monotonically increasing function $g : \mathbb{R} \to [0, \infty)$ and $R \in \mathbb{R}_+$, there holds:*

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}, W : \|W\|_F \le R} g\left(\left\| \sum_{i=1}^N \sigma_i \phi\left(W f_i\left(x\right)\right) \right\| \right) \le 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} g\left(R \cdot \left\| \sum_{i=1}^N \sigma_i f_i\left(x\right) \right\| \right) \tag{13}$$

Proof. Let $w_1, \cdots, w_h$ be the rows of W, we have

$$\begin{aligned}
\left\| \sum_{i=1}^N \sigma_i \phi\left(W f_i\left(x\right)\right) \right\|^2 &= \sum_{j=1}^h \left[\sum_{i=1}^N \sigma_i \phi(w_j f_i(x)) \right]^2 \\
&= \sum_{j=1}^h \|w_j\|^2 \left[\sum_{i=1}^N \sigma_i \phi\left(\frac{w_j^\top}{\|w_j\|} f_i\left(x\right) \right) \right]^2 && (\phi(ax) = a\phi(x))
\end{aligned}$$

Therefore, the supremum of this over all $w_1, \cdots, w_h$ such that $\|W\|_F^2 = \sum_{j=1}^h \|w_j\|^2 \le R^2$ must be attained when $\|w_j\| = R$ for some j and $\|w_i\| = 0$ for all $i \ne j$. So we have

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}, W : \|W\|_F \le R} g\left(\left\| \sum_{i=1}^N \sigma_i \phi\left(W f_i\left(x\right)\right) \right\| \right) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}, w : \|w\| = R} g\left(\left| \sum_{i=1}^N \sigma_i \phi\left(w^\top f_i\left(x\right)\right) \right| \right).$$

Since $g(|z|) \le g(z) + g(-z)$, this can be upper bounded by

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup g\left(\sum_{i=1}^{N} \sigma_i \phi\left(w^\top f_i\left(x\right)\right)\right) + \mathbb{E}_{\boldsymbol{\sigma}} \sup g\left(-\sum_{i=1}^{N} \sigma_i \phi\left(w^\top f_i\left(x\right)\right)\right)$$
$$= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup g\left(\sum_{i=1}^{N} \sigma_i \phi\left(w^\top f_i\left(x\right)\right)\right),$$

where the equality follows from the symmetry in the distribution of the σ_i random variables. The right hand side in turn can be upper bounded by

$$2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}, w: \|w\| = R} g\left(\sum_{i=1}^{N} \sigma_i w^\top f_i\left(x\right)\right) \leq 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}, w: \|w\| = R} g\left(\|w\| \left\|\sum_{i=1}^{N} \sigma_i f_i\left(x\right)\right\|\right)$$
$$= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} g\left(R \cdot \left\|\sum_{i=1}^{N} \sigma_i f_i\left(x\right)\right\|\right).$$

□

With this lemma in hand, we can prove lemma 2:

Proof. For $\lambda > 0$, the rademacher complexity can be upper bounded as

$$N\mathfrak{R}_N\left(\mathcal{Z}\right) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f_1, \cdots, f_n} \sum_{i=1}^{N} \sigma_i f_i(x)$$
$$\leq \frac{1}{\lambda} \log \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(\lambda \sum_{i=1}^{N} \sigma_i f_i(x)\right) \qquad \text{(Jensen's inequality)}$$
$$\leq \frac{1}{\lambda} \log \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(\underbrace{\sup_{i \in [n]} \|W_{i,l}\|_F}_{T_l} \left\|\lambda \sum_{i=1}^{N} \sigma_i \phi_{l-1} \underbrace{\left(W_{i,l-1} \phi_{l-2}\left(\ldots \phi_1\left(W_{i,1} x\right)\right)\right)}_{f_{i,l-1}(x)}\right\|\right)$$

We write this last expression as

$$\frac{1}{\lambda} \log \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(T_l \cdot \lambda \left\|\sum_{i=1}^{N} \sigma_i \phi_{l-1}\left(f_{i,l-1}(x)\right)\right\|\right)$$
$$\leq \frac{1}{\lambda} \log\left(2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(T_l \cdot T_{l-1} \cdot \lambda \left\|\sum_{i=1}^{N} \sigma_i f_{i,l-2}\left(x\right)\right\|\right)\right) \qquad \text{(Lemma 5)}$$
$$\leq \cdots \qquad \text{(Repeatedly apply Lemma 5)}$$
$$\leq \frac{1}{\lambda} \log\left(2^{l-2} \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(\lambda \cdot \prod_{i=1}^{l-1} T_i \cdot \left\|\sum_{i=1}^{N} \sigma_i \phi_1(W_{i,1} x)\right\|\right)\right)$$
$$\leq \frac{1}{\lambda} \log\left(2^{l-1} \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(\lambda \cdot \prod_{i=1}^{l-1} T_i \cdot \left\|\sum_{i=1}^{N} \sigma_i W_{i,1} x\right\|\right)\right)$$

Assume that $W_{i,1}^*, i \in [N]$ maximizes

$$\sup \exp\left(\lambda \cdot \prod_{i=1}^{l-1} T_i \cdot \left\|\sum_{i=1}^{N} \sigma_i W_{i,1} x\right\|\right).$$

Therefore,

$$\frac{1}{\lambda} \log\left(2^{l-1} \cdot \mathbb{E}_{\boldsymbol{\sigma}} \sup \exp\left(\lambda \cdot \prod_{i=1}^{l-1} T_i \cdot \left\|\sum_{i=1}^{N} \sigma_i W_{i,1} x\right\|\right)\right)$$

$$
\begin{aligned}
=&\frac{1}{\lambda}\log\left(2^{l-1}\cdot\mathbb{E}_{\boldsymbol{\sigma}}\exp\left(\lambda\cdot\underbrace{\prod_{i=1}^{l-1}T_i\cdot\left\|\sum_{i=1}^{N}\sigma_i W_{i,1}^* x\right\|}_{Z}\right)\right)\\
=&\frac{1}{\lambda}\log\left(2^{l-1}\cdot\mathbb{E}_{\boldsymbol{\sigma}}\exp\left(\lambda Z\right)\right)\\
=&\frac{(l-1)\log(2)}{\lambda}+\frac{1}{\lambda}\log\left\{\mathbb{E}_{\boldsymbol{\sigma}}\exp\left(\lambda Z\right)\right\}\\
=&\frac{(l-1)\log(2)}{\lambda}+\frac{1}{\lambda}\log\{\mathbb{E}\exp\lambda(Z-\mathbb{E}Z)\}+\mathbb{E}Z
\end{aligned}
$$

For $\mathbb{E}Z$, we have

$$
\begin{aligned}
\mathbb{E}Z=&\prod_{i=1}^{l-1}T_i\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{N}\sigma_i W_{i,1}^* x\right\|^2\right]}\\
=&\prod_{i=1}^{l-1}T_i\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i=j}^{N}\sigma_i\sigma_j\left(W_{i,1}^* x\right)^T\left(W_{j,1}^* x\right)\right]}\\
\leq&\prod_{i=1}^{l-1}T_i\left(T_1 B\sqrt{N}\right)\\
=&B\sqrt{N}\prod_{i=1}^{l}T_i
\end{aligned}
$$

Note that Z is a deterministic function of the *i.i.d.* random variables $\sigma_1,\cdots,\sigma_N$, and satisfies

$$
Z(\sigma_1,\cdots,\sigma_i,\cdots,\sigma_N)-Z(\sigma_1,\cdots,-\sigma_i,\cdots,\sigma_N)\leq 2B\underbrace{\prod_{i=1}^{l}T_i}_{T}.
$$

This means that Z satisfies a bounded-difference condition. According to Theorem 6.2 in Boucheron et al. (2013), Z is sub-Gaussian with variance factor

$$
\frac{1}{4}\sum_{i=1}^{N}(2BT)^2=NB^2T^2,
$$

and satisfies

$$
\frac{1}{\lambda}\log\{\mathbb{E}\exp\lambda(Z-\mathbb{E}Z)\}\leq\frac{1}{\lambda}\cdot\frac{\lambda^2}{2}NB^2T^2=\frac{\lambda}{2}NB^2T^2.
$$

Choosing $\lambda=\frac{\sqrt{2\log(2)l}}{BT\sqrt{N}}$ and using the above, we get that

$$
\frac{(l-1)\log(2)}{\lambda}+\frac{1}{\lambda}\log\{\mathbb{E}\exp\lambda(Z-\mathbb{E}Z)\}+\mathbb{E}Z\leq\left(\sqrt{(2\log 2)l}+1\right)BT\sqrt{N}
$$

Finally, we get

$$
\Re_N\left(\mathcal{Z}\right)\leq\frac{\left(\sqrt{(2\log 2)l}+1\right)BT}{\sqrt{N}}
$$

The proof is complete.

□

B PROOF OF TRANSFERABILITY ERROR

B.1 TRANSFERABILITY ERROR AND GENERALIZATION ERROR

For $z = (x, y)$, there holds

$$\begin{aligned} TE(z) = L_P(z^*) - L_P(z) &\leq L_P(z^*) - L_P(z) + (L_E(z) - L_E(z^*)) \\ &= (L_P(z^*) - L_E(z^*)) + (L_E(z) - L_P(z)) \\ &\leq \sup_{x \in \mathcal{B}_\epsilon(x)} (L_P(z) - L_E(z)) + \sup_{x \in \mathcal{B}_\epsilon(x)} (L_E(z) - L_P(z)) \\ &\leq \sup_{z \in \mathcal{Z}} (L_P(z) - L_E(z)) + \sup_{z \in \mathcal{Z}} (L_E(z) - L_P(z)). \\ &\leq 2 \sup_{z \in \mathcal{Z}} |L_P(z) - L_E(z)| . \end{aligned}$$

B.2 PROOF OF THEOREM 1

We prove a general version of the theorem as follows:

Theorem 3. *Consider the squared error loss $l(\theta, x, y) = [f(\theta; x) - y]^2$ for a data point $z = (x, y)$. Assume that the data is generated by a function $g(x)$ such that $y = g(x) + \rho$, where the zero-mean noise ρ has a variance of η^2 and is independent of x. Then there holds*

$$TE(z, \epsilon) = L_P(z^*) - \eta^2 - \underbrace{Var_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)}_{\textit{Diversity}} - \underbrace{[g(x) - \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)]^2}_{\textit{Attack}}. \tag{14}$$

Remark. *The irreducible error η^2 is constant because it arises from inherent noise and randomness in the data (Geman et al., 1992).*

Now we start our proof of it.

Proof. Given Eq. (5), it is equivalent to prove

$$L_P(z) = Var_\theta f(\theta; x) + [g(x) - \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)]^2 + \eta^2. \tag{15}$$

Note that

$$\begin{aligned} L_P(z) &= \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} [f(\theta; x) - y]^2 \\ &= \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} [f(\theta; x) - g(x) + g(x) - y]^2 \\ &= \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} \left[(f(\theta; x) - g(x))^2 + (g(x) - y)^2 + 2(g(x) - y)(f(\theta; x) - g(x)) \right]. \end{aligned}$$

Recall that $y = g(x) + \rho$ with $\mathbb{E}(\rho) = 0$ and $Var(\rho) = \eta^2$, we have

$$\mathbb{E}_{\theta \sim \mathcal{P}_\Theta} (g(x) - y)^2 = \eta^2,$$

and

$$\mathbb{E}_{\theta \sim \mathcal{P}_\Theta} [2(g(x) - y)(f(\theta; x) - g(x))] = -2\mathbb{E}(\rho)\mathbb{E}_{\theta \sim \mathcal{P}_\Theta} [f(\theta; x) - g(x)] = 0.$$

Therefore,

$$L_P(z) = \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} [f(\theta; x) - g(x)]^2 + \eta^2. \tag{16}$$

Likewise, we decompose the first term as

$$\begin{aligned} &\mathbb{E}_\theta [f(\theta; x) - g(x)]^2 \\ =&\mathbb{E}_\theta [f(\theta; x) - \mathbb{E}_\theta f(\theta; x) + \mathbb{E}_\theta f(\theta; x) - g(x)]^2 \\ =&\mathbb{E}_\theta \left[(f(\theta; x) - \mathbb{E}_\theta f(\theta; x))^2 + (\mathbb{E}_\theta f(\theta; x) - g(x))^2 \right. \\ &\left. - 2(f(\theta; x) - \mathbb{E}_\theta f(\theta; x))(\mathbb{E}_\theta f(\theta; x) - g(x)) \right] \\ =&\underbrace{\mathbb{E}_\theta (f(\theta; x) - \mathbb{E}_\theta f(\theta; x))^2}_{Var_\theta f(\theta; x)} + \underbrace{\mathbb{E}_\theta (\mathbb{E}_\theta f(\theta; x) - g(x))^2}_{(g(x) - \mathbb{E}_\theta (f(\theta; x))^2} \end{aligned}$$

$$- 2\underbrace{\mathbb{E}_\theta \left[f(\theta;x) - \mathbb{E}_\theta f(\theta;x))(\mathbb{E}_\theta f(\theta;x) - g(x)\right]}_{0},$$

with the derivations for the second and third term:

$$\begin{aligned}\mathbb{E}_\theta(f(\theta;x) - \mathbb{E}_\theta f(\theta;x))^2 &= \left(\mathbb{E}_\theta f(\theta;x)\right)^2 - 2g(x)\mathbb{E}_\theta f(\theta;x) + g^2(x) \\ &= \left(g(x) - \mathbb{E}_\theta(f(\theta;x)\right)^2,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_\theta \left[f(\theta;x) - \mathbb{E}_\theta f(\theta;x))(\mathbb{E}_\theta f(\theta;x) - g(x)\right] \\ =&(\mathbb{E}_\theta f(\theta;x))^2 - g(x)\mathbb{E}_\theta f(\theta;x) - (\mathbb{E}_\theta f(\theta;x))^2 + g(x)\mathbb{E}_\theta f(\theta;x) \\ =&0.\end{aligned}$$

As a result,

$$\mathbb{E}_\theta \left[f(\theta;x) - g(x)\right]^2 = \mathit{Var}_\theta f(\theta;x) + \left[g(x) - \mathbb{E}_{\theta\sim\mathcal{P}_\Theta} f(\theta;x)\right]^2. \tag{17}$$

Combining the above results and we complete the proof.

□

To prove Theorem 1, we just set $\rho = 0$ in the above general version of theorem.

Similarly, consider the empirical version of Theorem 1, we decompose $L_E(z)$ as follows:

Theorem 4 (Vulnerability-diversity Decomposition (empirical version))**.** *Consider the squared error loss $l(f(\theta;x),y) = [f(\theta;x) - y]^2$ for a data point $z = (x,y)$. Let $\hat{f}(\theta;x) = \frac{1}{N}\sum_{i=1}^N f(\theta_i;x)$ be the expectation of prediction over the distribution on the parameter space. Then there holds*

$$\begin{aligned}L_E(z) &= \frac{1}{N}\sum_{i=1}^N \ell(f(\theta_i;x),y) \\ &= \underbrace{l(\hat{f}(\theta;x),y)}_{\textit{Vulnerability}} + \underbrace{\frac{1}{N}\sum_{j=1}^N \left(f(\theta_i;x) - \frac{1}{N}\sum_{j=1}^N f(\theta_i;x)\right)^2}_{\textit{Diversity}}.\end{aligned}$$

The proof is similar to the above:

$$\begin{aligned}L_E(z) &= \frac{1}{N}\sum_{i=1}^N \left(f(\theta_i;x) - y\right)^2 \\ &= \frac{1}{N}\frac{1}{N}\sum_{i=1}^N \left(f(\theta_i;x) - \frac{1}{N}\sum_{i=1}^N f(\theta_i;x) + \frac{1}{N}\sum_{i=1}^N f(\theta_i;x) - y\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^N \left[\left(f(\theta_i;x) - \frac{1}{N}\sum_{i=1}^N f(\theta_i;x)\right)^2 + \left(\frac{1}{N}\sum_{i=1}^N f(\theta_i;x) - y\right)^2 + \right. \\ &\left. 2\left(f(\theta_i;x) - \frac{1}{N}\sum_{i=1}^N f(\theta_i;x)\right)\left(\frac{1}{N}\sum_{i=1}^N f(\theta_i;x) - y\right)\right] \\ &= \underbrace{l(\hat{f}(\theta;x),y)}_{\text{Vulnerability}} + \underbrace{\frac{1}{N}\sum_{j=1}^N \left(f(\theta_i;x) - \frac{1}{N}\sum_{j=1}^N f(\theta_i;x)\right)^2}_{\text{Diversity}} + \\ &\frac{2}{N}\sum_{i=1}^N \left(f(\theta_i;x) - \frac{1}{N}\sum_{i=1}^N f(\theta_i;x)\right)\left(\frac{1}{N}\sum_{i=1}^N f(\theta_i;x) - y\right).\end{aligned}$$

The last terms equals to 0 because

$$\begin{aligned}
&\sum_{i=1}^{N}\left(f(\theta_i;x)-\frac{1}{N}\sum_{i=1}^{N}f(\theta_i;x)\right)\left(\frac{1}{N}\sum_{i=1}^{N}f(\theta_i;x)-y\right)\\
=&\frac{1}{N}\left(\sum_{i=1}^{N}f(\theta_i;x)\right)^2-y\sum_{i=1}^{N}f(\theta_i;x)-\frac{1}{N}\left(\sum_{i=1}^{N}f(\theta_i;x)\right)^2+y\sum_{i=1}^{N}f(\theta_i;x)\\
=&0.
\end{aligned}$$

The proof is complete.

B.3 Proof of Theorem 2

We first define a divergence measure taken into account. Given a measurable space and two measures μ, ν which render it a measure space, we denote $\nu \ll \mu$ if ν is absolutely continuous with respect to μ. Hellinger integrals are defined below:

Definition 4 (Hellinger integrals (Hellinger, 1909))**.** *Let ν, μ be two probability measures on $(\Omega, \mathcal{F})$ and satisfy $\nu \ll \mu$, and $\varphi_\alpha : \mathbb{R}^+ \to \mathbb{R}$ be defined as $\varphi_\alpha(x) = x^\alpha$. Then the Hellinger integral of order α is given by*

$$H_\alpha(\nu\|\mu) = \int \left(\frac{d\nu}{d\mu}\right)^\alpha \mathrm{d}\mu.$$

It can be seen as a ϕ-Divergence with a specific parametrised choice of ϕ (Liese & Vajda, 2006). For $\alpha > 1$, the Hellinger integral measures the divergence between two probability distributions (Liese & Vajda, 2006). There holds $H_\alpha(\nu\|\mu) \in [1, +\infty), \alpha > 1$, and it equals to 1 if the two measures coincide (Shiryaev, 2016). Given such a divergence measure, we now provide the proof.

Proof. From Section B.1, we know that

$$\begin{aligned}
TE(z) = L_P(z^*) - L_P(z) &\le L_P(z^*) - L_P(z) + (L_E(z) - L_E(z^*))\\
&= (L_P(z^*) - L_E(z^*)) + (L_E(z) - L_P(z))\\
&\le \sup_{x\in\mathcal{B}_\epsilon(x)}(L_P(z) - L_E(z)) + \sup_{x\in\mathcal{B}_\epsilon(x)}(L_E(z) - L_P(z))\\
&\le \sup_{z\in\mathcal{Z}}(L_P(z) - L_E(z)) + \sup_{z\in\mathcal{Z}}(L_E(z) - L_P(z)).
\end{aligned}$$

Let $(\theta'_1, \ldots, \theta'_N) \sim \mathcal{P}'_{\Theta^N}$, where $\mathcal{P}'_{\Theta^N}$ be a distribution over the product space, and the *m*-th member is different from $\mathcal{P}_{\Theta^N}$, i.e., $(\theta'_1, \ldots, \theta'_m, \cdots, \theta'_N) = (\theta_1, \ldots, \theta'_m, \cdots, \theta_N)$, where $\theta'_m \neq \theta_m$. The training process of N surrogate models $f(\theta'_1), \cdots, f(\theta'_N)$ can be viewed as sampling the parameter sets $(\theta'_1, \ldots, \theta'_N)$ from the distribution $\mathcal{P}'_{\Theta^N}$.

We define

$$L_{E'}(z) = \frac{1}{N}\sum_{i=1}^{N}\ell(f(\theta'_i;x), y),$$

and

$$\begin{aligned}
\Phi_1(E) &= \sup_{z\in\mathcal{Z}}\left\{L_P(z) - L_E(z)\right\},\\
\Phi_1(E') &= \sup_{z\in\mathcal{Z}}\left\{L_P(z) - L_{E'}(z)\right\}.
\end{aligned}$$

We have

$$\begin{aligned}
\Phi_1(E) - \Phi_1(E') &= \sup_{z\in\mathcal{Z}}\left\{L_P(z) - L_E(z)\right\} - \sup_{z\in\mathcal{Z}}\left\{L_P(z) - L_{E'}(z)\right\}\\
&\le \sup_{z\in\mathcal{Z}}\left\{L_P(z) - L_E(z) - (L_P(z) - L_{E'}(z))\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{z\in\mathcal{Z}} \{L_{E'}(z) - L_E(z)\} \\
&= \frac{1}{N} \sup_{z\in\mathcal{Z}} \left[\sum_{i=1}^{N} \ell(f(\theta_i'; x), y) - \sum_{i=1}^{N} \ell(f(\theta_i; x), y) \right].
\end{aligned}
$$

By assuming that loss function ℓ is bounded by β, we have

$$
|\Phi_1(E) - \Phi_1(E')| \leq \frac{\beta}{N}.
$$

According to Theorem 1 in Esposito & Mondelli (2024), for all $\delta \in (0,1)$ and $\alpha > 1$, with probability at least $1-\delta$, we have

$$
\Phi_1(E) \leq \mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_1(E)] + \sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N} \ln \frac{2^{\frac{\alpha-1}{\alpha}} H_\alpha^{\frac{1}{\alpha}} \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)}{\delta}}. \tag{18}
$$

Denote $f(\theta_i; x)$ as $f_i(x)$ and $f(\theta_i'; x)$ as $f_i'(x)$. Then we estimate the upper bound of $\mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_1(E)]$ as follows:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_1(E)] &= \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left[\sup_{z\in\mathcal{Z}} (L_P(z) - L_E(z)) \right] \\
&= \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left[\sup_{z\in\mathcal{Z}} \mathbb{E}_{(\theta_1', \cdots, \theta_N') \sim \mathcal{P}'_{\Theta^N}} (L_{E'}(z) - L_E(z)) \right] \\
&\leq \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left[\sup_{z\in\mathcal{Z}} (L_{E'}(z) - L_E(z)) \right] \qquad \text{(Jensen inequality)} \\
&= \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \ell(f(\theta_i'; x), y) - \sum_{i=1}^{N} \ell(f(\theta_i; x), y) \right] \right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \left[\ell(f_i'(x), y) - \ell(f_i(x), y) \right] \right] \right\} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}'_{\Theta^N}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \ell(f_i'(x), y) \right] \right\} + \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right] \right\} \\
&= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right\} \\
&= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{z\in\mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right\} \\
&= 2\mathcal{R}_N(\mathcal{F}).
\end{aligned}
$$

Likewise, if we define

$$
\begin{aligned}
\Phi_2(E) &= \sup_{z\in\mathcal{Z}} \{L_E(z) - L_P(z)\}, \\
\Phi_2(E') &= \sup_{z\in\mathcal{Z}} \{L_{E'}(z) - L_P(z)\},
\end{aligned}
$$

then we have

$$
\begin{aligned}
\Phi_2(E) - \Phi_2(E') &= \sup_{z\in\mathcal{Z}} \{L_E(z) - L_P(z)\} - \sup_{z\in\mathcal{Z}} \{L_{E'}(z) - L_P(z)\} \\
&\leq \sup_{z\in\mathcal{Z}} \{L_E(z) - L_P(z) - (L_{E'}(z) - L_P(z))\} \\
&= \sup_{z\in\mathcal{Z}} \{L_E(z) - L_{E'}(z)\}
\end{aligned}
$$

$$= \frac{1}{N} \sup_{z \in \mathcal{Z}} \left[\sum_{i=1}^{N} \ell(f(\theta_i; x), y) - \sum_{i=1}^{N} \ell(f(\theta_i'; x), y) \right].$$

According to the assumption that loss function ℓ is bounded by β, we have

$$|\Phi_2(E) - \Phi_2(E')| \leq \frac{\beta}{N}.$$

According to Theorem 1 in Esposito & Mondelli (2024), for all $\delta \in (0, 1)$ and $\alpha > 1$, with probability at least $1 - \delta$, we have

$$\Phi_2(E) \leq \mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_2(E)] + \sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N} \ln \frac{2^{\frac{\alpha-1}{\alpha}} H_\alpha^{\frac{1}{\alpha}} \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta_i} \right)}{\delta}}. \tag{19}$$

We estimate the upper bound of $\mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_2(E)]$ as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_{\Theta^N}}[\Phi_2(E)] &= \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left[\sup_{z \in \mathcal{Z}} (L_E(z) - L_P(z)) \right] \\
&= \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left[\sup_{z \in \mathcal{Z}} \mathbb{E}_{(\theta_1', \cdots, \theta_N') \sim \mathcal{P}'_{\Theta^N}} (L_E(z) - L_{E'}(z)) \right] \\
&\leq \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left[\sup_{z \in \mathcal{Z}} (L_E(z) - L_{E'}(z)) \right] \qquad \text{(Jensen inequality)} \\
&= \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \ell(f(\theta_i; x), y) - \sum_{i=1}^{N} \ell(f(\theta_i'; x), y) \right] \right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}, \mathcal{P}'_{\Theta^N}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \left[\ell(f_i(x), y) - \ell(f_i'(x), y) \right] \right] \right\} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}'_{\Theta^N}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \ell(f_i'(x), y) \right] \right\} + \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \left[\sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right] \right\} \\
&= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\mathcal{P}_{\Theta^N}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right\} \\
&= 2 \cdot \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(f_i(x), y) \right\} \\
&= 2\mathcal{R}_N(\mathcal{F}).
\end{aligned}$$

Therefore, with probability at least $1 - \delta$, there holds

$$TE(z, \epsilon) = \Phi_1(E) + \Phi_2(E) \leq 4\mathcal{R}_N(\mathcal{F}) + \sqrt{\frac{2\alpha\beta^2}{(\alpha-1)N} \ln \frac{2^{\frac{\alpha-1}{\alpha}} H_\alpha^{\frac{1}{\alpha}} \left(\mathcal{P}_{X^n} \| \mathcal{P}_{\bigotimes_{i=1}^n X_i} \right)}{\delta}}.$$

The proof is complete.

□

C More Related Work

C.1 Transferable Adversarial Attack

Input transformation. Input transformation-based attacks have shown great effectiveness in improving transferability and can be combined with gradient-based attacks. Most input transformation

techniques rely on the fundamental idea of applying data augmentation strategies to prevent overfitting to the surrogate model (Gu et al., 2024). Such methods adopt various input transformations to further improve the transferability of adversarial examples (Wang et al., 2023b;a). For instance, random resizing and padding (Xie et al., 2019), downscaling (Lin et al., 2019), mixing (Wang et al., 2021), automated data augmentation (Yan et al., 2023), block shuffle and rotation (Wang et al., 2024), and dynamical transformation (Zhu et al., 2024).

Gradient-based optimization. The central concept of these methods is to develop optimization techniques in the generation of adversarial examples to achieve better transferability. Dong et al. (2018); Lin et al. (2019); Wang & He (2021) draw an analogy between generating adversarial examples and the model training process. Therefore, conventional optimization methods that improve model generalization can also benefit adversarial transferability. In gradient-based optimization methods, adversarial perturbations are directly optimized based on one or more surrogate models during inference. Some popular ideas include applying momentum (Dong et al., 2018), Nesterov accelerated gradient (Lin et al., 2019), scheduled step size (Gao et al., 2020) and gradient variance reduction (Wang & He, 2021; Xiong et al., 2022). There are also other elegantly designed techniques in recent years (Gubri et al., 2022b; Wang et al., 2022; Xiaosen et al., 2023; Li et al., 2024; Wu et al., 2024; Zhang et al., 2024b), such as collecting weights (Gubri et al., 2022b), modifying gradient calculation (Xiaosen et al., 2023) and applying integrated gradients (Ma et al., 2023).

Model ensemble attack. Motivated by the use of model ensembles in machine learning, researchers have developed diverse ensemble attack strategies to obtain transferable adversarial examples (Gu et al., 2024). It is a powerful attack that employs an ensemble of models to simultaneously generate adversarial samples. It can not only integrate with advanced gradient-based optimization methods, but also harness the unique strengths of each individual model (Tang et al., 2024). Some popular ensemble paradigms include loss-based ensemble (Dong et al., 2018), prediction-based (Liu et al., 2017), logit-based ensemble (Dong et al., 2018), and longitudinal strategy (Li et al., 2020). There is also some deep analysis to compare these ensemble paradigms (Zhang et al., 2024b). Moreover, advanced ensemble algorithms have been created to ensure better adversarial transferability (Zou et al., 2020; Gubri et al., 2022a; Xiong et al., 2022; Chen et al., 2023; Li et al., 2023; Wu et al., 2024; Chen et al., 2024b).

C.2 Statistical Learning Theory

Statistical learning theory forms the theoretical backbone of modern machine learning by providing rigorous frameworks for understanding model generalization (Vapnik, 1999). It introduces foundational concepts such as Rademacher complexity (Bartlett & Mendelson, 2002), VC dimension (Vapnik & Chervonenkis, 1971), structural risk minimization (Vapnik, 1998) . It has also been instrumental in the development of Support Vector Machines (Cortes & Vapnik, 1995) and kernel methods (Shawe-Taylor & Cristianini, 2004), which remain pivotal in supervised learning tasks. Recent advances extend statistical learning theory to deep learning, addressing challenges of high-dimensional data and model complexity (Bartlett et al., 2021). These contributions have significantly enhanced the capability to design robust learning algorithms that generalize well across diverse applications (Du & Swamy, 2013). In addition, there are also some other novel theoretical frameworks, such as information-theoretic analysis (Xu & Raginsky, 2017), PAC-Bayes bounds (Parrado-Hernández et al., 2012), transductive learning (Vapnik, 2006), and stability analysis (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010). Most of them derive a bound of the order $\mathcal{O}(\frac{1}{\sqrt{M}})$, while some others derive sharper bound of generalization (Li & Liu, 2021) of the order $\mathcal{O}(\frac{1}{M})$. Such theoretical analysis suggests that with the increase of the dataset volume, the model generalization will become better.

D Further Discussion

D.1 Analyze Empirical Model Ensemble Rademacher Complexity

In particular, we present detailed analysis for the simple and complex cases below, within the context of transferable model ensemble attack.

The simple input space. Firstly, consider the trivial case where the input space contains too simple examples so that all classifiers correctly classify $(x, y) \in \mathcal{Z}$. Then there holds

$$\mathcal{R}_N(\mathcal{Z}) = \ell(y, y) \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\frac{1}{N} \sum_{i=1}^{N} \sigma_i \right] = 0.$$

In this case, $\mathcal{Z}$ is simple enough for $f_1, \cdots, f_N$. Such $\mathcal{Z}$ corresponds to a $\mathcal{R}_N(\mathcal{Z})$ close to 0. However, it is important to note that an overly simplistic space $\mathcal{Z}$ may be impractical for model ensemble attack: the adversarial examples in such a space may not successfully attack the models from D, leading to a small value of $L_P(z^*)$. In other words, the existence of transferable adversarial examples implicitly imposes constraints on the minimum complexity of $\mathcal{Z}$.

The complex input space. Secondly, we consider the complex case. In particular, given arbitrarily N models in $\mathcal{H}$ and any assignment of $\boldsymbol{\sigma}$, a sufficiently complex $\mathcal{Z}$ contains all kinds of examples that make $\mathcal{R}_N(\mathcal{Z})$ large: (1) If $\sigma_i = +1$, there are adversarial examples that can successfully attack f_i and leads to a large $\sigma_i \ell(f_i(x), y)$; (2) If $\sigma_i = -1$, there exists some examples that can be correctly classified by f_i, leading to $\sigma_i \ell(f_i(x), y) = 0$. However, such a large $\mathcal{R}_N(\mathcal{Z})$ is also not appropriate for transferable model ensemble attack. It may include adversarial examples that perform well against $f_1, \cdots, f_N$ but are merely overfitted to the current N surrogate models (Rice et al., 2020; Yu et al., 2022). In other words, these examples might not effectively attack other models in $\mathcal{H}$, thereby limiting their adversarial transferability.

The above analysis suggests that an excessively large or small $\mathcal{R}_N(\mathcal{Z})$ is not suitable for adversarial transferability. So we are curious to investigate the correlation between $\mathcal{R}_N(\mathcal{Z})$ and adversarial transferability, which comes to the analysis about the general case in Section 3.4.

Explain robust overfitting. After a certain point in adversarial training, continued training significantly reduces the robust training loss of the classifier while increasing the robust test loss, a phenomenon known as robust overfitting (Rice et al., 2020; Yu et al., 2022) (also linked to robust generalization (Schmidt et al., 2018; Yin et al., 2019)). From the perspective in Section 3.4, the cause of this overfitting is the *limited complexity of the input space relative to the classifier* used to generate adversarial examples during training. The adversarial examples become too simple for the model, leading to overfitting. To mitigate this, we could consider generating more "hard" and "generalizable" adversarial examples to improve the model's generalization in adversarial training. For a less transferable adversarial example (x, y), it is associated with a small $L_P(z)$, which in turn makes $TE(z, \epsilon)$ large.

D.2 Other Opinions on "Diversity"

D.2.1 Other Definitions

In Yang et al. (2021), gradient diversity is defined using the cosine similarity of gradients between different models, and instance-level transferability is introduced, along with a bound for transferability. This work cleverly uses Taylor expansion to establish a theoretical connection between the success probability of attacking a single sample and the gradients of the models. In Kariyappa & Qureshi (2019), inspired by the concept of adversarial subspace (Tramèr et al., 2017), diversity is defined based on the cosine similarity of gradients across different models. The authors aim to encourage models to become more diverse, thereby achieving "no overlap in the adversarial subspaces," and provide intuitive insights to readers. Both papers define gradient diversity and explain its impact.

In contrast, our definition of diversity stems from the unified theoretical framework proposed in this paper. Specifically:

- We draw inspiration from statistical learning theory (Shalev-Shwartz et al., 2010; Bartlett & Mendelson, 2002) on generalization, defining transferability error accordingly.
- Additionally, we are motivated by ensemble learning (Abe et al., 2023; Wood et al., 2023), where we define diversity as the variation in outputs among different ensemble models.
- Intuitively, when different models exhibit significant differences in their outputs for the same sample, their gradient differences during training are likely substantial as well. This suggests a potential connection between our output-based definition of diversity and the

gradient-based definitions in (Shalev-Shwartz et al., 2010; Bartlett & Mendelson, 2002), which is worth exploring in future research.

Overall, our perspective differs from that of Shalev-Shwartz et al. (2010); Bartlett & Mendelson (2002). However, despite the differences in definitions, both our work and Shalev-Shwartz et al. (2010); Bartlett & Mendelson (2002) provide valuable explanations for phenomena in the field of adversarial transferability. Our work introduces a novel theoretical toolset to the field, offering researchers an alternative lens through which to understand adversarial transferability. We aim to inspire more theoretical insights and foster further advancements in this domain.

D.2.2 Conflicting Opinions

We observe a significant and intriguing disagreement within the academic community concerning the role of "diversity" in transferable model ensemble attacks:

- Some studies advocate for enhancing model diversity to produce more transferable adversarial examples. For instance, Li et al. (2020) applies feature-level perturbations to an existing model to potentially create a huge set of diverse "Ghost Networks". Li et al. (2023) emphasizes the importance of diversity in surrogate models and promotes attacking a Bayesian model to achieve desirable transferability. Tang et al. (2024) supports the notion of improved diversity, suggesting the generation of adversarial examples independently from individual models.
- In contrast, other researchers adopt a diversity-reduction strategy to enhance adversarial transferability. For example, Xiong et al. (2022) focuses on minimizing gradient variance among ensemble models to improve transferability. Meanwhile, Chen et al. (2023) introduces a disparity-reduced filter designed to decrease gradient variances among surrogate models in ensemble attacks.

Although all these studies reference "diversity," their perspectives appear to diverge. In this paper, we advocate for increasing the diversity of surrogate models. However, we also recognize that diversity-reduction approaches have their merits.

Consider the vulnerability-diversity decomposition of transferability error presented in Theorem 1. It suggests the presence of a vulnerability-diversity trade-off in transferable model ensemble attacks. In other words, we may need to prioritize either vulnerability or diversity to effectively reduce transferability error. Diversity-reduction approaches aim to stabilize the training process, thereby increasing the "bias." In contrast, diversity-promoting methods directly enhance "diversity." This analysis, framed within our unified theoretical framework, provides insight into the differing opinions regarding adversarial transferability in the academic community.

D.3 Compare with A Previous Bound

We compare Theorem 1 with Lemma 5 in Yang et al. (2021). We first restate Lemma 5 in Yang et al. (2021) and our Theorem 1. Our theoretical results and theirs offer complementary perspectives in the analysis of transferable adversarial attack.

Lemma 5 (Yang et al. (2021)). *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 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 and $\mathcal{A}(\cdot)$ is ρ-conservative (TV distance between the adversarial example distribution and clean data distribution is less than ρ, which is defined as Definition 7 and 8 in Yang et al. (2021)).

Lemma 5 states an intriguing conclusion: if two models exhibit low risk on the original data distribution and the distributional discrepancy between adversarial examples and the original data is small, the predictions of the two models on the same input will be close. In other words, for two well-performing models, if an attack strategy successfully targets one model, it is highly likely to succeed on the other. Lemma 5 thus describes the success rate of transferring an attack from one model to another. In contrast, Theorem 1 demonstrates that if the ensemble models exhibit significant

output differences on the same input, the resulting diverse ensemble is more effective at generating adversarial examples with reduced transferability.

To better clarify, let A denote the ensemble models generating adversarial examples and B the model being attacked. Comparing Lemma 5 and our work leads to the following reasoning:

- Assumptions: Suppose A and B both fit the original data distribution well (i.e., the risk of A and B is bounded by ϵ, as in Lemma 5).
- Vulnerability-diversity trade-off: As shown in our work, increasing ensemble diversity while keeping vulnerability constant reduces the transferability error of adversarial examples generated by the ensemble.
- Distributional gap: Many models in parameter space, such as A and B, are vulnerable to these adversarial examples. However, fitting both the original data distribution and the adversarial example distribution simultaneously becomes challenging, leading to a large distributional discrepancy.
- Impact on Lemma 5: This discrepancy enlarges ρ in Lemma 5, thereby loosening its "conservative condition" and weakening its theoretical guarantee of successful transferability. Consequently, adversarial transferability decreases, which could be interpreted as a potential contradiction.

We argue that no actual contradiction exists between Lemma 5 and our work. Instead, they provide complementary analyses:

- Upper bound interpretation: Lemma 5 provides an upper bound rather than an equality or lower bound. While an increase in ρ loosens this upper bound, it does not necessarily imply that the left-hand side (i.e., transferability success) will increase. The significance of an upper bound lies in the fact that a tighter right-hand side suggests the potential for a smaller left-hand side. However, a looser upper bound does not necessarily imply that the left-hand side will increase. Therefore, while increasing ensemble diversity may loosen the upper bound in Lemma 5, it does not contradict the fundamental interpretation of it.
- Complementary perspectives: While Lemma 5 analyzes the trade-off between ϵ (model fit to the original data) and ρ (distributional discrepancy), our work focuses on the trade-off between vulnerability and ensemble diversity. Together, they provide a comprehensive understanding of the factors influencing adversarial transferability.

We now further elucidate the relationship between our results and Lemma 5:

- Reducing Transferability Error: To minimize transferability error (as in our work), the adversarial transferability described by Lemma 5 may have stronger theoretical guarantees, requiring its upper bound to be tighter.
- Trade-off Between ϵ and ρ: To tighten the bound in Lemma 5, either ϵ or ρ must decrease. However, the two exhibit a trade-off:
  - If ϵ decreases, A and B fit the original data distribution better. However, beyond a certain point, the adversarial examples generated by A diverge significantly from the original data distribution, increasing ρ.
  - If ρ decreases, the adversarial example distribution becomes closer to the original data distribution. However, beyond a certain point, A exhibits similar losses on both distributions, resulting in a higher ϵ.

Therefore, Lemma 5 indicates the potential trade-off between ϵ and ρ in adversarial transferability, while our Theorem 1 emphasizes the trade-off between vulnerability and diversity. By integrating the perspectives from both Lemma 5 and our findings, these results illuminate different facets of adversarial transferability, offering complementary theoretical insights. This combined understanding deepens our knowledge of the factors influencing adversarial transferability and lays a solid foundation for future research in the field.

D.4 Extension of Theorem 2

Firstly, our proposed setting aligns with many realistic scenarios, as demonstrated in (Wu et al., 2024; Tang et al., 2024; Li et al., 2023; Xiong et al., 2022; Lin et al., 2019). Specifically, they encompass cases where both the surrogate model and the target model adopt the same architectures. It reflects the fact that the settings in this paper are commonly considered in prior studies.

Furthermore, our theoretical framework is not only rigorous but also highly adaptable, making it straightforward to effectively extend and be more general. For example, this can be achieved by redefining the model space, as shown in Appendix D.4.1, or by drawing insights from domain adaptation theory, as discussed in Appendix D.4.2. The simplicity and flexibility of our framework allow researchers to follow and build upon it seamlessly, fostering further innovation in addressing adversarial transferability challenges.

D.4.1 Defining the Model Space

The two issues raised above can be circumvented by redefining the model space. In particular, we consider N surrogate classifiers $f_1, \cdots, f_N$ trained to generate adversarial examples. ***Let D be the distribution over the surrogate models (for instance, the distribution of all the low-risk models), and $f_i \in D, i \in [N]$. The low-risk claim is in line with Lemma 5 in Yang et al. (2021), which assumes that the risk of surrogate model and target model is low (have risk at most ϵ). Therefore, the surrogate model and target model can be seen as drawing from the same distribution (such as a distribution of all the low-risk models).*** For a data point $z = (x, y) \in \mathcal{Z}$ and N classifiers for model ensemble attack, define the population risk $L_P(z)$ and the empirical risk $L_D(z)$ as

$$L_P(z) = \mathbb{E}_{f\sim D}[\ell(f(x), y)].$$
$$L_D(z) = \frac{1}{N} \sum_{i\in[N], f_i\in D} \ell(f_i(x), y).$$

Now here is an extension of Theorem 2 based on the above definition. ***The proof is almost the same as Appendix B.3, but the definition of distribution is different.***

Theorem 5 (Extension of Theorem 2)**.** *Let $\mathcal{P}_{D^N}$ be the joint distribution of $f_1, \cdots, f_N$, and $\mathcal{P}_{\bigotimes_{i=1}^N D}$ be the joint measure induced by the product of the marginals. If the loss function ℓ is bounded by $\beta \in R_+$ and $\mathcal{P}_{D^N} \ll \mathcal{P}_{\bigotimes_{i=1}^N D}$ for any function f_i, then for $\alpha > 1$ and $\gamma = \frac{\alpha}{\alpha-1}$, with probability at least $1-\delta$, there holds*

$$TE(z, \epsilon) \leq 4\mathcal{R}_N(\mathcal{Z}) + \sqrt{\frac{2\gamma\beta^2}{N} \ln \frac{2^{\frac{1}{\gamma}} H_\alpha^{\frac{1}{\alpha}} \left(\mathcal{P}_{D^N} \| \mathcal{P}_{\bigotimes_{i=1}^N D}\right)}{\delta}}. \tag{20}$$

The first term answers the question that more surrogate models and smaller complexity will lead to a smaller $\mathcal{R}_N(\mathcal{Z})$ and contributes to a tighter bound of $TE(z, \epsilon)$. The second term motivates us that if we reduce the interdependency among the ensemble components, then the upper bound of $TE(z, \epsilon)$ will be tighter. Recall that $H_\alpha(\mathcal{P}_{D^N} \| \mathcal{P}_{\bigotimes_{i=1}^N D})$ quantifies the divergence between the joint distribution $\mathcal{P}_{D^N}$ and product of marginals $\mathcal{P}_{\bigotimes_{i=1}^N D}$. The joint distribution captures dependencies while the product of marginals does not. So the divergence between them measures the degree of dependency among the N classifiers $f_1, \cdots, f_N$. As a result, *improving the diversity of $f_1, \cdots, f_N$ and reduce the interdependence among them is beneficial to adversarial transferability*.

Proof. We know that

$$\begin{aligned} TE(z) = L_P(z^*) - L_P(z) &\leq L_P(z^*) - L_P(z) + (L_D(z) - L_D(z^*)) \\ &= (L_P(z^*) - L_D(z^*)) + (L_D(z) - L_P(z)) \\ &\leq \sup_{x\in\mathcal{B}_\epsilon(x)} (L_P(z) - L_D(z)) + \sup_{x\in\mathcal{B}_\epsilon(x)} (L_D(z) - L_P(z)) \\ &\leq \sup_{z\in\mathcal{Z}} (L_P(z) - L_D(z)) + \sup_{z\in\mathcal{Z}} (L_D(z) - L_P(z)). \end{aligned}$$

We define

$$\Phi_1(D) = \sup_{z \in \mathcal{Z}} \{L_P(z) - L_D(z)\},$$
$$\Phi_1(D') = \sup_{z \in \mathcal{Z}} \{L_P(z) - L_{D'}(z)\},$$

where D' is also a distribution over the classifiers, and only one classifier in D and D' is different. And

$$L_{D'}(z) = \frac{1}{N} \sum_{\substack{i = 1, \cdots, N \\ f_i \sim D'}} \ell(f_i(x), y).$$

From the definition of D and D', we have

$$\begin{aligned}
\Phi_1(D) - \Phi_1(D') &= \sup_{z \in \mathcal{Z}} \{L_P(z) - L_D(z)\} - \sup_{z \in \mathcal{Z}} \{L_P(z) - L_{D'}(z)\} \\
&\leq \sup_{z \in \mathcal{Z}} \{L_P(z) - L_D(z) - (L_P(z) - L_{D'}(z))\} \\
&= \sup_{z \in \mathcal{Z}} \{L_{D'}(z) - L_D(z)\} \\
&= \frac{1}{N} \sup_{z \in \mathcal{Z}} \left[\sum_{\substack{i = 1, \cdots, N \\ f_i \sim D'}} \ell(f_i(x), y) - \sum_{\substack{i = 1, \cdots, N \\ f_i \sim D}} \ell(f_i(x), y) \right].
\end{aligned}$$

By assuming that loss function ℓ is bounded by β, we have

$$|\Phi_1(D) - \Phi_1(D')| \leq \frac{\beta}{N}.$$

According to Theorem 1 in Esposito & Mondelli (2024), for all $\delta \in (0, 1)$ and $\alpha > 1$, with probability at least $1 - \delta$, we have

$$\Phi_1(D) \leq \mathbb{E}_D[\Phi_1(D)] + \sqrt{\frac{\alpha \beta^2}{2(\alpha - 1)N} \ln \frac{2^{\frac{\alpha - 1}{\alpha}} H_\alpha^{\frac{1}{\alpha}} \left(\mathcal{P}_{D^N} \| \mathcal{P}_{\bigotimes_{i=1}^N D} \right)}{\delta}}. \tag{21}$$

Let $f \sim D$ and $f' \sim D'$ be different classifiers. Then we estimate the upper bound of $\mathbb{E}_D[\Phi_1(D)]$ as follows:

$$\begin{aligned}
\mathbb{E}_D[\Phi_1(D)] &= \mathbb{E}_D \left[\sup_{z \in \mathcal{Z}} (L_P(z) - L_D(z)) \right] \\
&= \mathbb{E}_D \left[\sup_{z \in \mathcal{Z}} \mathbb{E}_{f \sim D'} \left(L_{D'}(z) - L_D(z) \right) \right] \\
&\leq \mathbb{E}_{D, D'} \left[\sup_{z \in \mathcal{Z}} \left(L_{D'}(z) - L_D(z) \right) \right] && \text{(Jensen inequality)} \\
&= \mathbb{E}_{D, D'} \left\{ \sup_{z \in \mathcal{Z}} \frac{1}{N} \left[\sum_{\substack{i = 1, \cdots, N \\ f_i' \sim D'}} \ell(f_i(x), y) - \sum_{\substack{i = 1, \cdots, N \\ f_i \sim D}} \ell(f_i(x), y) \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{D,D'}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\left[\ell(f_i'(x),y)-\ell(f_i(x),y)\right]\right]\right\}\\
&= \mathbb{E}_{\boldsymbol{\sigma},D,D'}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\left[\ell(f_i'(x),y)-\ell(f_i(x),y)\right]\right]\right\}\\
&\le \mathbb{E}_{\boldsymbol{\sigma},D'}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\ell(f_i'(x),y)\right]\right\}+\mathbb{E}_{\boldsymbol{\sigma},D}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\ell(f_i(x),y)\right]\right\}\\
&= 2\cdot\mathbb{E}_{\boldsymbol{\sigma},D}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f_i(x),y)\right\}\\
&= 2\mathcal{R}_N(\mathcal{F}).
\end{aligned}
$$

Likewise, if we define

$$
\begin{aligned}
\Phi_2(D) &= \sup_{z\in\mathcal{Z}}\left\{L_D(z)-L_P(z)\right\},\\
\Phi_2(D') &= \sup_{z\in\mathcal{Z}}\left\{L_{D'}(z)-L_P(z)\right\},
\end{aligned}
$$

then we have

$$
\begin{aligned}
\Phi_2(D)-\Phi_2(D') &= \sup_{z\in\mathcal{Z}}\left\{L_D(z)-L_P(z)\right\}-\sup_{z\in\mathcal{Z}}\left\{L_{D'}(z)-L_P(z)\right\}\\
&\le \sup_{z\in\mathcal{Z}}\left\{L_D(z)-L_P(z)-(L_{D'}(z)-L_P(z))\right\}\\
&= \sup_{z\in\mathcal{Z}}\left\{L_D(z)-L_{D'}(z)\right\}\\
&= \frac{1}{N}\sup_{z\in\mathcal{Z}}\left[\sum_{\substack{i=1,\cdots,N\\ f_i\sim D}}\ell(f_i(x),y)-\sum_{\substack{i=1,\cdots,N\\ f_i\sim D'}}\ell(f_i(x),y)\right].
\end{aligned}
$$

According to the assumption that loss function ℓ is bounded by β, we have

$$
|\Phi_2(D)-\Phi_2(D')|\le\frac{\beta}{N}.
$$

According to Theorem 1 in Esposito & Mondelli (2024), for all $\delta\in(0,1)$ and $\alpha>1$, with probability at least $1-\delta$, we have

$$
\Phi_2(D)\le\mathbb{E}_D[\Phi_2(D)]+\sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N}\ln\frac{2^{\frac{\alpha-1}{\alpha}}H_\alpha^{\frac{1}{\alpha}}\left(\mathcal{P}_{D^N}\|\mathcal{P}_{\bigotimes_{i=1}^N D_i}\right)}{\delta}}. \tag{22}
$$

We estimate the upper bound of $\mathbb{E}_D[\Phi_2(D)]$ as follows:

$$
\begin{aligned}
\mathbb{E}_D[\Phi_2(D)] &= \mathbb{E}_D\left[\sup_{z\in\mathcal{Z}}(L_D(z)-L_P(z))\right]\\
&= \mathbb{E}_D\left[\sup_{z\in\mathcal{Z}}\mathbb{E}_{f\sim D'}\left(L_D(z)-L_{D'}(z)\right)\right]\\
&\le \mathbb{E}_{D,D'}\left[\sup_{z\in\mathcal{Z}}\left(L_D(z)-L_{D'}(z)\right)\right] \qquad \text{(Jensen inequality)}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{D,D'}\left\{\sup_{z\in\mathcal{Z}}\frac{-1}{N}\left[\sum_{\substack{i=1,\cdots,N\\ f_i'\sim D'}}\ell(f_i(x),y)-\sum_{\substack{i=1,\cdots,N\\ f_i\sim D}}\ell(f_i(x),y)\right]\right\}\\
&= \mathbb{E}_{D,D'}\left\{\sup_{z\in\mathcal{Z}}\frac{-1}{N}\left[\sum_{i=1}^{N}\left[\ell(f_i'(x),y)-\ell(f_i(x),y)\right]\right]\right\}\\
&= \mathbb{E}_{\boldsymbol{\sigma},D,D'}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\left[\ell(f_i'(x),y)-\ell(f_i(x),y)\right]\right]\right\}\\
&\le \mathbb{E}_{\boldsymbol{\sigma},D'}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\ell(f_i'(x),y)\right]\right\}+\mathbb{E}_{\boldsymbol{\sigma},D}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\left[\sum_{i=1}^{N}\sigma_i\ell(f_i(x),y)\right]\right\}\\
&= 2\cdot\mathbb{E}_{\boldsymbol{\sigma},D}\left\{\sup_{z\in\mathcal{Z}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f_i(x),y)\right\}\\
&= 2\mathcal{R}_N(\mathcal{F}).
\end{aligned}
$$

Therefore, with probability at least $1-\delta$, there holds

$$
TE(z,\epsilon)=\Phi_1(D)+\Phi_2(D)\le 4\mathcal{R}_N(\mathcal{F})+\sqrt{\frac{2\alpha\beta^2}{(\alpha-1)N}\ln\frac{2^{\frac{\alpha-1}{\alpha}}H_\alpha^{\frac{1}{\alpha}}\left(\mathcal{P}_{D^n}\|\mathcal{P}_{\bigotimes_{i=1}^n D_i}\right)}{\delta}}.
$$

The proof is complete. □

D.4.2 Extension to Different Parameter Distributions

In fact, the two issues mentioned at the beginning of Appendix D.4 can also be properly addressed using domain adaptation theory (Blitzer et al., 2007). Intuitively, there is a need for domain adaptation between the surrogate model and the target model. Mathematically, a feasible and straightforward approach is to define a divergence metric and apply domain adaptation theory. For instance,

Definition 5 ($\mathcal{X}$ divergence for transferable attack)**.** *Given a feature space $\mathcal{X}$ and a label space $\mathcal{Y}$, We denote the hypothesis space by $\mathcal{H}:\mathcal{X}\mapsto\mathcal{Y}$. Denote the parameter space of surrogate model and target model by Θ and Θ', respectively. Let $f(\theta;\cdot)\in\mathcal{H}$ be a classifier parameterized by θ, where $\theta\in\Theta$ or $\theta\in\Theta'$. Consider a metric loss function $\ell:\mathcal{Y}\times\mathcal{Y}\mapsto\mathbb{R}_0^+$. Then the $\mathcal{X}$ divergence between the surrogate model domain and the target model domain can be defined as:*

$$
d_{\mathcal{X}}\left(\mathcal{P}_\Theta,\mathcal{P}_{\Theta'}\right)=2\sup_{x\in\mathcal{X}}\left|\mathbb{E}_{\theta\sim\mathcal{P}_\Theta}\ell\left[f(\theta;x),y\right]-\mathbb{E}_{\theta\sim\mathcal{P}_{\Theta'}}\ell\left[f(\theta;x),y\right]\right|.
$$

It is a natural extension from $\mathcal{H}$ divergence in domain adaptation theory (Blitzer et al., 2007) to transferable adversarial attack. We consider such divergence and redefine the population risk $L_P(z)$ in Eq. (3) as

$$
L_P(z,\Theta)=\mathbb{E}_{\theta\sim\mathcal{P}_\Theta}[\ell(f(\theta;x),y)].
$$

Therefore, there is a connection between the surrogate model domain and target model domain:

$$
|L_P(z,\Theta')-L_P(z,\Theta)|\le\frac{1}{2}d_{\mathcal{X}}\left(\mathcal{P}_\Theta,\mathcal{P}_{\Theta'}\right).
$$

Substituting Eq. (12) into this inequality, we will obtain a general upper bound with an additional divergence term on the right-hand side:

$$
TE(z,\epsilon)\le 4\mathcal{R}_N(\mathcal{Z})+\sqrt{\frac{2\gamma\beta^2}{N}\ln\frac{2^{\frac{1}{\gamma}}H_\alpha^{\frac{1}{\alpha}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)}{\delta}}+d_{\mathcal{X}}\left(\mathcal{P}_\Theta,\mathcal{P}_{\Theta'}\right).
$$

According to this theory, the smaller the $d_{\mathcal{X}}$, the tighter the theoretical bound. Therefore, we need to let surrogate model domain be as close to the target model domain as possible. Such insight is in line with (Zhao et al., 2023), which shows that reducing model discrepancy (which corresponds to the divergence defined above) can make adversarial examples highly transferable.

Moreover, to further advance the field, leveraging advanced domain adaptation theories (e.g., Wang & Mao (2022); Zhang et al. (2019b)) could yield deeper theoretical insights and inspire new algorithm designs. In the revision, we provide a more detailed analysis, including:

- Extending our analysis to scenarios with different parameter spaces and distributions.
- Future work can be done by identifying suitable mathematical tools from the extensive domain adaptation literature (Redko et al., 2020) to analyze adversarial transferability more deeply and inform algorithm development.

These enhancements will significantly expand the impact of our work by:

- Being the first to draw an analogy between statistical learning theory and adversarial transferability, thereby introducing a new perspective to the field.
- Being the first to encourage researchers to consider domain adaptation for deeper analysis and algorithmic innovations in transferable adversarial attack.

D.5 Information-theoretic Analysis

Firstly, we formally define the Kullback-Leibler divergence (KL divergence), mutual information and total variation distance (TV distance).

Definition 6 (Kullback-Leibler Divergence)**.** *Given two probability distributions P and Q, the Kullback-Leibler (KL) divergence between P and Q is*

$$\mathrm{D}_{\mathrm{KL}}(P\|Q) = \int_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}dx.$$

We know that $\mathrm{D}_{\mathrm{KL}}(P\|Q) \in [0,+\infty]$, and $\mathrm{D}_{\mathrm{KL}}(P\|Q) = 0$ if and only if $P = Q$.

Definition 7 (Mutual Information)**.** *For continuous random variables X and Y with joint probability density function $p(x,y)$ and marginal probability density functions $p(x)$ and $p(y)$, the mutual information is defined as:*

$$I(X;Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy.$$

We know that $I(X;Y) \in [0,+\infty]$, and $I(X;Y) = 0$ if and only if X and Y are independent to each other.

Definition 8 (Total Variation Distance)**.** *Given two probability distributions P and Q, the Total Variation (TV) distance between P and Q is*

$$\mathrm{D}_{\mathrm{TV}}(P\|Q) = \frac{1}{2}\int_{x\in\mathcal{X}} |P(x) - Q(x)|\,dx.$$

We know that $\mathrm{D}_{\mathrm{TV}}(P\|Q) \in [0,1]$. Also, $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 0$ if and only if P and Q coincides, and $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 1$ if and only if P and Q are disjoint.

Note that the training process of N classifiers can be viewed as sampling the parameter sets $\overline{\theta}^N = (\overline{\theta}_1,\ldots,\overline{\theta}_N)$ from the distribution $\mathcal{P}_{\Theta^N}$, i.e., $\overline{\theta}^N \sim \mathcal{P}_{\Theta^N}$. We generate a transferable adversarial example using these N models and evaluate its performance on another N models $\theta^N = (\theta_1,\ldots,\theta_N)$, which is an independent copy of $\overline{\theta}^N$. For a data $z = (x,y) \in \mathcal{Z}$ and the parameter set θ^N, our aim is to bound **the difference of attack performance between the given N models $\overline{\theta}^N$ and N unknown models θ^N**. In other words, if

- An adversarial example z can effectively attack the given model ensemble.
- There is guarantee for the aforementioned difference of attack performance between known and unknown models.

Then there is adversarial transferability guarantee for z.

Here we provide further analysis from the perspective of information in deep learning. It is supported by both empirical observations (Alemi et al., 2016; Shwartz-Ziv & Tishby, 2017; Wu et al., 2020; Lorenzen et al., 2021; Hu et al., 2024) and theoretical insights (Tishby & Zaslavsky, 2015; Xu & Raginsky, 2017; Jeon & Van Roy, 2022; Kawaguchi et al., 2023; Wang & Mao, 2024).

Theorem 6. *Given N surrogate models $\overline{\theta}^N = (\overline{\theta}_1, \ldots, \overline{\theta}_N) \sim \mathcal{P}_{\Theta^N}$ as the ensemble components. Let $\theta^N = (\theta_1, \ldots, \theta_N) \sim \mathcal{P}_{\Theta^N}$ be the target models, which is an independent copy of $\overline{\theta}^N$. Assume the loss function ℓ is bounded by $\beta \in \mathbb{R}_+$ and $\mathcal{P}_{\Theta^N}$ is absolutely continuous with respect to $\mathcal{P}_{\bigotimes_{i=1}^N \Theta}$. For $\alpha > 1$ and adversarial example $z = (x, y) \sim \mathcal{P}_{\mathcal{Z}}$, Let*

$$\Delta_N(\theta, z) = \mathbb{E}_{\overline{\theta}^N \sim \mathcal{P}_{\Theta^N}} \left[\frac{1}{N} \sum_{i=1}^N \ell(f(\overline{\theta}_i; x), y) \right] - \frac{1}{N} \sum_{i=1}^N \ell(f(\theta_i; x), y).$$

Then there holds

$$\left| \mathbb{E}_{z, \theta^N \sim \mathcal{P}_{\mathcal{Z}, \Theta^N}} \Delta_N(\theta, z) \right| \leq 2\beta \cdot \mathrm{D_{TV}} \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right) + \sqrt{\frac{\alpha \beta^2}{2(\alpha - 1)N} \left(I\left(\overline{\theta}^N; z\right) + \frac{1}{\alpha} \log H_\alpha \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right) \right)},$$

where $\mathrm{D_{TV}}(\cdot\|\cdot)$, $I(\cdot\|\cdot)$ and $H_\alpha(\cdot\|\cdot)$ denotes TV distance, mutual information and Hellinger integrals, respectively.

In Theorem 6:

- $\Delta_N(\theta, z)$ quantifies how effectively the surrogate models represent all possible target models. Taking the expectation of $\Delta_N(\theta, z)$ over z and θ^N accounts for the inherent randomness in both adversarial examples and surrogate models.
- The mutual information $I\left(\overline{\theta}^N; z\right)$ quantifies how much information about the surrogate models is retained in the adversarial example. Intuitively, higher mutual information indicates that the adversarial example is overly tailored to the surrogate models, capturing specific features of these models. This overfitting reduces its ability to generalize and transfer effectively to other target models. By controlling the complexity of the surrogate models, the specific information captured by the adversarial example can be limited, encouraging it to rely on broader, more transferable patterns rather than model-specific details. This reduction in overfitting enhances the adversarial example's transferability to diverse target models.
- The total variation (TV) distance, $\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)$, and the Hellinger integral, $H_\alpha\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)$, capture the interdependence among the surrogate models.

Theorem 6 reveals that the following strategies contribute to a tighter bound:

- Increasing the number of surrogate models, i.e., increasing N;
- Reducing the model complexity of surrogate models, i.e., reducing $I\left(\overline{\theta}^N; z\right)$;
- Making the surrogate models more diverse, i.e., reducing $\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)$ and $H_\alpha\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)$.

A tighter bound ensures that an adversarial example maximizing the loss function on the surrogate models will also lead to a high loss on the target models, thereby enhancing transferability.

Proof. According to Donsker and Varadhan's variational formula, for any $\lambda \in \mathbb{R}$, there holds:

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{P}_{\mathcal{Z},\Theta^N} \| \mathcal{P}_{\mathcal{Z}} \otimes \mathcal{P}_{\Theta^N}) \geq \lambda \mathbb{E}_{z,\theta^N \sim \mathcal{P}_{\mathcal{Z},\Theta^N}} \Delta_N(\theta, z) - \log \mathbb{E}_{z \sim \mathcal{P}_{\mathcal{Z}}} \mathbb{E}_{\theta^N \sim \mathcal{P}_{\Theta^N}} \left[e^{\lambda \Delta_N(\theta,z)} \right]. \quad (23)$$

Fix $z \in \mathcal{Z}$,

$$\begin{aligned}
\mathbb{E}_{\theta^N \sim \mathcal{P}_{\Theta^N}} \left[e^{\lambda \Delta_N(\theta,z)} \right] &= \int e^{\lambda \Delta_N(\theta,z)} d\mathcal{P}_{\Theta^N} \\
&= \int e^{\lambda \Delta_N(\theta,z)} \frac{d\mathcal{P}_{\Theta^N}}{d\mathcal{P}_{\bigotimes_{i=1}^N \Theta}} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \\
&\leq \left(\int e^{\frac{\alpha}{\alpha-1} \lambda \Delta_N(\theta,z)} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right)^{\frac{\alpha-1}{\alpha}} \left(\int \left(\frac{d\mathcal{P}_{\Theta^N}}{d\mathcal{P}_{\bigotimes_{i=1}^N \Theta}} \right)^{\alpha} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right)^{\frac{1}{\alpha}} \\
&= \left(\int e^{\frac{\alpha}{\alpha-1} \lambda \Delta_N(\theta,z)} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right)^{\frac{\alpha-1}{\alpha}} H_{\alpha}^{\frac{1}{\alpha}}(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}). \quad (24)
\end{aligned}$$

The third line uses Hölder's inequality, while the last line follows Definition 4. Now we deal with the first term. Denote

$$\begin{aligned}
\Delta_1 &= \mathbb{E}_{\bar{\theta}^N \sim \mathcal{P}_{\Theta^N}} \left[\frac{1}{N} \sum_{i=1}^N \ell(f(\bar{\theta}_i; x), y) \right] - \mathbb{E}_{\bar{\theta}^N \sim \mathcal{P}_{\bigotimes_{i=1}^N \Theta}} \left[\frac{1}{N} \sum_{i=1}^N \ell(f(\bar{\theta}_i; x), y) \right], \\
\Delta_2 &= \mathbb{E}_{\bar{\theta}^N \sim \mathcal{P}_{\bigotimes_{i=1}^N \Theta}} \left[\frac{1}{N} \sum_{i=1}^N \ell(f(\bar{\theta}_i; x), y) \right] - \frac{1}{N} \sum_{i=1}^N \ell(f(\theta_i; x), y).
\end{aligned}$$

Notice that

$$\begin{aligned}
|\Delta_1| &= \left| \iint \cdots \int \left[\frac{1}{N} \sum_{i=1}^N \ell(f(\bar{\theta}_i; x), y) \right] \left[\mathcal{P}_{\Theta^N}(\bar{\theta}_1, \cdots, \bar{\theta}_N) - \mathcal{P}_{\bigotimes_{i=1}^N \Theta}(\bar{\theta}_1, \cdots, \bar{\theta}_N) \right] d\bar{\theta}_1 \cdots d\bar{\theta}_N \right| \\
&\leq \beta \iint \cdots \int \left| \mathcal{P}_{\Theta^N}(\bar{\theta}_1, \cdots, \bar{\theta}_N) - \mathcal{P}_{\bigotimes_{i=1}^N \Theta}(\bar{\theta}_1, \cdots, \bar{\theta}_N) \right| d\bar{\theta}_1 \cdots d\bar{\theta}_N \\
&= \beta \int \left| \mathcal{P}_{\Theta^N}(\bar{\theta}^N) - \mathcal{P}_{\bigotimes_{i=1}^N \Theta}(\bar{\theta}^N) \right| d\bar{\theta}^N \\
&\leq 2\beta \cdot \mathrm{D}_{\mathrm{TV}} \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right). \quad (25)
\end{aligned}$$

Also,

$$\begin{aligned}
\int \left(e^{\frac{\alpha}{\alpha-1} \lambda \Delta_2} \right) d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} &= \mathbb{E}_{\theta^N \sim \mathcal{P}_{\bigotimes_{i=1}^N \Theta}} \left[e^{\frac{\alpha}{\alpha-1} \lambda \Delta_2} \right] \\
&= \prod_{i=1}^N \mathbb{E}_{\theta_i \sim \mathcal{P}_{\Theta}} \left[\exp \left(\frac{\alpha \lambda}{\alpha - 1} \left(\mathbb{E}_{\bar{\theta}_i \sim \mathcal{P}_{\Theta}} \left[\frac{1}{N} \ell(f(\bar{\theta}_i; x), y) \right] - \frac{1}{N} \ell(f(\theta_i; x), y) \right) \right) \right] \\
&\leq \prod_{i=1}^N \exp \left(\frac{\alpha^2}{8(\alpha-1)^2 N^2} \lambda^2 \beta^2 \right). \\
&\leq \exp \left(\frac{\alpha^2}{8(\alpha-1)^2 N} \lambda^2 \beta^2 \right). \quad (26)
\end{aligned}$$

The third line is due to Hoeffding's Lemma (using it for each θ_i). Therefore, recall the fact that $\Delta_N(\theta, z) = \Delta_1 + \Delta_2$, we have

$$\begin{aligned}
\int e^{\frac{\alpha}{\alpha-1} \lambda \Delta_N(\theta,z)} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} &= \int \left(e^{\frac{\alpha}{\alpha-1} \lambda \Delta_1} \cdot e^{\frac{\alpha}{\alpha-1} \lambda \Delta_2} \right) d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \\
&\leq \exp \left(\frac{2\lambda\alpha\beta}{\alpha-1} \mathrm{D}_{\mathrm{TV}} \left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta} \right) \right) \int e^{\frac{\alpha}{\alpha-1} \lambda \Delta_2} d\mathcal{P}_{\bigotimes_{i=1}^N \Theta} \\
&\qquad \text{(Using (25))}
\end{aligned}$$

$$\leq \exp\left(\frac{2\lambda\alpha\beta}{\alpha-1}\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)+\frac{\alpha^2}{8(\alpha-1)^2N}\lambda^2\beta^2\right) \quad \text{(Using (26))}$$

With the above results, we obtain the following:

$$\log \mathbb{E}_{z\sim\mathcal{P}_{\mathcal{Z}}}\mathbb{E}_{\theta^N\sim\mathcal{P}_{\Theta^N}}\left[e^{\lambda\Delta_N(\theta,z)}\right]\leq 2\lambda\beta\cdot\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)+\frac{\alpha}{8(\alpha-1)N}\lambda^2\beta^2+\log H_\alpha^{\frac{1}{\alpha}}(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}).$$

Substitute the above into Eq. (23), we have

$$\frac{\alpha}{8(\alpha-1)N}\beta^2\lambda^2+\left(2\beta\cdot\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)-\mathbb{E}_{z,\theta^N\sim\mathcal{P}_{\mathcal{Z},\Theta^N}}\Delta_N(\theta,z)\right)\lambda+ \mathrm{D_{KL}}(\mathcal{P}_{\mathcal{Z},\Theta^N}\|\mathcal{P}_{\mathcal{Z}}\otimes\mathcal{P}_{\Theta^N})+\log H_\alpha^{\frac{1}{\alpha}}(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta})\geq 0.$$

Let the discriminant of the quadratic function with respect to λ be less than or equal to 0, leading to:

$$\left|2\beta\cdot\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)-\mathbb{E}_{z,\theta^N\sim\mathcal{P}_{\mathcal{Z},\Theta^N}}\Delta_N(\theta,z)\right|\leq \sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N}\left(\mathrm{D_{KL}}\left(\mathcal{P}_{\mathcal{Z},\Theta^N}\|\mathcal{P}_{\mathcal{Z}}\otimes\mathcal{P}_{\Theta^N}\right)+\frac{1}{\alpha}\log H_\alpha\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)\right)}. \quad (27)$$

In other words,

$$\left|\mathbb{E}_{z,\theta^N\sim\mathcal{P}_{\mathcal{Z},\Theta^N}}\Delta_N(\theta,z)\right|\leq 2\beta\cdot\mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)+ \sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N}\left(\mathrm{D_{KL}}\left(\mathcal{P}_{\mathcal{Z},\Theta^N}\|\mathcal{P}_{\mathcal{Z}}\otimes\mathcal{P}_{\Theta^N}\right)+\frac{1}{\alpha}\log H_\alpha\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^N\Theta}\right)\right)}.$$

Finally, substitute $I\left(\overline{\theta}^N;z\right)=\mathrm{D_{KL}}\left(\mathcal{P}_{\mathcal{Z},\Theta^N}\|\mathcal{P}_{\mathcal{Z}}\otimes\mathcal{P}_{\Theta^N}\right)$ into above and we can get the desired result. □

D.6 Compare with Generalization Error Bound

We note that a key distinction between transferability error and generalization error lies in the *independence assumption*. Conventional generalization error analysis relies on an assumption: each data point from the dataset is independently sampled (Zou & Liu, 2023; Hu et al., 2023). By contrast, the surrogate models $f_1,\cdots,f_N$ for ensemble attack are usually trained on the datasets with similar tasks, e.g., image classification. In this case, such models tend to correctly classify easy examples while misclassify difficult examples (Bengio et al., 2009). Consequently, such correlation indicates dependency (Lancaster, 1963), suggesting that ***we cannot assume these surrogate models behave independently for a solid theoretical analysis***. Additionally, there are alternative methods for analyzing concentration inequality in generalization error analysis that do not rely on the independence assumption (Kontorovich & Ramanan, 2008; Mohri & Rostamizadeh, 2008; Lei et al., 2019; Zhang et al., 2019a). However, such data-dependent analysis is either too loose (Lampert et al., 2018) (because it includes an additional additive factor that grows with the number of samples (Esposito & Mondelli, 2024)) or requires specific independence structure of data (Zhang & Amini, 2024) that may not align well with model ensemble attacks. Therefore, we uses the latest techniques of information theory (Esposito & Mondelli, 2024) about concentration inequality regarding dependency. To our best knowledge, it is the first mathematical tool in concentration inequality that fits our needs.

D.7 The Analogy between Generalization and Adversarial Transferability

Besides providing inspiration for model ensemble attacks, the theoretical evidence in this paper also offers new insights into another fascinating idea. Within the extensive body of research on

transferable adversarial attack algorithms accumulated over the years (Gu et al., 2024), we revisit a foundational analogy that is universally applicable in the adversarial transferability literature: *The transferability of an adversarial example is an analogue to the generalizability of the model* (Dong et al., 2018). In other words, the ideas that enhance model generalization in deep learning may also improve adversarial transferability (Lin et al., 2019). Over the past few years, this analogy has significantly inspired the development of numerous effective algorithms, which directly reference it in their papers (Lin et al., 2019; Wang et al., 2021; Wang & He, 2021; Xiong et al., 2022; Chen et al., 2024b). And some recent papers are also inspired by it (Chen et al., 2023; Wu et al., 2024; Wang et al., 2024; Tang et al., 2024). Thus, validating this influential analogy is indispensable for defining the future landscape of research in adversarial transferability. Interestingly, our paper sheds light on this insight in several ways.

First, the mathematical formulations in Lemma 1 is similar to generalization error (Vapnik, 1998; Bousquet & Elisseeff, 2002) , which also derives an objective as a difference between the population risk and the empirical risk. Such similarity between transferability error and generalization error suggests the possible validity of the analogy. Also, Lemma 2 is similar to the bound of the original Rademacher complexity (Golowich et al., 2018), which also suggests that obtaining a larger training set as well as a less complex model contribute a tighter bound of Rademacher complexity. Such similarities between transferability error and generalization error suggests the possible validity of the analogy. More importantly, if the analogy is correct, then recall that in the conventional framework of learning theory: (1) increasing the size of training set typically leads to a better generalization of the model (Bousquet & Elisseeff, 2002); (2) improving the diversity among ensemble classifiers makes it more advantageous for better generalization (Ortega et al., 2022); and (3) reducing the model complexity (Cherkassky, 2002) benefits the generalization ability. It is natural to ask: In model ensemble attack, do (1) incorporating more surrogate models, (2) making them more diverse, and (3) reducing their model complexity theoretically result in better adversarial transferability?

In Section 4, our theoretical framework provides consistently affirmative responses to the above question as well as the analogy. Considering a higher perspective, the theory is also instructive in two ways. On the one hand, from the perspective of a theoretical researcher, the extensive and advanced generalization theory may yield enlightening insights in the field of adversarial transferability. On the other hand, from an practitioner's point of view, ideas from deep learning algorithms can also be leveraged to develop more effective transferable attack algorithms.

D.8 Vulnerability-diversity Trade-off Curve

The relationship between vulnerability and diversity, as discussed in Section 5, merits deeper exploration. Drawing on the parallels between the vulnerability-diversity trade-off and the bias-variance trade-off (Geman et al., 1992), we find that insights from the latter may prove valuable for understanding the former, and warrant further investigation.

The classical bias-variance trade-off suggests that as model complexity increases, bias decreases while variance rises, resulting in a U-shaped test error curve. However, recent studies have revealed additional phenomena and provided deeper analysis (Neal et al., 2018; Neal, 2019; Derumigny & Schmidt-Hieber, 2023), such as the double descent (Belkin et al., 2019; Nakkiran et al., 2021).

Our experiments indicate that diversity does not follow the same pattern as variance in classical bias-variance trade-off. Nonetheless, there are indications within the bias-variance trade-off literature that suggest similar behavior might occur. For instance, Yang et al. (2020) proposes that variance exhibits a bell-shaped curve, initially increasing and then decreasing as network width grows. Additionally, Lin & Dobriban (2021) offers a meticulous understanding of variance through detailed decomposition, highlighting the influence of factors such as initialization, label noise, and training data. Overall, the trend of variance in model ensemble attack remains a valuable area for future research. We may borrow insights from machine learning literature to get a better understanding of this.

D.9 Insight for Model Ensemble Defense

While our paper primarily focuses on analyzing model ensemble attacks, our theoretical findings can also provide valuable insights for model ensemble defenses:

From a theoretical perspective. The vulnerability-diversity decomposition introduced for model ensemble attacks can likewise be extended to model ensemble defenses. Mathematically, this results in a decomposition similar to conclusions in ensemble learning (see Proposition 3 in Wood et al. (2023) and Theorem 1 in Ortega et al. (2022)), which shows that within the adversarial perturbation region,

$$\text{Expected loss} \leq \text{Empirical ensemble loss} - \text{Diversity}.$$

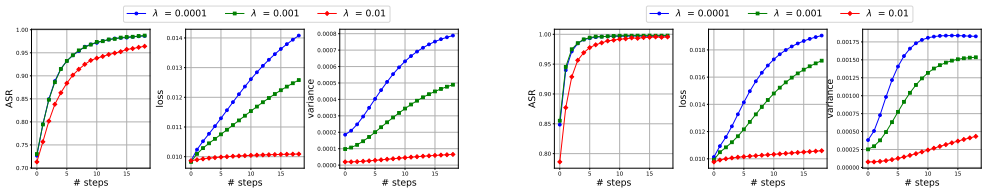
Thus, to improve model robustness (reduce the expected loss within the perturbation region), the core strategy involves minimizing the ensemble defender's loss or increasing diversity.

However, there is also an inherent trade-off between these two objectives: when the ensemble loss is sufficiently small, the model may overfit to the adversarial region, potentially reducing diversity; conversely, when diversity is maximized, the model may underfit the adversarial region, potentially increasing the ensemble loss. Therefore, from this perspective, our work provides meaningful insights for adversarial defense that warrant further analysis.

From an algorithmic perspective. We can consider recently proposed diversity metrics, such as Vendi score (Friedman & Dieng, 2022) and EigenScore (Chen et al., 2024a). Following the methodology outlined in Deng & Mu (2023), diversity can be incorporated into the defense optimization objective to strike a balance between diversity and ensemble loss. By finding an appropriate trade-off between these two factors, the effectiveness of ensemble defense may be enhanced.

E EVALUATION ON THE CIFAR-100 DATASET

Following the same setting in our experiments, we further validate the vulnerability-diversity decomposition on the CIFAR-100 (Krizhevsky et al., 2009) dataset. The results are shown in fig. 6.



(a) MLP (b) CNN

Figure 6: Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the CIFAR-100 dataset.

As the model becomes stronger (i.e., a smaller λ), the three metrics (ASR, loss and variance) increases, validating the soundness of vulnerability-diversity decomposition.