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:

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

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

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

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

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*

$$\Re_N(\mathcal{Z}) \leq \frac{\sqrt{m}BB'C}{\sqrt{N}}.$$

Proof. We have

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

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

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

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 $\mathcal{F}$ and any convex and monotonically increasing function $g : \mathbb{R} \to [0, \infty)$, 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

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

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,

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

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

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

$$
\begin{aligned}
\mathbb{E}Z &= \prod_{i=1}^{l-1}T_i\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{N}\sigma_i W_{i,1}^{*}x\right\|^2\right]}\\
&= \prod_{i=1}^{l-1}T_i\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i=j}^{N}\sigma_i\sigma_j\left(W_{i,1}^{*}x\right)^T\left(W_{j,1}^{*}x\right)\right]}\\
&\leq \prod_{i=1}^{l-1}T_i\left(T_1B\sqrt{N}\right)\\
&= B\sqrt{N}\prod_{i=1}^{l}T_i
\end{aligned}
$$

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

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

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*

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

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:

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

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:

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

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

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

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

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

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

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:

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

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., 2024).

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 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., 2019). 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.3 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., 2024). 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.4 Conflicting Opinions on "Diversity"

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.5 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.