A. Negative Transfer Definition

In this section we show how the negative transfer condition (NTC) in Eq.(3) is derived.

The intuitive definition given earlier in Section 3 does not leads to a rigorous definition. There are two key questions that are not clear: (1) Should negative transfer be defined to be algorithm-specific? (2) What is the negative impact being compared with?

First, if negative transfer is completely algorithm-agnostic, then its definition would be independent to which transfer learning algorithm is being used. Mathematically, this may yield the following:

$$\min_{A} R_{P_T}(A(\mathcal{S}, \mathcal{T})) > \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})). \tag{12}$$

However, it is easy to see that this condition is never satisfied. To show this, given source data $\mathcal{S}$ and target data $\mathcal{T}$, consider an algorithm A_1 that minimizes the expected risk on the RHS:

$$A_1 \in \operatorname*{argmin}_{A'} \ R_{P_T}(A'(\emptyset, \mathcal{T})).$$

Then we can always construct a new algorithm A_1' such that $A_1'(\mathcal{S}, \mathcal{T})) = A'(\emptyset, \mathcal{T})$, i.e. A_1' always ignores the source data. As a result, we must have:

$$\begin{aligned} \min_{A} R_{P_T}(A(\mathcal{S}, \mathcal{T})) &\leq R_{P_T}(A_1'(\mathcal{S}, \mathcal{T})) \\ &= R_{P_T}(A_1(\emptyset, \mathcal{T})) \\ &= \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})) \end{aligned} \tag{13}$$

Therefore, the condition defined in Eq.(12) is never true and we conclude that negative transfer must be algorithm-specific. This answers the first question.

Given the answer, the condition in Eq.(12) could be modified to consider only a specific transfer algorithm A, i.e.,

$$R_{P_T}(A(\mathcal{S}, \mathcal{T})) > \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})). \tag{14}$$

However, there are still two problems with this definitions:

(a) This condition cannot be measured in practice since we cannot evaluate the RHS even at test time;

(b) An algorithm that does not utilize any source at all still satisfies the condition, which is counterintuitive. For instance, consider a degenerated algorithm A_2 such that $A_2(\mathcal{S}, \mathcal{T}) = A_2(\emptyset, \mathcal{T})$ and $R_{P_T}(A_2(\emptyset, \mathcal{T})) > \min_{A'} R_{P_T}(A'(\mathcal{S}, \mathcal{T}))$. This algorithm does not perform any meaningful transfer from the source, but negative transfer occurs in this case according to Eq. (14) since:

$$R_{P_T}(A_2(\mathcal{S}, \mathcal{T})) = R_{P_T}(A_2(\emptyset, \mathcal{T})) > \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})).$$

Therefore, it is misleading to only compare with the best possible algorithm and we propose the following definition:

Definition 1 (Negative Transfer). *Given a source dataset $\mathcal{S}$, a target dataset $\mathcal{T}$ and a transfer learning algorithm A, the negative transfer condition (NTC) is defined as:*

$$R_{P_T}(A(\mathcal{S}, \mathcal{T})) > R_{P_T}(A(\emptyset, \mathcal{T})) \geq \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})), \tag{15}$$

which is exactly Eq.(3) since the "$\geq$" constraint on the right side is true for any A. This definition of NTC resolves the two questions mentioned above. Furthermore, it is consistent with the intuitive definition and is also tractable at test time.