Supplementary Material: Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability

Contents Summary

A. An Example Illustrating the Necessity of both α_1, α_2 in Characterizing the Relation Between Adversarial Transferability and Knowledge Transferability

α_1 and α_2 (Definition 1&2) represent complementary aspects of the adversarial transferability: α_1 can be understood as how often the adversarial attack transfers, while α_2 encodes directional information of the output deviation caused by adversarial attacks. Recall that $\alpha_1, \alpha_2 \in [0, 1]$ (higher values indicate better adversarial transferability). As we show in our theoretical results reveal that high α_1 alone is not enough, *i.e.*, both the proposed metrics are necessary to characterize adversarial transferability and the relation between adversarial and knowledge transferabilities.

We provide a one-dimensional example showing that large α_1 only is not enough to indicate high knowledge transferability. Suppose the ground truth target function $f_T(x) = x^2$, and the source function $f_S(x) = \texttt{sgn}(x) \cdot x^2$ where $\texttt{sgn}(\cdot)$ denotes the sign function. Let the adversarial loss be the deviation in function output, and the data distribution be the uniform distribution on $[-1, 1]$. As we can see, the direction that makes either f_T or f_S deviates the most is always the same, *i.e.*, in this example even with $\alpha_1 = 1$ achieves its maximum and adversarial attacks always transfer, regardless of the choice of $f_1 \to f_2$ or $f_2 \to f_1$. However, there does not exist an affine function g (*i.e.*, fine-tuning) making $g \circ f_S$ close to f_T on $[-1, 1]$. Indeed, one can verify that $\alpha_2 = 0$ in this case (either $f_1 \to f_2$ or $f_2 \to f_1$), which contributes to the low knowledge transferability. However, if we move the data distribution to $[0, 2]$, we can have $\alpha_1 = \alpha_2 = 1$ (either $f_1 \to f_2$ or $f_2 \to f_1$) indicating high adversarial transferability, and indeed it achieves $f_S = f_T$ showing perfect knowledge transferability.

B. Detailed Discussion About the Direction of Adversarial Transfer From $f_S \to f_T$ in Subsection 3.3

In this section, we present a detailed discussion, in addition to subsection 3.3, about the connection between function matching distance and knowledge transfer distance when the direction of adversarial transfer is from $f_S \to f_T$.

Recall that, to complete the story, it remains to connect the function matching distance to knowledge transferability. As the adversarial transfer is symmetric (*i.e.*, either from $f_S \to f_T$ or $f_T \to f_S$), we are able to use the placeholders $\star, \diamond \in \{S, T\}$ all the way through. However, as the knowledge transfer is asymmetric (*i.e.*, $f_S \to y$ to the target ground truth), we need to instantiate the direction of adversarial transfer to further our discussion. We have discussed the direction of adversarial transfer from $f_T \to f_S$ in the main paper, where we show that the function matching distance of this direction, *i.e.*,

$$\min_{g \in \mathbb{G}} \|f_T - g \circ f_S\|_{\mathcal{D}, \boldsymbol{H}_T}, \tag{16}$$

can both upper and lower bound the knowledge transfer distance, *i.e.*,

$$\min_{g \in \mathbb{G}} \|y - g \circ f_S\|_{\mathcal{D}, \boldsymbol{H}_T}. \tag{17}$$

The direction of adversarial transfer from $f_S \to f_T$ corresponds to $(\star, \diamond) = (S, T)$. Accordingly, the function matching distance (equation 13) becomes

$$\min_{g \in \mathbb{G}} \|f_S - g \circ f_T\|_{\mathcal{D}, \boldsymbol{H}_S}. \tag{18}$$

Since the affine transformation g acts on the target reference model, it can not be directly viewed as a surrogate transfer loss. However, interesting interpretations can be found in this direction, depending on the output dimension of $f_S : \mathbb{R}^n \to \mathbb{R}^m$ and $f_T : \mathbb{R}^n \to \mathbb{R}^d$.

In this subsection in the appendix we provide detailed discussion on the connection between the function matching distance of the direction of adversarial transfer from $f_S \to f_T$ (equation 18) and the knowledge transfer distance (equation 17). We build this connection by providing the relationships between the two directions of function matching distance, *i.e.*, equation 16 and equation 18. That is being said, since we know equation 17 and equation 16 are tied together, we only need to provide relationships between equation 16 and equation 18 to show the connection between equation 18 and equation 17.

Suppose $g : \mathbb{R}^d \to \mathbb{R}^m$ is full rank, and loosely speaking we can derive the following intuitions.

- If $d < m$, then g is injective and there exists $g^{-1} : \mathbb{R}^m \to \mathbb{R}^d$ such that $g^{-1} \circ g$ is the identity function. That is, if g can map f_T to closely track f_S, then reversely g^{-1} can map f_S to f_T, showing equation 18 upper bounds equation 16 in some sense.
- If $d > m$, then g is surjective. By symmetry, equation 16 upper bounds equation 18 in some sense.
- It is when $m = d$ that equation 16 and equation 18 coincide.

Formally, we have the following theorem.

Theorem B.1. *Denote $\tilde{g}_{T,S} : \mathbb{R}^m \to \mathbb{R}^d$ as the optimal solution of equation 16, and $\tilde{g}_{S,T} : \mathbb{R}^d \to \mathbb{R}^m$ as the optimal solution of equation 18. Suppose the two optimal affine maps $\tilde{g}_{T,S}, \tilde{g}_{S,T}$ are both full-rank. For $\boldsymbol{v} \in \mathbb{R}^m$, denote the matrix representation of $\tilde{g}_{T,S}$ as $\tilde{g}_{T,S}(\boldsymbol{v}) = \tilde{\boldsymbol{W}}_{T,S}\boldsymbol{v} + \tilde{\boldsymbol{b}}_{T,S}$. Similarly, for $\boldsymbol{w} \in \mathbb{R}^d$, denote the matrix representation of $\tilde{g}_{S,T}$ as $\tilde{g}_{S,T}(\boldsymbol{w}) = \tilde{\boldsymbol{W}}_{S,T}\boldsymbol{w} + \tilde{\boldsymbol{b}}_{S,T}$. We have the following statements.*

If $d < m$, then $\tilde{g}_{S,T}$ is injective, and we have:

$$\|f_T - \tilde{g}_{T,S} \circ f_S\|_{\mathcal{D}, \boldsymbol{H}_T} \leq \sqrt{\|(\tilde{\boldsymbol{W}}_{S,T}^\top \tilde{\boldsymbol{W}}_{S,T})^{-1}\|_F \cdot \|\boldsymbol{H}_T\|_F} \cdot \|f_S - \tilde{g}_{S,T} \circ f_T\|_{\mathcal{D}}. \tag{19}$$

If $d > m$, then $\tilde{g}_{T,S}$ is injective, and we have:

$$\|f_S - \tilde{g}_{S,T} \circ f_T\|_{\mathcal{D}, \boldsymbol{H}_S} \leq \sqrt{\|(\tilde{\boldsymbol{W}}_{T,S}^\top \tilde{\boldsymbol{W}}_{T,S})^{-1}\|_F \cdot \|\boldsymbol{H}_S\|_F} \cdot \|f_T - \tilde{g}_{T,S} \circ f_S\|_{\mathcal{D}}. \tag{20}$$

If $d = m$, then both $\tilde{g}_{S,T}$ and $\tilde{g}_{T,S}$ are bijective, and we have both (19) and (20) stand.

That is, when the direction of adversarial transfer is from $f_S \to f_T$, the indicating relation between the function matching distance if this direction (equation 18) and knowledge transferability would possibly be unidirectional, depending on the dimensions.

C. Proofs in Section 2

In this section, we present proofs for Proposition 2.1 and Proposition 2.2.

C.1. Proof of Proposition 2.1

Proposition C.1 (Proposition 2.1 Restated)**.** *The $\alpha_2^{f_1 \to f_2}$ can be reformulated as*

$$(\alpha_2^{f_1 \to f_2})^2 = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2} \left[\theta_{f_1 \to f_1}(\boldsymbol{x}_1, \boldsymbol{x}_2) \theta_{f_1 \to f_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) \right],$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2 \overset{i.i.d.}{\sim} \mathcal{D}$, and

$$\theta_{f_1 \to f_1}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1), \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2) \rangle$$
$$\theta_{f_1 \to f_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1), \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \rangle$$

Proof. Recall that we want to show

$$\left\| \mathbb{E}_{\boldsymbol{x}} [\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x})^\top] \right\|_F^2 = (\alpha_2^{f_1 \to f_2})^2 = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2} \left[\theta_{f_1 \to f_1}(\boldsymbol{x}_1, \boldsymbol{x}_2) \theta_{f_1 \to f_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) \right],$$

and the proof of this proposition is done by applying some trace tricks, as shown below.

$$\begin{aligned}
\theta_{f_1 \to f_1}(\boldsymbol{x}_1, \boldsymbol{x}_2) \theta_{f_1 \to f_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \langle \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1), \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2) \rangle \cdot \langle \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1), \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \rangle \\
&= \langle \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2), \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \rangle \cdot \langle \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1), \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \rangle \\
&= \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \\
&= \operatorname{tr} \left(\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \right) \\
&= \operatorname{tr} \left(\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \right) \qquad (21)
\end{aligned}$$

Plugging equation 21 into equation 25, we have

$$\begin{aligned}
(\alpha_2^{f_1 \to f_2})^2 &= \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2} \left[\operatorname{tr} \left(\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \right) \right] \\
&= \operatorname{tr} \left(\mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2} \left[\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \right] \right) \\
&= \operatorname{tr} \left(\mathbb{E}_{\boldsymbol{x}_1} \left[\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_1) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_1)^\top \right] \cdot \mathbb{E}_{\boldsymbol{x}_2} \left[\widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}_2) \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}_2)^\top \right] \right), \qquad (22)
\end{aligned}$$

where the last equality is because that $\boldsymbol{x}_1, \boldsymbol{x}_2$ are $i.i.d.$ samples from the same distribution.

Therefore, we can re-write the $\boldsymbol{x}_1, \boldsymbol{x}_2$ to be the same $\boldsymbol{x} \sim \mathcal{D}$ and realize that the two matrices are in fact the same one.

$$\begin{aligned}
(22) &= \operatorname{tr} \left(\mathbb{E}_{\boldsymbol{x}} \left[\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x})^\top \right] \cdot \mathbb{E}_{\boldsymbol{x}} \left[\widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x}) \widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x})^\top \right] \right) \\
&= \left\| \mathbb{E}_{\boldsymbol{x}} [\widehat{\Delta_{f_1 \to f_1}}(\boldsymbol{x}) \widehat{\Delta_{f_1 \to f_2}}(\boldsymbol{x})^\top] \right\|_F^2.
\end{aligned}$$

□

C.2. Proof of Proposition 2.2

Proposition C.2 (Proposition 2.2 Restated)**.** *The adversarial transferability metrics* $\alpha_1^{f_1\to f_2}$, $\alpha_2^{f_1\to f_2}$ *and* $(\alpha_1 * \alpha_2)^{f_1\to f_2}$ *are in* $[0,1]$.

Proof. Let us begin with

$$\alpha_1^{f_1\to f_2}(\boldsymbol{x}) = \frac{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})))}{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_2,\epsilon}(\boldsymbol{x})))}.$$

Recall that $\ell_{adv}(\cdot) \geq 0$, and the definition of adversarial attack:

$$\boldsymbol{\delta}_{f,\epsilon}(\boldsymbol{x}) \;=\; \underset{\|\boldsymbol{\delta}\|\leq\epsilon}{\arg\max}\; \ell_{adv}(f(\boldsymbol{x}), f(\boldsymbol{x}+\boldsymbol{\delta})),$$

and we can see that by definition,

$$0 \leq \ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x}))) \leq \ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_2,\epsilon}(\boldsymbol{x}))).$$

Therefore,

$$0 \leq \frac{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})))}{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_2,\epsilon}(\boldsymbol{x})))} \leq 1,$$

where we define $0/0 = 0$ if necessary.

Hence, $\alpha_1^{f_1\to f_2} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\alpha_1^{f_1\to f_2}(\boldsymbol{x})]$ is also in $[0,1]$.

Next, we use Proposition 2.1 to prove the same property for $\alpha_2^{f_1\to f_2}$. Note that

$$(\alpha_2^{f_1\to f_2})^2 = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\langle\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_2)}\rangle \cdot \langle\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_2)}\rangle\right] \tag{23}$$

is the expectation of the product of two inner products, where each inner product is of two unit-length vector. That is being said, $\langle\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_2)}\rangle \in [-1,1]$ and $\langle\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_2)}\rangle \in [-1,1]$. Therefore, we know that

$$\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\langle\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_2)}\rangle \cdot \langle\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_2)}\rangle\right] \in [-1,1].$$

In addition, we know from equation 23 that it is non-negative, and hence

$$(\alpha_2^{f_1\to f_2})^2 \in [0,1].$$

As $\alpha_2^{f_1\to f_2}$ itself is also non-negative by definition, we can see that $\alpha_2^{f_1\to f_2} \in [0,1]$.

Finally, we move to prove $(\alpha_1 * \alpha_2)^{f_1\to f_2} \in [0,1]$. Recall that

$$(\alpha_1 * \alpha_2)^{f_1\to f_2} = \left\|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\alpha_1^{f_1\to f_2}(\boldsymbol{x})\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x})}\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x})}^\top]\right\|_F.$$

If we see $\alpha_1^{f_1\to f_2}(\boldsymbol{x})\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x})}$ as a whole, we can show exactly the same as the Proposition 2.1 that

$$((\alpha_1 * \alpha_2)^{f_1\to f_2})^2 = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2)\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\right], \tag{24}$$

where

$$\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \langle\alpha_1^{f_1\to f_2}(\boldsymbol{x}_1)\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_1)}, \alpha_1^{f_1\to f_2}(\boldsymbol{x}_2)\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_2)}\rangle$$
$$\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \langle\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_2)}\rangle.$$

Similarly, as $\alpha_1^{f_1\to f_2}(\boldsymbol{x}) \in [0,1]$, we can see that $\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2)\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) \in [-1,1]$, and hence

$$\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2)\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\right] \in [-1,1].$$

Noting that equation 24 is non-negative, we conclude that

$$((\alpha_1 * \alpha_2)^{f_1\to f_2})^2 \in [0,1].$$

Since $(\alpha_1 * \alpha_2)^{f_1\to f_2}$ itself is non-negative as well, we can see that $(\alpha_1 * \alpha_2)^{f_1\to f_2} \in [0,1]$.

Therefore, the three adversarial transferability metrics are all within $[0,1]$. □

D. Proofs in Section 3

In this section, we prove the two theorems and the two propositions presented in section 3, which are our main theories.

D.1. Proof of Theorem 3.1

We introduce two lemmas before proving Theorem 3.1.

Lemma D.1. *The square of the gradient matching distance is*

$$\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 = \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \langle \boldsymbol{P}^\top \boldsymbol{H}_\star \boldsymbol{P}, \boldsymbol{J}^\dagger\rangle,$$

where $g\in\mathbb{G}$ are affine transformations, and

$$\boldsymbol{P} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top \nabla f_\diamond(\boldsymbol{x})\right], \qquad \boldsymbol{J} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top \nabla f_\diamond(\boldsymbol{x})\right].$$

Proof.

$$\begin{aligned}\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 &= \min_{\boldsymbol{W}} \quad \|\nabla f_\star^\top - \boldsymbol{W}\nabla f_\diamond^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \\ &= \min_{\boldsymbol{W}} \quad \mathbb{E}_{\boldsymbol{x}\in\mathcal{D}}\|\nabla f_\star(\boldsymbol{x})^\top - \boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2, \end{aligned} \tag{25}$$

where $\boldsymbol{W}$ is a matrix.

We can see that (25) is a convex program, where the optimal solution exists in a closed-form form, as shown in the following. Denote $l(\boldsymbol{W}) = \|\nabla f_\star(\boldsymbol{x})^\top - \boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2$, we have

$$\begin{aligned} l(\boldsymbol{W}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\nabla f_\star(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 + \|\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 - 2\langle \nabla f_\star(\boldsymbol{x})^\top, \boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\rangle_{\boldsymbol{H}_\star}\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\nabla f_\star(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 + \operatorname{tr}\left(\nabla f_\diamond(\boldsymbol{x})\boldsymbol{W}^\top\boldsymbol{H}_\star\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\right) - 2\operatorname{tr}\left(\nabla f_\star(\boldsymbol{x})\boldsymbol{H}_\star\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\nabla f_\star(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 + \operatorname{tr}\left(\boldsymbol{H}_\star\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\boldsymbol{W}^\top\right) - 2\operatorname{tr}\left(\boldsymbol{H}_\star\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\star(\boldsymbol{x})\right)\right]. \end{aligned}$$

Taking the derivative of $l(\cdot)$ w.r.t. $\boldsymbol{W}$, we have

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{W}} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[2\boldsymbol{H}_\star\left(\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top - \nabla f_\star(\boldsymbol{x})^\top\right)\nabla f_\diamond(\boldsymbol{x})\right] \\ &= 2\boldsymbol{H}_\star\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{W}\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x}) - \nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right] \\ &= 2\boldsymbol{H}_\star\left(\boldsymbol{W}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right]\right). \end{aligned} \tag{26}$$

Since $l(\cdot)$ is convex, if there exists a $\tilde{\boldsymbol{W}}$ such that $\frac{\partial l}{\partial \boldsymbol{W}}\big|_{\boldsymbol{W}=\tilde{\boldsymbol{W}}} = \boldsymbol{0}$ then we know that $\tilde{\boldsymbol{W}}$ is an optimal solution. Luckily, we can find such solution easily by using pseudo inverse, *i.e.*,

$$\begin{aligned}\tilde{\boldsymbol{W}} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right]\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right]\right)^\dagger \\ &= \boldsymbol{P}\boldsymbol{J}^\dagger, \end{aligned} \tag{27}$$

where we denote $\boldsymbol{P} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right]$ and $\boldsymbol{J} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right]$.

We can verify that such $\tilde{\boldsymbol{W}}$ indeed make the partial derivative (equation 26) zero. In equation 26, we have

$$\tilde{\boldsymbol{W}}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right] = \boldsymbol{P}\boldsymbol{J}^\dagger\boldsymbol{J} - \boldsymbol{P}. \tag{28}$$

To continue, we can see from Lemma E.2 that $\ker(\boldsymbol{J}) \subseteq \ker(\boldsymbol{P})$ which means $\operatorname{rowsp}(\boldsymbol{P}) \subseteq \operatorname{rowsp}(\boldsymbol{J})$, where $\ker(\cdot)$ denotes the kernel of a matrix, and $\operatorname{rowsp}(\cdot)$ denotes the row space of a matrix. Therefore, by definition of the pseudo-inverse, we can see that $\boldsymbol{P}\boldsymbol{J}^\dagger\boldsymbol{J} = \boldsymbol{P}$, *i.e.*, (28) $= \boldsymbol{0}$, and hence $\tilde{\boldsymbol{W}}$ is indeed the optimal solution.

Plugging (27) into (25), we have the optimal value as

$$\begin{aligned}
(25) &= l(\tilde{\boldsymbol{W}}) \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\nabla f_\star(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 + \operatorname{tr}\left(\boldsymbol{H}_\star\tilde{\boldsymbol{W}}\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\tilde{\boldsymbol{W}}^\top\right) - 2\operatorname{tr}\left(\boldsymbol{H}_\star\tilde{\boldsymbol{W}}\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\star(\boldsymbol{x})\right)\right] \\
&= \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 + \operatorname{tr}\left(\boldsymbol{H}_\star\tilde{\boldsymbol{W}}\boldsymbol{J}\tilde{\boldsymbol{W}}^\top - 2\boldsymbol{H}_\star\tilde{\boldsymbol{W}}\boldsymbol{P}^\top\right) \\
&= \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 + \operatorname{tr}\left(\boldsymbol{H}_\star\boldsymbol{P}\boldsymbol{J}^\dagger\boldsymbol{J}\boldsymbol{J}^\dagger\boldsymbol{P}^\top - 2\boldsymbol{H}_\star\boldsymbol{P}\boldsymbol{J}^\dagger\boldsymbol{P}^\top\right) \\
&= \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \operatorname{tr}\left(\boldsymbol{H}_\star\boldsymbol{P}\boldsymbol{J}^\dagger\boldsymbol{P}^\top\right) \\
&= \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \langle\boldsymbol{P}^\top\boldsymbol{H}_\star\boldsymbol{P}, \boldsymbol{J}^\dagger\rangle.
\end{aligned}$$

□

Next, we present another lemma to analyze the term $\boldsymbol{P}^\top\boldsymbol{H}_\star\boldsymbol{P}$.

Lemma D.2. *In this lemma, we break down the matrix representation of $\boldsymbol{P}^\top\boldsymbol{H}_\star\boldsymbol{P}$ into pieces relating to the output deviation caused by the generalized adversarial attacks (defined in equation 10)*

$$\boldsymbol{P}^\top\boldsymbol{H}_\star\boldsymbol{P} = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{i.i.d.}{\sim}\mathcal{D}}\sum_{i,j=1}^{n}\left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top\boldsymbol{H}_\star\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)\cdot\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top\right).$$

Proof. Denote a symmetric decomposition of the positive semi-definitive matrix $\boldsymbol{H}_\star$ as

$$\boldsymbol{H}_\star = \boldsymbol{T}^\top\boldsymbol{T},$$

where $\boldsymbol{T}$ is of the same dimension of $\boldsymbol{H}_\star$. We note that the choice of decomposition does not matter.

Then, plugging in the definition of $\boldsymbol{P}$, we can see that

$$\begin{aligned}
\boldsymbol{P}^\top\boldsymbol{H}_\star\boldsymbol{P} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\star(\boldsymbol{x})\right]\cdot\boldsymbol{T}^\top\boldsymbol{T}\cdot\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\star(\boldsymbol{x})\boldsymbol{T}^\top\right]\cdot\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})\right].
\end{aligned} \tag{29}$$

A key observation to connect the above equation to the adversarial attack (equation 5) is that,

$$\begin{aligned}
\boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x}) &= \operatorname*{arg\,max}_{\|\boldsymbol{\delta}\|_2\le\epsilon}\|\nabla f_\star(\boldsymbol{x})^\top\boldsymbol{\delta}\|_{\boldsymbol{H}_\star} \\
&= \operatorname*{arg\,max}_{\|\boldsymbol{\delta}\|_2\le\epsilon}\|\nabla f_\star(\boldsymbol{x})^\top\boldsymbol{\delta}\|_{\boldsymbol{H}_\star}^2 \\
&= \operatorname*{arg\,max}_{\|\boldsymbol{\delta}\|_2\le\epsilon}\boldsymbol{\delta}^\top\nabla f_\star(\boldsymbol{x})\boldsymbol{H}_\star\nabla f_\star(\boldsymbol{x})^\top\boldsymbol{\delta} \\
&= \operatorname*{arg\,max}_{\|\boldsymbol{\delta}\|_2\le\epsilon}\|\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top\boldsymbol{\delta}\|_2^2.
\end{aligned}$$

That is being said, the adversarial attack is the right singular vector corresponding to the largest singular value (in absolute value) of $\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top$.

Similarly, we can see the singular values $\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x})\in\mathbb{R}^n$, defined as the descending (in absolute value) singular values of the Jacobian $\nabla f_\star(\boldsymbol{x})^\top\in\mathbb{R}^{\cdot\times n}$ in the $\boldsymbol{H}_\star$ inner product space (equation 8), are the singular values of $\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top$.

With this perspective, if we write down the singular value decomposition of $\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top$, *i.e.*,

$$\boldsymbol{T}\nabla f_\star(\boldsymbol{x})^\top = \boldsymbol{U}_\star(\boldsymbol{x})\Sigma_\star(\boldsymbol{x})\boldsymbol{V}_\star^\top(\boldsymbol{x}),$$

we can observe that:

1. $\Sigma_\star(\boldsymbol{x})$ is diagonalized singular values $\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x})$;
2. The i^{th} column of $\boldsymbol{V}_\star(\boldsymbol{x})$ is the i^{th} generalized attack $\boldsymbol{\delta}^{(i)}_{f_\star}(\boldsymbol{x})$ (defined in equation 9);

3. The i^{th} column of $\boldsymbol{U}_\star(\boldsymbol{x})\Sigma(\boldsymbol{x})$ is $\boldsymbol{T}\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})$ where $\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})$ is the output deviation (defined in equation 10);

4. The i^{th} column of $\nabla f_\diamond(\boldsymbol{x})^\top \boldsymbol{V}_\star(\boldsymbol{x})$ is the output deviation $\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x})$ (defined in equation 10).

With the four key observations, we can break down the Jacobian matrices as

$$\begin{aligned}\nabla f_\diamond(\boldsymbol{x})^\top \nabla f_\star(\boldsymbol{x})\boldsymbol{T}^\top &= \left(\Delta^{(1)}_{f_\star\to f_\diamond}(\boldsymbol{x})\dots\Delta^{(n)}_{f_\star\to f_\diamond}(\boldsymbol{x})\right)\begin{pmatrix}\Delta^{(1)}_{f_\star\to f_\star}(\boldsymbol{x})^\top\boldsymbol{T}^\top\\ \vdots\\ \Delta^{(n)}_{f_\star\to f_\star}(\boldsymbol{x})^\top\boldsymbol{T}^\top\end{pmatrix}\\ &= \sum_{i=1}^n \Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x})\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})^\top\boldsymbol{T}^\top.\end{aligned}$$

Therefore, plugging it into the equation 29, we have

$$\begin{aligned}(29) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{i=1}^n \Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x})\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})^\top\boldsymbol{T}^\top\right]\cdot\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{i=1}^n \boldsymbol{T}\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x})^\top\right]\\ &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}}\left[\sum_{i=1}^n\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top\boldsymbol{T}^\top\right)\sum_{j=1}^n\left(\boldsymbol{T}\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top\right)\right]\\ &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}}\sum_{i,j=1}^n\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top\boldsymbol{H}_\star\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top\right)\\ &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}}\sum_{i,j=1}^n\left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top\boldsymbol{H}_\star\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)\cdot\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top\right),\end{aligned}$$

where the last equality is due to that $\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top\boldsymbol{H}_\star\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)$ is a scalar value.

□

Equipped with Lemma D.1 and Lemma D.2, we are able to prove the Theorem 3.1.

Theorem D.1 (Theorem 3.1 Restated)**.** *Given the target and source models $f_\star, f_\diamond$, where $(\star,\diamond)\in\{(S,T),(T,S)\}$, the gradient matching distance (equation 7) can be written as*

$$\min_{g\in\mathbb{G}}\quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star} = \left(1-\frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|^2_{\mathcal{D},\boldsymbol{H}_\star}\cdot\|\boldsymbol{J}^\dagger\|^{-1}_{\boldsymbol{H}_\diamond}}\right)^{\frac{1}{2}}\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star},$$

where the expectation is taken over $\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}$, and

$$\begin{aligned}\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}) &= \sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x})\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x})\odot\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})\\ \boldsymbol{J} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\nabla f_\diamond(\boldsymbol{x})^\top\nabla f_\diamond(\boldsymbol{x})].\end{aligned}$$

Moreover, $\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is a matrix, and its element in the i^{th} row and j^{th} column is

$$\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{(i,j)} = \langle\widehat{\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\star},\widehat{\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\star}\rangle\cdot\langle\widehat{\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\diamond},\widehat{\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\diamond}\rangle_{\widehat{\boldsymbol{J}^\dagger}|_{\boldsymbol{H}_\diamond}}.$$

Proof. Combining the result from Lemma D.1 and Lemma D.2, and applying the linearity of the inner product, we have

$$
\begin{aligned}
&\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \\
=&\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \left\langle \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}} \sum_{i,j=1}^{n} \left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top \boldsymbol{H}_\star \Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)\cdot\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top\right), \boldsymbol{J}^\dagger\right\rangle \\
=&\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}} \sum_{i,j=1}^{n} \left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top \boldsymbol{H}_\star \Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)\cdot\left\langle\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top, \boldsymbol{J}^\dagger\right\rangle \\
=&\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}} \sum_{i,j=1}^{n} \left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top \boldsymbol{H}_\star \Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)\cdot\operatorname{tr}\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)^\top \boldsymbol{J}^\dagger\right) \\
=&\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 - \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}} \sum_{i,j=1}^{n} \underbrace{\left(\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)^\top \boldsymbol{H}_\star \Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)\right)}_{X_1}\cdot\underbrace{\left(\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{J}^\dagger \Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)\right)}_{X_2}.
\end{aligned}
\tag{30}
$$

As the generalized first adversarial transferability $\boldsymbol{A}_1$ is about the magnitude of the output deviation (defined in equation 11), and we can separate the $\boldsymbol{A}_1$ out from the above equation. Then, what left should be about the directions about the output deviation, which we will put into the matrix $\boldsymbol{A}_2$, *i.e.*, the generalized second adversarial transferability.

Recall that the generalized the first adversarial transferability is a n-dimensional vector $\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})$ including the adversarial losses of all of the generalized adversarial attacks, where the i^{th} element in the vector is

$$
\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})^{(i)} = \frac{\|\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}{\|\nabla f_\diamond(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}.
$$

Moreover, to connect the magnitude of the output deviation to the generalized singular values (equation 9), we have

$$
\|\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x})\|_{\boldsymbol{H}_\star} = \|\nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}^{(i)}_{f_\star}(\boldsymbol{x})\|_{\boldsymbol{H}_\star} = \sigma^{(i)}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}),
$$

and similarly,

$$
\|\nabla f_\diamond(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond} = \|\nabla f_\diamond(\boldsymbol{x})\boldsymbol{\delta}^{(1)}_{f_\diamond}(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond} = \sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}).
$$

Therefore, we can finally rewrite the X_1, X_2 in equation 30 as

$$
\begin{aligned}
X_1 &= \sigma^{(i)}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_1)\sigma^{(j)}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_2)\cdot\langle\widehat{\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\star}, \widehat{\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\star}\rangle \\
X_2 &= \boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_1)^{(i)}\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_2)^{(j)}\cdot\langle\widehat{\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\diamond}, \widehat{\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\diamond}\rangle_{\widehat{\boldsymbol{J}^\dagger|_{\boldsymbol{H}_\diamond}}}\cdot\sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_1)\sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_2)\|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}.
\end{aligned}
$$

Recall the $(i,j)^{th}$ entry of the matrix $\boldsymbol{A}_2$ is

$$
\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{(i,j)} = \langle\widehat{\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\star}, \widehat{\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\star}\rangle\cdot\langle\widehat{\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\diamond}, \widehat{\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\diamond}\rangle_{\widehat{\boldsymbol{J}^\dagger|_{\boldsymbol{H}_\diamond}}}.
$$

We can write

$$
X_1X_2 = \sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_1)\sigma^{(i)}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_1)\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_1)^{(i)}\cdot\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{(i,j)}\cdot\sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_2)\sigma^{(j)}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_2)\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_2)^{(j)}\|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}.
$$

Plugging the above into equation 30, and rearranging the double summation, we have

$$
\begin{aligned}
&(30) = \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \\
&-\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\overset{\text{i.i.d.}}{\sim}\mathcal{D}}\left[(\sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_1)\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_1)\odot\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_1))^\top\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)(\sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x}_2)\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}_2)\odot\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}_2))\right]\|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}.
\end{aligned}
\tag{31}
$$

Denoting

$$
\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}) = \sigma^{(1)}_{f_\diamond,\boldsymbol{H}_\diamond}(\boldsymbol{x})\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x})\odot\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}),
$$

and rearranging equation 31 give us the Theorem 3.1.

□

D.2. Proof of Proposition 3.1

From the proof of Theorem 3.1 in the above subsection, we can see why this proposition holds.

Proposition D.1 (Proposition 3.1 Restated)**.** *In Theorem 3.1,*

$$0 \leq \frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \cdot \|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}^{-1}} \leq 1.$$

Proof. Recall Theorem 3.1 states

$$\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star} = \left(1 - \frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \cdot \|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}^{-1}}\right)^{\frac{1}{2}} \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}.$$

We can see that the ≤ 1 part stands, since $\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}$ is always non-negative.

The ≥ 0 part can be proved by observing

$$\begin{aligned}\left(1 - \frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \cdot \|\boldsymbol{J}^\dagger\|_{\boldsymbol{H}_\diamond}^{-1}}\right)^{\frac{1}{2}} \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star} &= \min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star} \\ &\leq \|\nabla f_\star^\top - \nabla(0\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star} = \|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}\end{aligned}$$

□

D.3. Proof of Theorem 3.2

We introduce two lemmas before proving Theorem 3.2.

Lemma D.3. *Assume that function $h(\cdot)$ satisfies the β-smoothness under $\|\cdot\|_{\boldsymbol{H}_\star}$ norm (Assumption 1), and assume there is a vector $\boldsymbol{x}_0$ in the same space as $\boldsymbol{x}\sim\mathcal{D}$ such that $h(\boldsymbol{x}_0)=0$. Given $\tau>0$, there exists $\boldsymbol{x}'$ as a function of $\boldsymbol{x}$ such that $\|\boldsymbol{x}-\boldsymbol{x}'\|_2\leq\tau$, and*

$$\|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star}^2 \leq 2\left(\|\nabla h(\boldsymbol{x}')^\top\|_{\boldsymbol{H}_\star}^2 + \beta^2\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2-\tau\right)_+^2\right)\cdot\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2,$$

where the $(\cdot)_+$ is an operator defined by $\forall x\in\mathbb{R}$: $(x)_+ = x$ if $x\geq 0$ and $(x)_+ = 0$ otherwise.

Proof. To begin with, we note that the assumption of $h(\boldsymbol{x}_0)=0$ is only used for this lemma, and the assumption will be naturally guaranteed when we invoke this lemma in the proof of Theorem 3.2.

With the smoothness assumption, we know that $h(\cdot)$ has continuous gradient. Thus, we have

$$\|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} = \|h(\boldsymbol{x}) - h(\boldsymbol{x}_0)\|_{\boldsymbol{H}_\star} = \|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top(\boldsymbol{x}-\boldsymbol{x}_0)\|_{\boldsymbol{H}_\star},$$

where the last equation is by mean value theorem and thus $\xi\in(0,1)$.

Then, noting that $\|\cdot\|_{\boldsymbol{H}_\star}$ and $\|\cdot\|_2$ are compatible (Lemma E.1), we have

$$\|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top(\boldsymbol{x}-\boldsymbol{x}_0)\|_{\boldsymbol{H}_\star} \leq \|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top\|_{\boldsymbol{H}_\star}\cdot\|(\boldsymbol{x}-\boldsymbol{x}_0)\|_2.$$

Now we discuss two cases to define a random variable $\boldsymbol{x}'$ as a function of $\boldsymbol{x}$.

If $(1-\xi)\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\leq\tau$, we define $\boldsymbol{x}'$ as

$$\boldsymbol{x}' = \boldsymbol{x}_0 + \xi(\boldsymbol{x}-\boldsymbol{x}_0),$$

and we can see that $\|\boldsymbol{x}'-\boldsymbol{x}\|_2\leq\tau$.

Otherwise, *i.e.*, $(1-\xi)\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 > \tau$, we apply triangle inequality to derive

$$\begin{aligned}
&\|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top(\boldsymbol{x}-\boldsymbol{x}_0)\|_{\boldsymbol{H}_\star} \\
&\quad= \|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top - \nabla h(\boldsymbol{x}-\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}))^\top + \nabla h(\boldsymbol{x}-\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}))^\top\|_{\boldsymbol{H}_\star} \\
&\quad\le \underbrace{\|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top - \nabla h(\boldsymbol{x}-\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}))^\top\|_{\boldsymbol{H}_\star}}_{X} + \|\nabla h(\boldsymbol{x}-\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}))^\top\|_{\boldsymbol{H}_\star},
\end{aligned}$$

where we define

$$\boldsymbol{x}' = \boldsymbol{x}-\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}).$$

By definition, in this case $\|\boldsymbol{x}'-\boldsymbol{x}\|_2 \le \tau$ as well. We then treat X: it can be bounded using β-smoothness, *i.e.*,

$$\begin{aligned}
X &\le \beta\|\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0)-\boldsymbol{x}+\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}))\|_2 \\
&= \beta\|\tau(\widehat{\boldsymbol{x}-\boldsymbol{x}_0}) - (1-\xi)(\boldsymbol{x}-\boldsymbol{x}_0))\|_2 \\
&= \beta\left|\tau - (1-\xi)\cdot\|(\boldsymbol{x}-\boldsymbol{x}_0)\|_2\right| \\
&= \beta\left((1-\xi)\cdot\|(\boldsymbol{x}-\boldsymbol{x}_0)\|_2 - \tau\right),
\end{aligned}$$

where the last step is because we are exactly considering the case of $(1-\xi)\cdot\|(\boldsymbol{x}-\boldsymbol{x}_0)\|_2 > \tau$.

Therefore, combining the two cases together, we can write

$$\|\nabla h(\boldsymbol{x}_0+\xi(\boldsymbol{x}-\boldsymbol{x}_0))^\top\|_{\boldsymbol{H}_\star} \le \beta\left((1-\xi)\cdot\|(\boldsymbol{x}-\boldsymbol{x}_0)\|_2-\tau\right)_+ + \|\nabla h(\boldsymbol{x}')^\top\|_{\boldsymbol{H}_\star},$$

where $\|\boldsymbol{x}-\boldsymbol{x}'\| \le \tau$.

Combining the above, we have

$$\|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} \le \left(\|\nabla h(\boldsymbol{x}')^\top\|_{\boldsymbol{H}_\star} + \beta\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2-\tau\right)_+\right)\cdot\|\boldsymbol{x}-\boldsymbol{x}_0\|_2.$$

Take the square on both sides, and apply the Cauchy-Schwarz inequality, we have the lemma proved.

$$\begin{aligned}
\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} &\le \left(\|\nabla h(\boldsymbol{x}')^\top\|_{\boldsymbol{H}_\star} + \beta\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2-\tau\right)_+\right)^2\cdot\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \\
&\le 2\left(\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2-\tau\right)^2_+\right)\cdot\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2.
\end{aligned}$$

□

Lemma D.4. *Assume that function $h(\cdot)$ satisfies the β-smoothness under $\|\cdot\|_{\boldsymbol{H}_\star}$ norm (Assumption 1). Given $\tau > 0$, there exists $\boldsymbol{x}_i'$ as a function of $\boldsymbol{x}$ for $\forall i \in [n]$ such that $\|\boldsymbol{x}-\boldsymbol{x}_i'\|_2 \le \tau$, and*

$$\tau^2\cdot\|\nabla h(\boldsymbol{x})^\top\|^2_{\boldsymbol{H}_\star} \le 3\left(\sum_{i=1}^n \|h(\boldsymbol{x}_i')\|^2_{\boldsymbol{H}_\star} + n\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} + n\tau^4\beta^2\right).$$

Proof. Denote the dimension of $\boldsymbol{x}$ as n, and let $\boldsymbol{U}$ be an orthogonal matrix in $\mathbb{R}^{n\times n}$, where we denote its column vectors as $\boldsymbol{u}_i \in \mathbb{R}^n$ for $i \in [n]$. Applying the mean value theorem, there exists $\xi_i \in (0,1)$ such that

$$\begin{aligned}
h(\boldsymbol{x}+\tau\boldsymbol{u}_i) - h(\boldsymbol{x}) &= \nabla h(\boldsymbol{x}+\tau\xi_i\boldsymbol{u}_i)^\top\tau\boldsymbol{u}_i \\
&= \tau\left(\nabla h(\boldsymbol{x})^\top\boldsymbol{u}_i + (\nabla h(\boldsymbol{x}+\tau\xi_i\boldsymbol{u}_i)^\top - \nabla h(\boldsymbol{x})^\top)\boldsymbol{u}_i\right).
\end{aligned}$$

Rearranging the equality, we have

$$\nabla h(\boldsymbol{x})^\top\boldsymbol{u}_i = \frac{1}{\tau}\boldsymbol{\gamma}_i,$$

where we denote

$$\boldsymbol{\gamma}_i = h(\boldsymbol{x} + \tau \boldsymbol{u}_i) - h(\boldsymbol{x}) - \tau(\nabla h(\boldsymbol{x} + \tau\xi_i \boldsymbol{u}_i)^\top - \nabla h(\boldsymbol{x})^\top)\boldsymbol{u}_i.$$

Collecting each $\boldsymbol{\gamma}_i$ for $i \in [n]$ into a matrix $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1 ... \boldsymbol{\gamma}_n]$, we can re-formulate the above equality as

$$\begin{aligned}\tau \nabla h(\boldsymbol{x})^\top \boldsymbol{U} &= \boldsymbol{\Gamma} \\ \tau \nabla h(\boldsymbol{x})^\top &= \boldsymbol{\Gamma}\boldsymbol{U}^\top,\end{aligned}$$

where the last equality is because that $\boldsymbol{U}$ is orthogonal.

Taking the $\|\cdot\|^2_{\boldsymbol{H}_\star}$ on both sides, with some linear algebra manipulation we can derive

$$\begin{aligned}\tau^2 \cdot \|\nabla h(\boldsymbol{x})^\top\|^2_{\boldsymbol{H}_\star} &= \|\boldsymbol{\Gamma}\boldsymbol{U}^\top\|^2_{\boldsymbol{H}_\star} \\ &= \mathrm{tr}(\boldsymbol{U}\boldsymbol{\Gamma}^\top \boldsymbol{H}_\star \boldsymbol{\Gamma}\boldsymbol{U}^\top) = \mathrm{tr}(\boldsymbol{\Gamma}^\top \boldsymbol{H}_\star \boldsymbol{\Gamma}) = \mathrm{tr}(\boldsymbol{H}_\star \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top) \\ &= \mathrm{tr}(\boldsymbol{H}_\star \sum_{i=1}^n \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top) = \sum_{i=1}^n \mathrm{tr}(\boldsymbol{H}_\star \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top) = \sum_{i=1}^n \mathrm{tr}(\boldsymbol{\gamma}_i^\top \boldsymbol{H}_\star \boldsymbol{\gamma}_i) \\ &= \sum_{i=1}^n \|\boldsymbol{\gamma}_i\|^2_{\boldsymbol{H}_\star}. \end{aligned} \tag{32}$$

Taking $\|\boldsymbol{\gamma}_i\|_{\boldsymbol{H}_\star}$ to work on further, we can derive its upper bound as

$$\begin{aligned}\|\boldsymbol{\gamma}_i\|_{\boldsymbol{H}_\star} &= \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i) - h(\boldsymbol{x}) - \tau(\nabla h(\boldsymbol{x} + \tau\xi_i \boldsymbol{u}_i)^\top - \nabla h(\boldsymbol{x})^\top)\boldsymbol{u}_i\|_{\boldsymbol{H}_\star} \\ &\le \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} + \tau\|(\nabla h(\boldsymbol{x} + \tau\xi_i \boldsymbol{u}_i)^\top - \nabla h(\boldsymbol{x})^\top)\boldsymbol{u}_i\|_{\boldsymbol{H}_\star} \\ &\le \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} + \tau\|\nabla h(\boldsymbol{x} + \tau\xi_i \boldsymbol{u}_i)^\top - \nabla h(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star} \\ &\le \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} + \tau^2\beta\xi_i \\ &\le \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} + \tau^2\beta, \end{aligned} \tag{33}$$

where the first inequality is by triangle inequality, the second inequality is by Lemma E.1 and the fact that $\|\boldsymbol{u}_i\|_2 = 1$, the third inequality is done by applying the β-smoothness assumption, and the last inequality is by the fact that $\xi_i \in (0, 1)$ from the mean value theorem.

Plugging the equation 33 into equation 32, we have

$$\begin{aligned}\tau^2 \cdot \|\nabla h(\boldsymbol{x})^\top\|^2_{\boldsymbol{H}_\star} &\le \sum_{i=1}^n \left(\|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|_{\boldsymbol{H}_\star} + \tau^2\beta\right)^2 \\ &\le \sum_{i=1}^n 3\left(\|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|^2_{\boldsymbol{H}_\star} + \|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} + \tau^4\beta^2\right) \\ &= 3\sum_{i=1}^n \|h(\boldsymbol{x} + \tau \boldsymbol{u}_i)\|^2_{\boldsymbol{H}_\star} + 3n\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} + 3n\tau^4\beta^2, \end{aligned}$$

where the inequality is done Cauchy-Schwarz inequality.

Denoting $\boldsymbol{x}'_i = \boldsymbol{x} + \tau \boldsymbol{u}_i$, we have the lemma proved.

□

Theorem D.2 (Theorem 3.2 Restated)**.** *Given a data distribution $\mathcal{D}$ and $\tau > 0$, there exist distributions $\mathcal{D}_1, \mathcal{D}_2$ such that the type-1 Wasserstein distance $W_1(\mathcal{D}, \mathcal{D}_1) \le \tau$ and $W_1(\mathcal{D}, \mathcal{D}_2) \le \tau$ satisfying*

$$\begin{aligned}\tfrac{1}{2B^2}\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} &\le \|\nabla h'^\top_{\star,\diamond}\|^2_{\mathcal{D}_1,\boldsymbol{H}_\star} + \beta^2(B - \tau)^2_+ \\ \tfrac{1}{3n}\|\nabla h'^\top_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} &\le \tfrac{2}{\tau^2}\|h_{\star,\diamond}\|^2_{\mathcal{D}_2,\boldsymbol{H}_\star} + \beta^2\tau^2,\end{aligned}$$

where n is the dimension of $\boldsymbol{x} \sim \mathcal{D}$, and $B = \inf_{\boldsymbol{x}_0 \in \mathbb{R}^n} \sup_{\boldsymbol{x} \in supp(\mathcal{D})} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ is the radius of the supp$(\mathcal{D})$. The $(\cdot)_+$ is an operator defined by $\forall x \in \mathbb{R}$: $(x)_+ = x$ if $x \ge 0$ and $(x)_+ = 0$ otherwise.

Proof. Let us begin with recalling the definition of $h_{\star,\diamond}$ and $h'_{\star,\diamond}$.

The optimal affine transformation $g \in \mathbb{G}$ in the function matching distance (13) is $\tilde{g}$, and one of the optimal $g \in \mathbb{G}$ in the gradient matching distance is (14) $\tilde{g}'$. Accordingly, we denote

$$h_{\star,\diamond} := f_\star - \tilde{g} \circ f_\diamond \quad \text{and} \quad h'_{\star,\diamond} := f_\star - \tilde{g}' \circ f_\diamond,$$

and we can see that the gradient matching distance and the function matching distance can be written as

$$(13) = \|h_{\star,\diamond}\|_{\mathcal{D},\boldsymbol{H}_\star} \quad \text{and} \quad (14) = \|\nabla {h'_{\star,\diamond}}^\top\|_{\mathcal{D},\boldsymbol{H}_\star}. \tag{34}$$

The first inequality. Then, we can prove the first inequality using Lemma D.3.

Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ be a free variable, and then set $\boldsymbol{b} = h'_{\star,\diamond}(\boldsymbol{x}_0)$. Noting that $\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star}$ by definition is the minimum of this function distance, we have

$$\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} \le \|h'_{\star,\diamond} - \boldsymbol{b}\|^2_{\mathcal{D},\boldsymbol{H}_\star}. \tag{35}$$

Denoting $h := h'_{\star,\diamond} - \boldsymbol{b}$, we can see $h(\boldsymbol{x}_0) = 0$. Therefore, h can be used to invoke Lemma D.3. That is, there exists $\boldsymbol{x}'$ as a function of $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \le \tau$, and

$$\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} \le 2\left(\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2 \left(\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 - \tau\right)^2_+\right) \cdot \|\boldsymbol{x} - \boldsymbol{x}_0\|^2_2,$$

Taking the expectation of $\boldsymbol{x} \sim \mathcal{D}$ of the both sides, and denote the induced distribution for $\boldsymbol{x}'$ as $\mathcal{D}_1$, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} \le 2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left(\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2 \left(\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 - \tau\right)^2_+\right) \cdot \|\boldsymbol{x} - \boldsymbol{x}_0\|^2_2.$$

Recall that $\boldsymbol{x}_0$ is a free variable, we can tighten the bound by

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} \le \inf_{\boldsymbol{x}_0\in\mathbb{R}^n} 2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left(\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2 \left(\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 - \tau\right)^2_+\right) \cdot \|\boldsymbol{x} - \boldsymbol{x}_0\|^2_2. \tag{36}$$

Note that we can have tighter but similar results if we keep the $\inf \boldsymbol{x}_0 \in \mathbb{R}^n$. However, by plugging in the radius

$$B = \inf_{\boldsymbol{x}_0\in\mathbb{R}^n} \sup_{\boldsymbol{x}\in\text{supp}(\mathcal{D})} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$$

we can make the presentation much more simplified without losing its core messages.

That is,

$$(36) \le 2\left(\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}_1}\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2 \left(B - \tau\right)^2_+\right) B^2.$$

Combining the above inequality and equation 35, and noting that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} = \|h\|^2_{\mathcal{D},\boldsymbol{H}_\star}$$
$$\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}_1}\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} = \|\nabla h^\top\|^2_{\mathcal{D}_1,\boldsymbol{H}_\star},$$

we have

$$\begin{aligned}
\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} &\le \|h'_{\star,\diamond} - \boldsymbol{b}\|^2_{\mathcal{D},\boldsymbol{H}_\star} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|h(\boldsymbol{x})\|^2_{\boldsymbol{H}_\star} \\
&\le 2\left(\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}_1}\|\nabla h(\boldsymbol{x}')^\top\|^2_{\boldsymbol{H}_\star} + \beta^2 \left(B - \tau\right)^2_+\right) B^2 \\
&= 2\left(\|\nabla h^\top\|^2_{\mathcal{D}_1,\boldsymbol{H}_\star} + \beta^2 \left(B - \tau\right)^2_+\right) B^2
\end{aligned}$$

Noting that h and $h'_{\star,\diamond}$ only differs by a constant shift $\boldsymbol{b}$, we can see $\nabla h = \nabla h'_{\star,\diamond}$. Therefore, by replacing $\nabla h^\top$ by $\nabla h'^\top_{\star,\diamond}$ we finally have the first inequality in Theorem 3.2

$$\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} \le 2\left(\|\nabla h'^\top_{\star,\diamond}\|^2_{\mathcal{D}_1,\boldsymbol{H}_\star} + \beta^2 \left(B - \tau\right)^2_+\right) B^2.$$

It remains to show the Wasserstein distance between $\mathcal{D}_1$ and $\mathcal{D}$. As $\boldsymbol{x}'$ is a function of the random variable $\boldsymbol{x} \sim \mathcal{D}$ with $\|\boldsymbol{x}' - \boldsymbol{x}\|_2 \leq \tau$, and $\mathcal{D}_1$ is the induced distribution of $\boldsymbol{x}'$ as a function of $\boldsymbol{x}$, we can see that by the definition of type-1 Wasserstein distance between $\mathcal{D}$ and $\mathcal{D}_1$ is bounded by τ.

Denote $\mathbb{J}(\mathcal{D}, \mathcal{D}')$ as the set of all joint distributions that have marginals $\mathcal{D}$ and $\mathcal{D}'$, and recall the definition of type-1 Wasserstein distance is

$$W_1(\mathcal{D}, \mathcal{D}_1) = \inf_{\mathcal{J} \in \mathbb{J}(\mathcal{D}, \mathcal{D}_1)} \int \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \, \mathrm{d}\mathcal{J}(\boldsymbol{x}, \boldsymbol{x}').$$

Denote $\mathcal{J}_0$ as the joint distribution such that in $(\boldsymbol{x}, \boldsymbol{x}') \sim \mathcal{J}$ we always have $\boldsymbol{x}'$ being a function of $\boldsymbol{x}$ as how $\boldsymbol{x}'$ is defined. We can see that

$$\begin{aligned} W_1(\mathcal{D}, \mathcal{D}_1) &= \inf_{\mathcal{J} \in \mathbb{J}(\mathcal{D}, \mathcal{D}_1)} \int \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \, \mathrm{d}\mathcal{J}(\boldsymbol{x}, \boldsymbol{x}') \leq \int \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \, \mathrm{d}\mathcal{J}_0(\boldsymbol{x}, \boldsymbol{x}') \leq \int \tau \, \mathrm{d}\mathcal{J}_0(\boldsymbol{x}, \boldsymbol{x}') \\ &= \tau. \end{aligned} \tag{37}$$

Therefore, we have the first inequality in the theorem proved .

The second inequality. Invoking Lemma D.4 with $h_{\star,\diamond}$, and rearranging the inequality, we have

$$\tfrac{1}{3n}\|\nabla h_{\star,\diamond}(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 \leq \tfrac{2}{\tau^2}\left(\sum_{i=1}^n \tfrac{1}{2n}\|h_{\star,\diamond}(\boldsymbol{x}_i')\|_{\boldsymbol{H}_\star}^2 + \tfrac{1}{2}\|h_{\star,\diamond}(\boldsymbol{x})\|_{\boldsymbol{H}_\star}^2\right) + \tau^2\beta^2.$$

Taking the expectation on both sides, we have

$$\tfrac{1}{3n}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|\nabla h_{\star,\diamond}(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2 \leq \tfrac{2}{\tau^2}\underbrace{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left(\sum_{i=1}^n \tfrac{1}{2n}\|h_{\star,\diamond}(\boldsymbol{x}_i')\|_{\boldsymbol{H}_\star}^2 + \tfrac{1}{2}\|h_{\star,\diamond}(\boldsymbol{x})\|_{\boldsymbol{H}_\star}^2\right)}_{X} + \tau^2\beta^2. \tag{38}$$

Note that X can be reformulated to be the expectation of an induced distribution from $x \sim \mathcal{D}$, since $\boldsymbol{x}_i'$ is a pre-defined function of $\boldsymbol{x}$. Denote $\mathcal{D}_2$ as the distribution induced by the following sampling process: first, sample $\boldsymbol{x} \sim \mathcal{D}$; then,

$$\begin{aligned} &\boldsymbol{x}' = \boldsymbol{x} \quad \text{with probability } \tfrac{1}{2} \\ &\boldsymbol{x}' = \boldsymbol{x}_i' \quad \text{with probability } \tfrac{1}{2n} \text{ for } \forall i \in [n]. \end{aligned}$$

Therefore, we can write X as

$$X = \|h_{\star,\diamond}\|_{\mathcal{D}_2,\boldsymbol{H}_\star}^2. \tag{39}$$

Similarly to equation 37, it also holds that $W_1(\mathcal{D}, \mathcal{D}_2) \leq \tau$.

To finally complete the proof, noting that $\|\nabla h_{\star,\diamond}'^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2$ is the minimum of this gradient distance (equation 34), we have

$$\|\nabla h_{\star,\diamond}'^\top\|_{\mathcal{D},\boldsymbol{H}_\star}^2 \leq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|\nabla h_{\star,\diamond}(\boldsymbol{x})^\top\|_{\boldsymbol{H}_\star}^2. \tag{40}$$

Combining equation 38, equation 39 and equation 40, we have the second inequality proved.

Hence, we have proved Theorem 3.2. □

D.4. Proof of Theorem 3.3

Theorem D.3 (Theorem 3.3 Restated)**.** *The surrogate transfer loss (16) and the true transfer loss (17) are close, with an error of* $\|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}$.

$$-\|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T} \leq (17) - (16) \leq \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}$$

Proof. Let us begin by recall the definition of the surrogate transfer loss (16) and the true transfer loss (17).

$$(16) := \min_{g\in\mathbb{G}} \|f_T - g\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}$$
$$(17) := \min_{g\in\mathbb{G}} \|y - g\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}.$$

Denote

$$\tilde{g}' := \arg\min_{g\in\mathbb{G}} \|f_T - g\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}$$
$$\tilde{g} := \arg\min_{g\in\mathbb{G}} \|y - g\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}.$$

First, we show an upper bound for (16).

$$(16) \le \|f_T - \tilde{g}\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} \le \|y - \tilde{g}\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} + \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T} = (17) + \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}, \tag{41}$$

where the last inequality is by triangle inequality.

Similarly, we can derive its lower bound.

$$\begin{aligned}(16) = \|f_T - \tilde{g}'\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} &\ge \|y - \tilde{g}'\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} - \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}\\ &\ge \min_{g\in\mathbb{G}} \|y - g\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} - \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T} = (17) - \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T},\end{aligned} \tag{42}$$

where the first inequality is by triangle inequality.

Combining equation 41 and equation 42, we have the proposition proved.

□

D.5. Proof of Theorem B.1

Theorem D.4 (Theorem B.1 Restated)**.** *Denote $\tilde{g}_{T,S} : \mathbb{R}^m \to \mathbb{R}^d$ as the optimal solution of equation 16, and $\tilde{g}_{S,T} : \mathbb{R}^d \to \mathbb{R}^m$ as the optimal solution of equation 18. Suppose the two optimal affine maps $\tilde{g}_{T,S}, \tilde{g}_{S,T}$ are both full-rank. For $\boldsymbol{v}\in\mathbb{R}^m$, denote the matrix representation of $\tilde{g}_{T,S}$ as $\tilde{g}_{T,S}(\boldsymbol{v}) = \tilde{\boldsymbol{W}}_{T,S}\boldsymbol{v} + \tilde{\boldsymbol{b}}_{T,S}$. Similarly, for $\boldsymbol{w}\in\mathbb{R}^d$, denote the matrix representation of $\tilde{g}_{S,T}$ as $\tilde{g}_{S,T}(\boldsymbol{w}) = \tilde{\boldsymbol{W}}_{S,T}\boldsymbol{w} + \tilde{\boldsymbol{b}}_{S,T}$. We have the following statements.*

If $d < m$, then $\tilde{g}_{S,T}$ is injective, and we have:

$$\|f_T - \tilde{g}_{T,S}\circ f_S\|_{\mathcal{D},\boldsymbol{H}_T} \le \sqrt{\|(\tilde{\boldsymbol{W}}_{S,T}^\top\tilde{\boldsymbol{W}}_{S,T})^{-1}\|_F \cdot \|\boldsymbol{H}_T\|_F}\cdot\|f_S - \tilde{g}_{S,T}\circ f_T\|_{\mathcal{D}}. \tag{19}$$

If $d > m$, then $\tilde{g}_{T,S}$ is injective, and we have:

$$\|f_S - \tilde{g}_{S,T}\circ f_T\|_{\mathcal{D},\boldsymbol{H}_S} \le \sqrt{\|(\tilde{\boldsymbol{W}}_{T,S}^\top\tilde{\boldsymbol{W}}_{T,S})^{-1}\|_F \cdot \|\boldsymbol{H}_S\|_F}\cdot\|f_T - \tilde{g}_{T,S}\circ f_S\|_{\mathcal{D}}. \tag{20}$$

If $d = m$, then both $\tilde{g}_{S,T}$ and $\tilde{g}_{T,S}$ are bijective, and we have both (19) and (20) stand.

Proof. Observing the symmetry, we only need to prove the following claim.

Claim. *For $\star,\diamond\in\{S,T\}$ and $\star\neq\diamond$, if $\tilde{g}_{\star,\diamond}$ is injective, then*

$$\|f_\diamond - \tilde{g}_{\diamond,\star}\circ f_\star\|_{\mathcal{D},\boldsymbol{H}_\diamond}^2 \le \|(\tilde{\boldsymbol{W}}_{\star,\diamond}^\top\tilde{\boldsymbol{W}}_{\star,\diamond})^{-1}\|_F\cdot\|\boldsymbol{H}_\diamond\|_F\cdot\|f_\star - \tilde{g}_{\star,\diamond}\circ f_\diamond\|_{\mathcal{D}}^2.$$

Proof of the Claim. We have mostly done with this claim with Lemma E.3. Noting that $\tilde{g}_{\diamond,\star}$ is the minimizer of $\min_{g\in\mathbb{G}}\|f_\diamond - g\circ f_\star\|_{\mathcal{D},\boldsymbol{H}_\diamond}^2$, we have

$$\begin{aligned}\|f_\diamond - \tilde{g}_{\diamond,\star}\circ f_\star\|_{\mathcal{D},\boldsymbol{H}_\diamond}^2 &\le \|f_\diamond - \tilde{g}_{\star,\diamond}^{-1}\circ f_\star\|_{\mathcal{D},\boldsymbol{H}_\diamond}^2 = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|f_\diamond(\boldsymbol{x}) - \tilde{g}_{\star,\diamond}^{-1}(f_\star(\boldsymbol{x}))\|_{\boldsymbol{H}_\diamond}^2\right]\\ &\le \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|(\tilde{\boldsymbol{W}}_{\star,\diamond}^\top\tilde{\boldsymbol{W}}_{\star,\diamond})^{-1}\|_F\cdot\|\boldsymbol{H}_\diamond\|_F\cdot\|f_\star(\boldsymbol{x}) - \tilde{g}_{\star,\diamond}(f_\diamond(\boldsymbol{x}))\|_2^2\right]\\ &= \|(\tilde{\boldsymbol{W}}_{\star,\diamond}^\top\tilde{\boldsymbol{W}}_{\star,\diamond})^{-1}\|_F\cdot\|\boldsymbol{H}_\diamond\|_F\cdot\|f_\star - \tilde{g}_{\star,\diamond}\circ f_\diamond\|_{\mathcal{D}}^2,\end{aligned}$$

where the second inequality is by invoking Lemma E.3. □

Taking the square root of this claim, and applying ($\diamond = T, \star = S$) or ($\diamond = S, \star = T$), we immediatly have the first two statements about the case of $d < m$ or $d > m$. Finally, noting that when $m = d$, both $\tilde{g}_{S,T}$ and $\tilde{g}_{T,S}$ are bijective and thus also injective, we can see that both (19) and (20) stand. □

E. Auxiliary Lemmas

Lemma E.1 (Compatibility of $\|\cdot\|_{\boldsymbol{H}}$ and $\|\cdot\|_2$)**.** *Let $\boldsymbol{H} \in \mathbb{R}^{m\times m}$ be a positive semi-definite matrix, and denote $\boldsymbol{H} = \boldsymbol{T}^\top \boldsymbol{T}$ as its symmetric decomposition with $\boldsymbol{T} \in \mathbb{R}^{m\times m}$. For $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{v} \in \mathbb{R}^n$, we have*

$$\|\boldsymbol{W}\boldsymbol{v}\|_{\boldsymbol{H}} \le \|\boldsymbol{W}\|_{\boldsymbol{H}} \cdot \|\boldsymbol{v}\|_2.$$

Proof.

$$\begin{aligned}\|\boldsymbol{W}\boldsymbol{v}\|_{\boldsymbol{H}}^2 &= \boldsymbol{v}^\top \boldsymbol{W}^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{W}\boldsymbol{v} = \|\boldsymbol{T}\boldsymbol{W}\boldsymbol{v}\|_2^2 \\ &\le \|\boldsymbol{T}\boldsymbol{W}\|_2^2 \cdot \|\boldsymbol{v}\|_2^2 \le \|\boldsymbol{T}\boldsymbol{W}\|_F^2 \cdot \|\boldsymbol{v}\|_2^2,\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. Then, we can continue as

$$\|\boldsymbol{T}\boldsymbol{W}\|_F^2 = \mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{W}) = \mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{H}\boldsymbol{W}) = \|\boldsymbol{W}\|_{\boldsymbol{H}}^2.$$

Combining the above two parts, we have the lemma proved. □

Lemma E.2 (Expectation Preserves the Inclusion Relationship Between Linear Spaces)**.** *Given a distribution $\boldsymbol{x} \sim \mathcal{D}$ in $\mathbb{R}^n$, we denote the associated probability measure as μ. Given linear maps $\boldsymbol{M}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{N}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^d$, noting that they are both functions of $\boldsymbol{x}$, we have the following statement.*

$$\ker\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{M}_x^\top \boldsymbol{M}_x\right) \subseteq \ker\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{N}_x^\top \boldsymbol{M}_x\right),$$

where $\ker(\cdot)$ *denotes the kernel space of a given liner map.*

Proof. It suffice to show for $\forall \boldsymbol{v} \in \ker\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{M}_x^\top \boldsymbol{M}_x\right)$, we also have $\boldsymbol{v} \in \ker\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{N}_x^\top \boldsymbol{M}_x\right)$.

Denote $\boldsymbol{P} := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{M}_x^\top \boldsymbol{M}_x$, and let $\boldsymbol{v} \in \ker(\boldsymbol{P})$, we have

$$\boldsymbol{P}\boldsymbol{v} = \boldsymbol{0}.$$

Noting that $\boldsymbol{P}$ is positive semi-definite, we have the following equivalent statements.

$$\boldsymbol{v} \in \ker(\boldsymbol{P}) \quad \Longleftrightarrow \quad \boldsymbol{v}^\top \boldsymbol{P}\boldsymbol{v} = 0,$$

where the ' $\Longrightarrow$ ' direction is trivial, and the ' $\Longleftarrow$ ' direction can be proved by decomposing $\boldsymbol{P} = \boldsymbol{T}^\top \boldsymbol{T}$ as two matrices and noting that

$$\boldsymbol{v}^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{v} = 0 \quad \Longrightarrow \quad \|\boldsymbol{T}\boldsymbol{v}\|_2^2 = 0 \quad \Longrightarrow \quad \boldsymbol{T}\boldsymbol{v} = \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{v} = \boldsymbol{0} \quad \Longrightarrow \boldsymbol{P}\boldsymbol{v} = \boldsymbol{0}.$$

Therefore, we have

$$\begin{aligned} & \boldsymbol{v}^\top \boldsymbol{P}\boldsymbol{v} = 0 \\ \Longrightarrow \quad & \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{v}^\top \boldsymbol{M}_x^\top \boldsymbol{M}_x \boldsymbol{v}\right] = 0 \\ \Longrightarrow \quad & \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\boldsymbol{M}_x\boldsymbol{v}\|_2^2\right] = 0 \\ \Longrightarrow \quad & \int \|\boldsymbol{M}_x\boldsymbol{v}\|_2^2 \,\mathrm{d}\mu = 0, \end{aligned}$$

which implies $\boldsymbol{M}_x\boldsymbol{v} = \boldsymbol{0}$ almost everywhere w.r.t. μ.

Therefore, applying $\boldsymbol{v}$ to $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{N}_x^\top \boldsymbol{M}_x]$ and we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{N}_x^\top \boldsymbol{M}_x\right]\boldsymbol{v} &= \int \boldsymbol{N}_x^\top \boldsymbol{M}_x \boldsymbol{v}\,\mathrm{d}\mu \\ &= \int_{a.e.} \boldsymbol{N}_x^\top \mathbf{0}\,\mathrm{d}\mu \\ &= \mathbf{0},\end{aligned}$$

which means $\boldsymbol{v} \in \ker\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{N}_x^\top \boldsymbol{M}_x\right)$.

□

Lemma E.3 (Inverse an Injective Linear Map)**.** *Given a full-rank injective affine transformation $g : \mathbb{R}^m \to \mathbb{R}^d$, we denote its matrix representation as $g(\boldsymbol{v}) = \boldsymbol{W}\boldsymbol{v} + \boldsymbol{b}$ where $\boldsymbol{v} \in \mathbb{R}^m, \boldsymbol{W} \in \mathbb{R}^{d\times m}, \boldsymbol{b} \in \mathbb{R}^d$. The inverse of g is $g^{-1} : \mathbb{R}^d \to \mathbb{R}^m$ defined by $g^{-1}(\boldsymbol{w}) := \boldsymbol{W}^\dagger\boldsymbol{w} - \boldsymbol{W}^\dagger\boldsymbol{b}$ for $\boldsymbol{w} \in \mathbb{R}^d$,* i.e.*, $g^{-1} \circ g$ is the identity function. Moreover, given a positive semi-definite matrix $\boldsymbol{H}$, for $\forall \boldsymbol{v} \in \mathbb{R}^m$ and $\forall \boldsymbol{w} \in \mathbb{R}^d$, we have*

$$\sqrt{\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\|_F \cdot \|\boldsymbol{H}\|_F} \cdot \|\boldsymbol{w} - g(\boldsymbol{v})\|_2 \geq \|\boldsymbol{v} - g^{-1}(\boldsymbol{w})\|_{\boldsymbol{H}}.$$

Proof. First, let us verify that $g^{-1} \circ g$ is the identity function. The conditions of g being full-rank and injective are equivalent to $\boldsymbol{W}$ being full-rank and $d \geq m$. That is being said, $\boldsymbol{W}^\top\boldsymbol{W}$ is invertible and $\boldsymbol{W}^\dagger = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$. Therefore, for $\forall \boldsymbol{v} \in \mathbb{R}^m$, we have

$$\begin{aligned}g^{-1} \circ g(\boldsymbol{v}) &= \boldsymbol{W}^\dagger(\boldsymbol{W}\boldsymbol{v} + \boldsymbol{b}) - \boldsymbol{W}^\dagger\boldsymbol{b} = \boldsymbol{W}^\dagger\boldsymbol{W}\boldsymbol{v} \\ &= (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{v} = \boldsymbol{v}.\end{aligned}$$

That is, $g^{-1} \circ g$ is indeed the identity function.

Next, to prove the inequality, let us start from the right-hand-side of the inequality.

$$\begin{aligned}\|\boldsymbol{v} - g^{-1}(\boldsymbol{w})\|_{\boldsymbol{H}} &= \|g^{-1} \circ g(\boldsymbol{v}) - g^{-1}(\boldsymbol{w})\|_{\boldsymbol{H}} \\ &= \|\boldsymbol{W}^\dagger(g(\boldsymbol{v}) - \boldsymbol{w})\|_{\boldsymbol{H}} \\ &\leq \|\boldsymbol{W}^\dagger\|_{\boldsymbol{H}} \cdot \|g(\boldsymbol{v}) - \boldsymbol{w}\|_2, \end{aligned} \tag{43}$$

where the inequality is done by applying Lemma E.1.

To complete the prove, we can see that

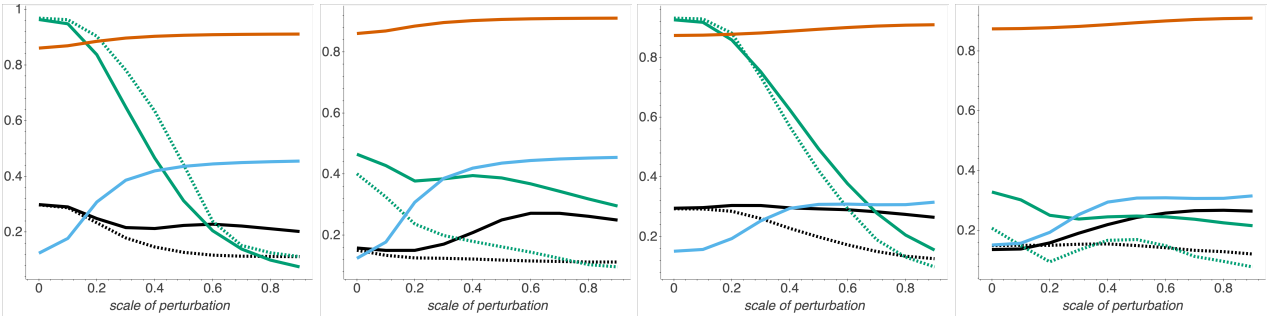
$$\begin{aligned}\|\boldsymbol{W}^\dagger\|_{\boldsymbol{H}}^2 &= \|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\|_{\boldsymbol{H}}^2 = \mathrm{tr}(\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{H}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top) \\ &= \mathrm{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{H}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{W}) = \mathrm{tr}((\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{H}) \\ &= \langle(\boldsymbol{W}^\top\boldsymbol{W})^{-1}, \boldsymbol{H}\rangle \\ &\leq \|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\|_F \cdot \|\boldsymbol{H}\|_F. \end{aligned} \tag{44}$$

Plugging the square root of equation 44 into equation 43, we have the lemma proved. □

F. Additional Details of Synthetic Experiments

In this section, we complete the description of the settings and methods used in the synthetic experiments. Moreover, we report two additional sets of results in cross-architecture scenarios.

In the main paper (section 4), the synthetic experiments are done on the setting where source models have the same architecture as the target model, *i.e.*, all the models are one-hidden-layer neural networks with width $m = 100$. A natural question is what would the results be if using different architectures? That is, the architecture of the source models are different from the target model. To answer this question, we present two additional sets of synthetic experiments where the width of the source models is $m = 50$ or $m = 200$, different from the target model (width $m = 100$).



(a) width=50, $\boldsymbol{\delta}_{f_\star}^{(1)}$ (b) width=50, $\boldsymbol{\delta}_{f_\star}^{(2)}$ (c) width=200, $\boldsymbol{\delta}_{f_\star}^{(1)}$ (d) width=200, $\boldsymbol{\delta}_{f_\star}^{(2)}$

Figure 3. In this figure, 'width' is the width of the source models (one-hidden-layer neural networks). As defined in equation 9, $\boldsymbol{\delta}_{f_\star}^{(1)}$ corresponds to the regular adversarial attacks, while $\boldsymbol{\delta}_{f_\star}^{(2)}$ the secondary adversarial attack. That is, $\boldsymbol{\delta}_{f_\star}^{(2)}$ represents the other information in the adversarial transferring process compared with the first. The x-axis shows the scale of perturbation $t \in [0, 1]$ that controls how much the source model deviates from its corresponding reference source model. There are in total 6 quantities reported. Specifically, $\alpha_1^{f_T \to f_S}$ is **black solid**; $\alpha_1^{f_S \to f_T}$ is **black dotted**; $\alpha_2^{f_T \to f_S}$ is **green solid**; $\alpha_2^{f_S \to f_T}$ is **green dotted**; the gradient matching loss is **red solid**; and the knowledge transferability distance is **blue solid**.

As we have presented in the main paper about the description of the methods and models used in this experiment, here we present the detailed description of the settings and the datasets being used.

Settings. We follow the small-ϵ setting used in the theory, *i.e.*, the adversarial attack are constrained to a small magnitude, so that we can use its first-order Talyor approximation.

Dataset. Denote a radial basis function as $\phi_i(\boldsymbol{x}) = e^{-\|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2/(\sigma_i)^2}$, and for each input data we form its corresponding M-dimensional feature vector as $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x})]^\top$. We set the dimension of $\boldsymbol{x}$ to be 50. For each radial basis function $\phi_i(\boldsymbol{x}), i \in [M]$, $\boldsymbol{\mu}_i$ is sampled from $U(-0.5, 0.5)^{50}$, and σ_i^2 is sampled from $U(0, 100)$. We use $M = 100$ radial basis functions so that the feature vector is 100-dimensional. Then, we set the target ground truth to be $y(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{x}) + \boldsymbol{b}$ where $\boldsymbol{W} \in \mathbb{R}^{10\times 100}, \boldsymbol{b} \in \mathbb{R}^{10}$ are sampled from $U(-0.5, 0.5)$ element-wise. We generate $N = 5000$ samples of $\boldsymbol{x}$ from a Gaussian mixture formed by 10 Gaussians with different centers but the same covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}$. The centers are sampled randomly from $U(-0.5, 0.5)^{50}$. That is, the dataset $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ consists of $N = 5000$ sample from the distribution, where $\boldsymbol{x}_i$ is 50-dimensional, $\boldsymbol{y}_i$ is 10-dimensional. The ground truth target $\boldsymbol{y}_i$ are computed using the ground truth target function $y(\boldsymbol{x}_i)$. That is, we want our neural networks to approximate $y(\cdot)$ on the Gaussian mixture.

Methods of Additional Experiments. Note that we have provided the detailed description of the methods used in the main paper synthetic experiments. Here, we present the methods for two additional sets of synthetic experiments, using the same dataset and settings, but different architectures. In the main paper, the source model and the target model are of the same architecture, and the source models are perturbed target model. Here, we use the same target model f_T (width $m = 100$) trained on the dataset D, but two different architectures for source models. That is, the source models and the target model are of different width.

To derive the source models, we first train two reference source models on D with width $m = 50$ and $m = 200$. For each of the reference models, denoting the weights of the model as $\boldsymbol{W}$, we randomly sample a direction $\boldsymbol{V}$ where each entry of $\boldsymbol{V}$ is sampled from $U(-0.5, 0.5)$, and choose a scale $t \in [0, 1]$. Subsequently, we perturb the model weights of the clean source model as $\boldsymbol{W}' := \boldsymbol{W} + t\boldsymbol{V}$, and define the source model f_S to be a one-hidden-layer neural network with weights $\boldsymbol{W}'$. Then, we compute each of the quantities we care about, including α_1, α_2 from both $f_S \to f_T$ and $f_T \to f_S$, the gradient matching distance (equation 7), and the actual knowledge transfer distance (equation 17). We use the standard ℓ_2 loss as the adversarial loss function.

Results. We present four sets of result in Figure 3. The indication relations between adversarial transferability and knowledge transferability can be observed in the cross-architecture setting. Moreover: 1. the metrics α_1, α_2 are more meaningful if using the regular attacks; 2. the gradient matching distance tracks the actual knowledge transferability loss; 3. the directions of $f_T \to f_S$ and $f_S \to f_T$ are similar.

G. Details of the Empirical Experiments

All experiments are run on a single GTX2080Ti.

G.1. Datasets

G.1.1. Image Datasets

- **CIFAR10**:[1]: it consists of 60000 32×32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.
- **STL10**:[2]: it consists of 13000 labeled 96×96 colour images in 10 classes, with 1300 images per class. There are 5000 training images and 8000 test images. 500 training images (10 pre-defined folds), 800 test images per class.

G.1.2. NLP Datasets

- **IMDB**:[3] Document-level sentiment classification on positive and negative movie reviews. We use this dataset to train the target model.
- **AG's News (AG)**: Sentence-level classification with regard to four news topics: World, Sports, Business, and Science/Technology. Following Zhang et al. (2015), we concatenate the title and description fields for each news article. We use this dataset to train the source model.
- **Fake News Detection (Fake)**: Document-level classification on whether a news article is fake or not. The dataset comes from the Kaggle Fake News Challenge[4]. We concatenate the title and news body of each article. We use this dataset to train the source model.
- **Yelp**: Document-level sentiment classification on positive and negative reviews (Zhang et al., 2015). Reviews with a rating of 1 and 2 are labeled negative and 4 and 5 positive. We use this dataset to train the source model.

G.2. Adversarial Trasnferability Indicating Knowledge Transferability

G.2.1. Image

For all the models, both source and target, in the Cifar10 to STL10 experiment, we train them by SGD with momentumn and learning rate 0.1 for 100 epochs. For knowledge tranferability, we randomly reinitialize and train the source models' last layer for 10 epochs on STL10. Then we generate adversarial examples with the target model on the validation set and measure the adversarial transferability by feeding these adversarial examples to the source models. We employ two adversarial attacks in this experiments and show that they achieve the same propose in practice: First, we generate adversarial examples by 50 steps of projected gradient descent and epsilon 0.1 (Results shown in Table 1). Then, we generate adversarial examples by the more efficient FGSM with epsilon 0.1 (Results shown in Table 6) and show that we can efficiently identify candidate models without the expensive PGD attacks.

To further visualize the averaged relation presented in Table 1 and 6, we plot scatter plots Figure 5 and Figure 4 with per sample α_1 as x axis and per sample transfer loss as y axis. Transfer loss is the cross entropy loss predicted by the source model with last layer fine-tuned on STL10. The Pearson score indicates strong correlation between adversarial transferability and knowledge transferability.

We note that in the figures where we report per-sample α_1, although ideally $\alpha_1 \in [0, 1]$, we can observe that for some samples they have $\alpha_1 > 1$ due to the attacking algorithm is not ideal in practice. However, the introduced sample-level noise does not affect the overall results, *e.g.*, see the averaged results in our tables, or the overall correlation in these figures.

G.2.2. NLP

In the NLP experiments, to train source and target models, we finetune BERT-base models on different datasets for 3 epochs with learning rate equal to $5e-5$ and warm-up steps equal to the 10% of the total training steps. For knowledge

[1] https://www.cs.toronto.edu/~kriz/cifar.html
[2] https://cs.stanford.edu/~acoates/stl10/
[3] https://datasets.imdbws.com/
[4] https://www.kaggle.com/c/fake-news/data

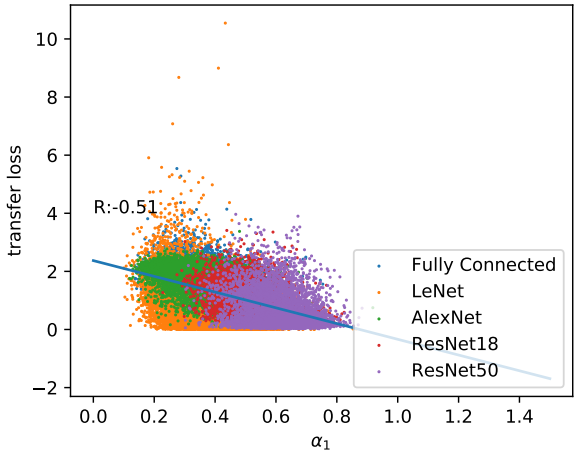


Figure 4. Distribution of per sample knowledge transfer loss and α_1. The adversarial samples are generated by PGD. The Pearson score shows strong negative correlation between α_1 and the knowledge transfer loss. The higher the transfer loss is, the lower the knowledge transferability is, and the lower the α_1 is.

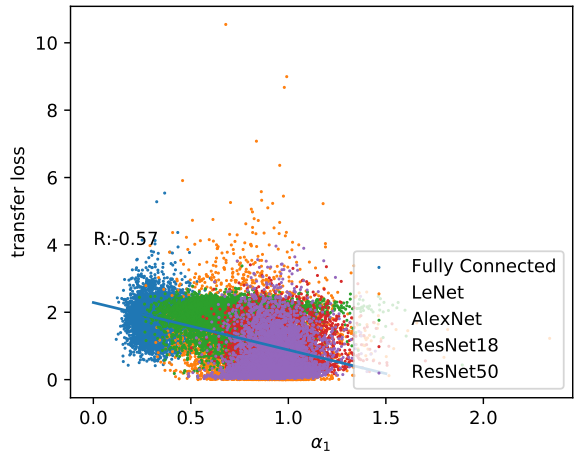


Figure 5. Distribution of per-sample knowledge transfer loss and α_1. The adversarial samples are generated by FGSM. The Pearson score shows negative strong correlation between α_1 and transfer loss. The higher the transfer loss is, the lower the knowledge transferability is, the lower the α_1 should be.

tranferability, we random initialize the last layer of source models and fine-tune all layers of BERT for 1 epoch on the targeted dataset (IMDB). Based on the test data from the target model, we generate $1,000$ textual adversarial examples via the state-of-the-art adversarial attacks T3 (Wang et al., 2020) with adversarial learning rate equal to 0.2, maximum iteration steps equal to 100, and $c = \kappa = 100$.

G.2.3. Ablation studies on controlling adversarial transferability

We conduct series of experiments on controlling adversarial transferability between source models and target model by promoting their Loss Gradient Diversity. Demontis et al. (2019) shows that for two models f_S and f_T, the cosine similarity between their loss gradient vectors $\nabla_x \ell_{f_S}$ and $\nabla_x \ell_{f_T}$ could be a significant indicator measuring two models' adversarial transferability. Moreover, Kariyappa & Qureshi (2019) claims that adversarial transferability betwen two models could be well controlled by regularizing the cosine similairity between their loss gradient vectors. Inspired by this, we train several

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.279	0.117	0.0103
AlexNet	45.65	0.614	0.208	0.0863
LeNet	55.09	0.803	0.298	0.205
ResNet18	76.60	1.000	0.405	0.410
ResNet50	77.92	0.962	0.392	0.368

Table 6. Knowledge transferability (Knowledge Trans.) among different model architectures. Adversarial examples are generated using FGSM attacks. Our correlation analysis shows Pearson score of -0.57 between the transfer loss and α_1. Lower transfer loss corresponds to higher transfer accuracy. More details can be found in Figure 5

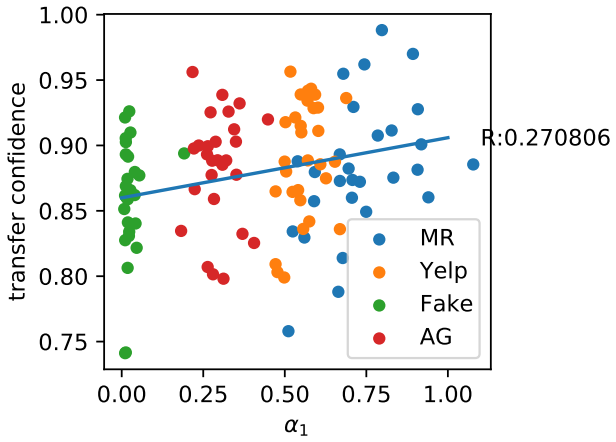


Figure 6. Distribution of per-batch knowledge transfer confidence and α_1. The Pearson score shows positive correlation between α_1 and transfer confidence. The higher the confidence, the higher the knowledge transferability.

Table 7. Knowledge transferability (Knowledge Trans.) among different source models (controlling adversarial transferability by promoting Loss Gradient Diversity). Adversarial transferability is measured by using the adversarial examples generated against the Target Model to attack the Source Models and estimate α_1 and α_2.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
$\rho = 0.0$	73.91	0.394	0.239	0.103
$\rho = 0.5$	73.11	0.385	0.246	0.102
$\rho = 1.0$	72.47	0.371	0.244	0.100
$\rho = 2.0$	71.62	0.370	0.244	0.100
$\rho = 5.0$	72.16	0.378	0.240	0.098

source models f_S to one target model f_T with following training loss:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{CE}}(f_S(\boldsymbol{x}), y)) + \rho \cdot \mathcal{L}_{\cos}(\nabla_{\hat{\boldsymbol{x}}} \ell_{f_S}, \nabla_{\hat{\boldsymbol{x}}} \ell_{f_T})$$

where $\mathcal{L}_{\text{CE}}$ refers to cross-entropy loss and $\mathcal{L}_{\cos}(\cdot, \cdot)$ the cosine similarity metric. $\boldsymbol{x}$ presents *source domain* instances while $\hat{\boldsymbol{x}}$ presents *target domain* instances. We explore $\rho \in \{0.0, 0.5, 1.0, 2.0, 5.0\}$ and finetune each source model for 50 epochs with learning rate as 0.01. For knowledge transferability, we random initialize the last layer of each source model and finetune it on STL-10 for 10 epochs with learning rate as 0.01. During the adversarial example generation, we utilize standard ℓ_∞ PGD attack with perturbation scale $\epsilon = 0.1$ and 50 attack iterations with step size as $\epsilon/10$.

Table 7 shows the relationship between knowledge transferability and adversarial transferability of different source model trained by different ρ. With the increasing of ρ, the adversarial transferabiltiy between source model and target model decreases ($\alpha_1, \alpha_1 * \alpha_2$ become smaller), and the knowledge transferability also decreases. We also plot the α_1 with its corresponding transfer loss on each instance, as shown in Figure 7. The negative correlation between α_1 and transfer loss confirms our theoretical insights.

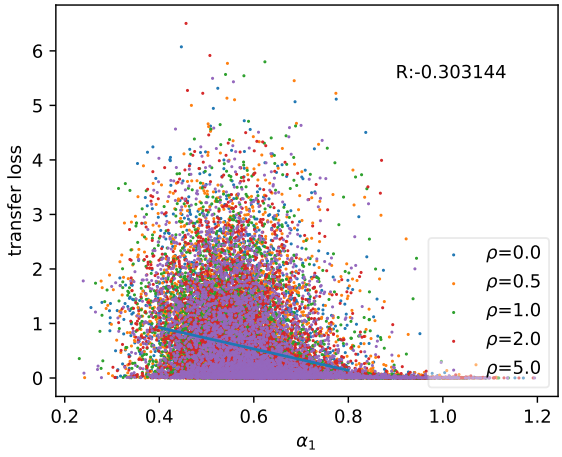


Figure 7. Distribution of per-sample knowledge transfer loss and α_1. The Pearson score shows negative correlation between α_1 and transfer loss. The higher the loss is, the lower the knowledge transferability is, the lower the α_1 should be.

G.3. Knowledge Trasnferability Indicating Adversarial Transferability

G.3.1. IMAGE

We follow the same setup in the previous image experiment for source model training, transfer learning as well as generation of adversarial examples. However, there is one key difference: Instead of generating adversarial examples on the target model and measuring adversarial transferability on source models, we generate adversarial examples on each source model and measure the adversarial transferability by feeding these adversarial examples to the target model.

Similarly, we also visualize the results (Table 3) and compute the Pearson score. Due to the significant noise introduced by per-sample calculation, the R score is not as significant as figure 5, but the trend is still correct and valid, which shows that higher knowledge transferability indicates higher adversarial transferability.

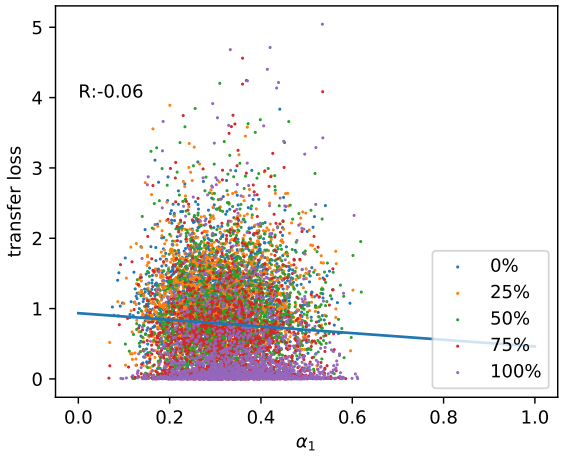


Figure 8. Distribution of per-sample knowledge transfer loss and α. The Pearson score shows negative strong correlation between α and transfer loss. The higher the loss is, the lower the knowledge transferability is, and the lower the α_1 is.

G.3.2. NLP

We follow the same setup to train the models and generate textual adversarial examples as §G.2 in the NLP experiments. We note that to measure the adversarial transferability, we generate $1,000$ adversarial examples on each source model based on the test data from the target model, and measure the adversarial transferability by feeding these adversarial examples to the target model.

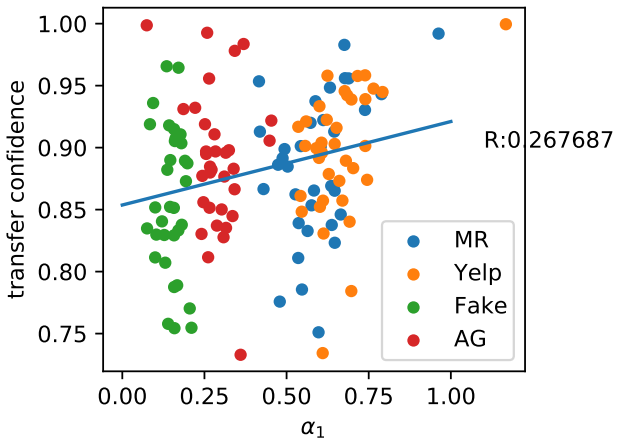


Figure 9. Distribution of per-batch knowledge transfer confidence and α_1. The Pearson score shows positive correlation between α_1 and transfer confidence. The higher the confidence, the higher the knowledge transferability.