Agnostic Transfer Learning

Qian Yu, Salman Avestimehr

Abstract

How can one extract knowledge for solving new problems based on possibly related existing information, especially for general practical scenarios where no explicit similarity is assumed between them? We formulate a theoretical framework, called agnostic transfer learning, which is the first formulation to study the fundamental limits of learning algorithms for this problem. We show that although a "no-free-lunch" theorem appears to suggest that there is an inherent trade-off for different problem scenarios: to achieve better performances when the source of side information is "more relevant" to the target, one has to sacrifice the performances when they are "less relevant". Interestingly, there is a simple algorithm that automatically adjusts to different problem scenarios, and we prove it universally achieves the optimum loss within a constant factor by developing a minimax error bound.

1 Introduction

An important component of the human learning process involves transferring existing knowledge to apply them in new environments. For example, it is in general much easier to learn a second language after establishing the first language, as they do not need to be learned from scratch. This knowledge transfer capability is also a goal to be achieved in machine learning research, for common scenarios that the size of the dataset for the target learning task is limited, but a large amount of data or well-trained models are available from related tasks.

This problem has been approached theoretically from two directions. On the one hand, algorithmic solutions are developed, mainly focusing on a setting called domain adaptation, where a model or hypothesis is to be learned for a target data distribution, but a significant fraction or all available data are sampled from a possibly different distribution, called the source distribution. For example, see [3, 11, 30, 31]. In these results, explicit similarity conditions between the datasets, such as the closeness of unlabeled distributions, covariate shift, or closeness of labeling functions, are assumed or used as a measurement. Algorithms are mostly designed based on empirical risk minimization and re-weighting techniques to achieve smaller losses when certain similarity requirements are satisfied.

On the other hand, impossibility theorems have been developed to understand the fundamental limits of learning algorithms. One result is the no-free-lunch theorem in supervised learning [34], which states that learning algorithms will not benefit from side information that is not correlated to the target dataset. In other words, for transfer learning, without assuming any similarities between the source and target domain, reducing learning error for certain problem scenarios results in an increased error in other cases. Explicit examples are also provided in [13] showing that in some worst-case scenarios, learnability can not be achieved if certain similarity assumptions are not satisfied.

However, a gap between existing theoretical approaches and practical learning scenarios is that real-world datasets may not always be similar in any predetermined mathematical sense (e.g., with small L_1 distances between the data distributions) even if they share a significant amount of common information. Under these scenarios, directly applying conventional approaches such as weighted empirical risk minimization may fail to exploit existing knowledge from the source dataset and potentially lead to bad performances. An example is that learning to read ciphered text is significantly easier with an established language model [33], even though the ciphered text can be defined on an alphabet that is completely disjoint to that of the natural language, indicating large divergence between data distributions.

In this work, we establish a theoretical framework, called *agnostic transfer learning*, to study and develop algorithms for this general transfer learning scenario. In particular, we study the fundamental limits when no similarity assumptions are made, and to overcome the technical challenges for characterizing the (potentially complicated) optimal trade-offs for learning errors in different problem scenarios.

By developing a learning algorithm and proving lower bounds, we completely characterized the achievable region of learning errors among all possible algorithms. We show that as a fundamental trade-off, given any realization of side information, a hypothesis class $\mathcal{H}$, and a parameter $\delta < \frac{1}{4}$, even if an additional information is provided that the risk function is minimized by a certain subset of hypotheses of size $r > 1$, one can only guarantee achieving a learning error of $\Omega(\min\{\sqrt{\frac{\log \frac{r}{\delta}}{n}}, 1\})$ with n data points in the target dataset, with a probability of $1 - \delta$. Interestingly, we show that there are algorithms that universally achieves this fundamental limit:

Theorem (Informal)**.** *For any sequence of hypotheses $h_1, h_2, ...$, there is a learning algorithm that achieves the minimax optimal learning error $\Theta(\min\{\sqrt{\frac{\log \frac{r}{\delta}}{n}}, 1\})$ whenever the risk in target domain is minimized by any hypotheses h_r for $r > 1$ and an error of 0 when the risk is minimized by h_1, with a probability of $1 - \delta$.*

Intuitively, the above result shows that given any instance of prior knowledge, and any interpretation of this knowledge that ranks the hypothesis class in an order, even if there is a mismatch in the sense that the hypothesis minimizing the

risk function is only ranked within the first r elements, the proposed algorithm still automatically achieves an optimal loss (within a constant factor) as if it knows the identity and removes any potential hypotheses that are ranked lower.

Practical application. While we focus on a general theoretical setting, our results directly apply to a practical scenario where a local model is to be customized on a personal device to improve user's experience. The user has access to a small amount of local data (target domain) and a class of candidates provided by some learning agents that are trained based on external datasets. The user device would like to decide on an explicit model, but does not know which of the candidates provides a closer estimate to the underlying distribution of the target domain. However, besides the available target data, she also has a list of personal preferences from the user based on prior experiences (the knowledge $Z = (h_1, h_2, ...)$ defined in Section 3). Our result shows an interesting fact that the local device could tailor the model to user's preference without significantly penalizing the learning performance for cases where the provided ordering of the candidates are not aligned to their generalization loss on the target domain. The proposed method also allows benefits of on-device learning [5, 6, 14], to provide user privacy and operate under limited computing budgets.

The rest of this paper is organized as follows. Section 2 provides a more detailed review of prior works. Section 3 formally introduces the settings and formulates the agnostic transfer learning framework. Section 4 presents the main results, which are proved in Section 5 and 6. Section 7 includes a discussion on generalization of the proposed agnostic transfer learning framework and future directions.

2 Prior works

Compared to prior results in transfer learning, works in domain adaptation (e.g, [2, 3, 9–11, 25, 30, 31], as well as multiple-source adaptation [12, 20, 26, 27]) focuses on developing algorithms to provide non-trivial learning bounds when the source and target distributions are close under some explicit metrics (e.g., with small discrepancy, or satisfies the covariate shift assumption). Here we take a different approach to show that non-trivial results could still be obtained with minimum assumptions, and complete characterization can be obtained.

There have been several other lines of works with a similar flavor. The authors of [15, 18, 23] considered a setting named hypothesis transfer learning, where the similarity between source and target domain is measured the transfer-exponent, which bounds errors of same hypotheses on different datasets. In particular, [18] presented a minimax learning bound for this setting. [7] consider a more explicit and stronger requirement between different domains, using a measure called relative signal exponent, which assumes a weaker version of covariant shift and that both domains share the same hard-decision maximum likelihood estimator. [21, 22, 32] assume that the distributions of all domains are generated with the same representation, which effectively reduces the complexity

of the hypothesis classes and enables improved learning bounds. There are also scenarios where the knowledge is to be transferred from pre-trained models, and to generate new models for the target domain. [24] uses linear combinations of auxiliary models and input features to show that a faster rate of generalization can be achieved if the models are combined with good coefficients. [17] presented an experimental approach to combine models with the same structures layer-wisely.

More recently, [19] presented an alternative no-free-lunch theorem for transfer learning, which instead focuses on the fundamental limit due to discrepancy between multiple source domains, rather than the discrepancy between source and target domain as considered in our paper. They have shown that even with bounded transfer exponent between source and target, the benefits of including multiple source datasets could still be limited if the algorithm is required to be adaptive.

Transfer learning has also been approached using Bayesian interference, to build a prior for the target domain with knowledge learned from earlier training tasks [16]. For example, in [8, 29], Gaussian priors are constructed by estimating their associated covariances; [1] provides an asymptotic analysis for learn priors using a hyper-prior and with sufficiently many training task. Compared to Bayesian inference, our setting does not assume a prior distribution or similar quantitative measures. Instead, our algorithm operated in a non-Bayesian setting, which only relying on a possibly mismatched rankings of the hypotheses determined by the user.

In another related direction called model selection [4, 28], the goal is to select a subset of "simpler" hypotheses for restricting the learning algorithms to avoid over-fitting. Unlike conventional transfer learning problems such as domain adaptation, model selection focuses on a supervised learning setup where no additional datasets are available, but the algorithms are build given a partial ordering of the entire hypothesis set which defines the simplicity, and a belief that the data distribution can be reasonably fit by certain subsets of simpler hypotheses defined in this sense. Given the formulation provided in this paper, model selection can fundamentally be viewed as a transfer learning problem, by observing that any definition of simplicity is essentially some prior knowledge (the variable Z defined in Section 3) learned from earlier experiences, and the goal is similarly to minimize some risk function given a dataset utilizing this side information. Note that the algorithmic upper bound provided in this paper can be viewed as a modified version of structural risk minimization (see [28], Chapter 4), and we completed the characterization be proving matching lower bounds.

Our presented lower bound also applies to standard supervised learning. In particular, lower bounds that were proposed for, or can be applied to supervised learning in the literature are mostly point-wise, in the sense that they focused on proving the unachievability for one specific pair of the learning error and the error probability for achieving it [18, 28]. We provided a uniform lower bound on the learning error for the worst case data distribution, for general values of the error probability required to be achieved.

3 Preliminaries

We consider a general setting, where the learner has access to n data points from the target domain, denoted $\{(x_i, y_i)\}_{i=1}^n$. Each data point is defined on sets $\mathcal{X}$ and $\mathcal{Y}$ and is i.i.d. sampled from a distribution P. The learner is also provided with some prior knowledge, denoted as a variable Z, which is an ordered sequence of hypotheses $h_1, h_2, ...$ that belongs to the hypotheses set $\mathcal{H}$.

Given a (possibly infinite) hypothesis class[1] $\mathcal{H}$ and a loss function $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to [0, 1]$, the goal of a learning algorithm is to minimize the following risk function:

$$R(h) = \mathbb{E}_P[l(x, y, h)].$$

Or equivalently, minimizing the *learning error* defined as follows

$$\epsilon(h) = R(h) - \inf_{h_0 \in \mathcal{H}} R(h_0).$$

We denote the set of all possible loss functions by $\mathcal{L}$. The algorithm can exploit side information provided by Z arbitrarily. In particular, an algorithm is characterized by a (possibly random) function $\pi_Z : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{L} \to \mathcal{H}$ parameterized by Z, that maps the values of target data and the given loss function to a hypothesis.

We consider an agnostic setting, where no assumptions are imposed on distribution P and loss l, and we study the learning errors achievable by any algorithm under scenarios where the risk is minimized by different hypothesis. Given each $h \in \mathcal{H}$, we denote $\mathcal{P}_h$ the subclass of scenarios where h is the best hypothesis. Explicitly,

$$\mathcal{P}_h \triangleq \{(P, l) \mid \epsilon(h) = 0\}.$$

We aim to find learning algorithms that achieves low learning error with some high probability $1 - \delta$ in each regime, characterized by a vector ϵ_h^*, called the *learning error function*, given any parameter $\delta < \frac{1}{4}$. Formally, we say an algorithm π_Z achieves a learning error function ϵ_h^* with a probability of $1 - \delta$, if and only if

$$\mathbb{P}_P[\epsilon(\pi_Z(\{(x_i, y_i)\}_{i=1}^n, l)) \leq \epsilon_h^*] \geq 1 - \delta$$

for any hypothesis h and any $(P, l) \in \mathcal{P}_h$.

4 Main Results

We summarize our main results in following theorems.

[1] We assume that the sets $\mathcal{X}$ and $\mathcal{Y}$ are sufficiently large whenever needed, so the hypotheses are distinguishable.

Theorem 1 (Algorithmic Upper Bound)**.** *Given parameter δ, a learning error function $\{\epsilon_h^*\}_{h\in\mathcal{H}}$ is achievable with a probability of $1-\delta$ if there is a sequence $h_1, h_2...$, in $\mathcal{H}$, such that*

$$\epsilon_h^* \geq \begin{cases} 0 & h = h_1 \\ O\left(\min\left\{\sqrt{\frac{\log\frac{k}{\delta}}{n}}, 1\right\}\right) & h = h_k \text{ for any } k > 1 \\ O(1) & \text{otherwise.} \end{cases}$$

Theorem 2 (Matching Lower Bound)**.** *Given any parameter $\delta < \frac{1}{4}$, a learning error function $\{\epsilon_h^*\}_{h\in\mathcal{H}}$ is achievable with a probability of $1-\delta$ only if there is a sequence $h_1, h_2...$, in $\mathcal{H}$, such that*

$$\epsilon_h^* \geq \begin{cases} 0 & h = h_1 \\ \Omega\left(\min\left\{\sqrt{\frac{\log\frac{k}{\delta}}{n}}, 1\right\}\right) & h = h_k \text{ for any } k > 1 \\ \Omega(1) & \text{otherwise.} \end{cases}$$

The proofs of Theorem 1 and 2 can be found in Section 5 and 6.

Remark 1 (Algorithmic Interpretation). The algorithmic result in Theorem 1 can be interpreted as follows. Given any side information Z and any pseudo-risk function based on Z that assign each hypothesis a distinct value, we show that one can find an algorithm that is consistent with this pseudo-ranking, in the sense that for any problem scenario where the best hypothesis has a lower pseudo-risk (i.e., a hypothesis h_k with a smaller k), the proposed algorithm achieves a lower learning error with the same probability.

Remark 2 (Universal Optimality). Theorem 2 shows that for any learning algorithm, if one consider the problem scenarios where the risk function is minimized by one of the k-"best" hypothesis indicated by the side information Z, then the minimum learning error achievable with a probability of $1-\delta$ in the worst case is $\Omega(\min\sqrt{\frac{\log\frac{k}{\delta}}{n}}, 1)$ for any $k > 1$. This lower bound matching Theorem 1 shows that the proposed algorithm automatically and universally achieves the minimax learning error for any level of mismatch between the target domain and the prior knowledge Z, without requiring accessing that information.

Remark 3 (Comparison to no-free-lunch Theorem). Perhaps the most interesting aspect of the above results is that we are able to find an algorithm that universally achieves the minimax learning error for all scenarios where the loss is minimized by different hypotheses, albeit a no-free-lunch theorem can be proved (the proof is provided in Appendix A):

Theorem 3 (No-free-lunch for transfer learning)**.** *Let $\{(P_i, l_i)\}_{i=1}^k$ be any set of problem scenarios, if Z is a random variable independent of the identity i, then for any algorithm π_Z that uses both the target dataset and prior knowledge Z does not outperform algorithms that only uses the target set in terms of the weighted average learning loss in these scenarios. Formally, we can always find*

an algorithm, denoted by π, such that

$$\begin{aligned}&\sum_i w_i \text{Dist}_{P_i}[\epsilon(\pi_Z(\{(x_i,y_i)\}_{i=1}^n, l_i))]\\ =&\sum_i w_i \text{Dist}_{P_i}[\epsilon(\pi(\{(x_i,y_i)\}_{i=1}^n, l_i))]\end{aligned} \tag{1}$$

for any weight function $\{w_i\}_{i=1}^k \in \mathbb{R}_+^k$, where Dist *denotes the distribution function of a random variable.*

The significance of no-free-lunch theorem is that for general frameworks where no assumptions are made on the relationship between the target distribution P and prior knowledge Z, any algorithm which strictly improves the learning performance with the help of the side information Z for some problem scenarios must be at the cost of performances in other scenarios. However, we show that there is a universally optimal algorithm which can order-wise reduce the learning error down to 0 for scenarios that are more "favored" by the side information while incurring additional errors in all other cases by at most a constant factor, as if it generates "free information" on the level of mismatch between the available dataset and side information.

5 Algorithmic Upper Bound

In this section, we propose a general class of learning algorithms to achieve the learning bound stated in Theorem 1 (see Algorithm 1). In particular, we need to show that given the side information $Z = (h_1, h_2, ...)$, for any $k \in \mathbb{N}_+$ and any problem scenario (P, l) where the risk function in target domain is minimized by h_k, the presented algorithm returns a model $\hat{h}$ that achieves a learning error $\epsilon(\hat{h}) = \epsilon^*(k)$ defined as follows

$$\epsilon^*(k) \triangleq \begin{cases} 0 & k = 1 \\ 4\sqrt{\frac{\log \frac{k}{\delta}}{n}} & \text{otherwise.} \end{cases}$$

with a probability of $1 - \delta$.

Algorithm 1 Agnostic Transfer Learning

1: Input: Target dataset $\{(x_i, y_i)\}_{i=1}^n$, loss function l, prior knowledge $Z = (h_1, h_2, ...)$, a target failure probability δ
2: **Define** $R_{\text{Emp}}(h) = \frac{1}{n}\sum_{i=1}^n l(x_i, y_i, h)$
3: **Define** $\epsilon^*(k) = \begin{cases} 0 & k = 1 \\ 4\sqrt{\frac{\log \frac{k}{\delta}}{n}} & \text{otherwise} \end{cases}$
4: **Find** $\hat{k} = \text{argmin}_j R_{\text{Emp}}(h_j) + \frac{1}{2}\epsilon^*(j)$
5: **Return** $\hat{h} = h_{\hat{k}}$

The main idea of the algorithmic construction is to modify the empirical risk minimization approach by adding a carefully designed regularization term, such that hypotheses with a smaller index determined by the side information are more favored in the optimization stage to achieve smaller learning errors when they minimizes the target risk function $R(h)$, while the ones with higher indices are not significantly penalized.

The detailed analysis is presented as follows. Let $\hat{h}$ be the hypothesis returned by the algorithm. For any $k \in \mathbb{N}_+$ and any problem scenario $(P, l) \in \mathcal{P}_{h_k}$, we consider the probability that the proposed algorithm does not achieve the learning error $\epsilon^*(k)$. Applying the union bound, we have

$$\begin{aligned}
&\mathbb{P}[\epsilon(\hat{h}) > \epsilon^*(k)] \\
&\leq \mathbb{P}[\exists j \neq k : \tilde{R}(h_j) \leq \tilde{R}(h_k) \text{ and } \epsilon(h_j) > \epsilon^*(k)] \\
&\leq \sum_{j \neq k} \mathbb{P}[\tilde{R}(h_j) \leq \tilde{R}(h_k) \text{ and } \epsilon(h_j) > \epsilon^*(k)] \\
&\leq \sum_{j \neq k} \mathbb{P}[\tilde{R}(h_j) \leq \tilde{R}(h_k) + \epsilon(h_j) - \epsilon^*(k)]
\end{aligned}$$

where $\tilde{R}(h_j) \triangleq R_{\mathrm{Emp}}(h_j) + \frac{1}{2}\epsilon^*(j)$ for each $j \in \mathbb{N}_+$ is the regularized empirical risk used in the proposed algorithm. Recall that the expectation of $R_{\mathrm{Emp}}(h)$ is the risk function $R(h)$, and $\epsilon(h) = R(h) - R(h_k)$. Thus,

$$\begin{aligned}
&\mathbb{P}[\epsilon(\hat{h}) > \epsilon^*(k)] \\
&\leq \sum_{j \neq k} \mathbb{P}[R_{\mathrm{Emp}}(h_j) - R_{\mathrm{Emp}}(h_k) \leq \\
&\quad - \frac{1}{2}\epsilon^*(j) - \frac{1}{2}\epsilon^*(k) + \epsilon(h_j)] \\
&= \sum_{j \neq k} \mathbb{P}[R_{\mathrm{Emp}}(h_j) - R_{\mathrm{Emp}}(h_k) \leq \\
&\quad \mathbb{E}[R_{\mathrm{Emp}}(h_j) - R_{\mathrm{Emp}}(h_k)] - \frac{1}{2}\epsilon^*(j) - \frac{1}{2}\epsilon^*(k)].
\end{aligned}$$

Note that $R_{\mathrm{Emp}}(h_j) - R_{\mathrm{Emp}}(h_k)$ is the average of n i.i.d. random variables $l(x_i, y_i, h_j) - l(x_i, y_i, h_k)$ that are bounded within $[-1, 1]$, we can apply Hoeffd-

ing's inequality and obtain the following inequalities:

$$\begin{aligned}
\mathbb{P}[\epsilon(\hat{h}) > \epsilon^*(k)] &\leq \sum_{j \neq k} \exp(-n\frac{(\epsilon^*(j)+\epsilon^*(k))^2}{8}) \\
&\leq \sum_{j<k} \exp(-n\frac{\epsilon^*(k)^2}{8}) \\
&\quad + \sum_{j>k} \exp(-n\frac{\epsilon^*(j)^2}{8}) \\
&\leq (k-1)(\frac{\delta}{k})^2 + \sum_{j>k}(\frac{\delta}{j})^2 \\
&\leq (\frac{\pi^2}{6} - 1)\delta^2 \\
&\leq \delta.
\end{aligned}$$

To conclude, we have proved that Algorithm 1 achieves the stated learning error bound with a probability of at least $1-\delta$, which completes the proof of Theorem 1.

Remark 4 (Fine-tuning General Algorithms for Minimax Optimality). The proposed algorithm provides a general theoretical approach for adapting any learning algorithm into a dominant version, which achieves a same or strictly improved learning error function (modulo a constant factor) for agnostic transfer learning. Concisely, given any learning algorithm and for any fixed side information Z, one can observe the worst case learning error achievable with probability $1-\delta$ among problem scenarios $(P, l) \in \mathcal{P}_h$ for each hypothesis h. Based on the impossibility result stated in Theorem 2, there are at most finitely many hypotheses that allows for achieving a learning error that is lower than $O(1)$. By sorting these hypotheses, one can construct a sequence $h_1, h_2, ...$ with their corresponding worst case learning errors in non-decreasing order, which is then used in Algorithm 1.

As a demonstrating example, consider the domain adaptation algorithm mentioned in [3], which applies for scenarios where the side information Z is an additional dataset (called the source dataset) where the elements are sampled from a different distribution Q. The algorithm presented in [3] computes a linear combination of the empirical risks in the two domains and finds the hypothesis minimizing this function. This can be interpreted as using the empirical risk in source domain for regularization.

One can show that by applying the approach described in Remark 4, we essentially replaced the empirical risk regularizer (denoted $R_{Q,\mathrm{EMP}}$) by a more

structured function

$$\phi(h) \triangleq \begin{cases} 0 & h \text{ minimizes } R_{Q,\mathrm{EMP}} \\ 2\sqrt{\frac{\log \frac{k}{\delta}}{n}} & h \text{ has the } k\text{-th smallest } R_{Q,\mathrm{EMP}} \\ & \text{for } k > 1 \\ 2 & \text{otherwise.} \end{cases}$$

This structure allows for a strict improvement even when the source dataset can be arbitrarily large, because in the most general case there is no guarantee on the values of the risks achieved by each hypothesis in the source domain. As a result, directly using the empirical risk on source domain will not guarantee achieving minimax learning error in the agnostic setting, while the structured regularizer provided in the refined version achieves universal optimailty.

6 Matching Lower Bound

We now prove lower bounds on the achievable learning errors stated in Theorem 2. Equivalently, it suffices to prove that for any finite subset $\mathcal{H}' \subseteq \mathcal{H}$ with a size $|\mathcal{H}'| > 1$ and any function $\{\epsilon_h^*\}_{h\in\mathcal{H}}$ achievable with a probability of $1-\delta$, we have $\max_{h_\in \mathcal{H}'} \epsilon_h^* \geq \frac{1}{4}\min\{\sqrt{\frac{\log \frac{|\mathcal{H}'|}{\delta}}{n}}, 1\}$. Intuitively, this statement shows that if the risk function is known to be minimized by a certain subclass of hypotheses, and such information is leaked to the learning algorithm, the lower bound to be proved still holds true.

Fixing any subset $\mathcal{H}'$, we consider a class of problem scenarios, denoted $\{(P_{h'}, l_{h'})\}_{h'\in\mathcal{H}'}$, such that for each h', the loss $l(x_i, y_i, h)$ is an i.i.d. Bernoulli random variable with an expectation of $\frac{1+(-1)^{\mathbb{1}[h=h']}\epsilon}{2}$ for some $\epsilon > 0$. Clearly, $(P_{h'}, l_{h'}) \in \mathcal{P}_{h'}$ for each $h' \in \mathcal{H}'$. Hence, if $\epsilon > \frac{1}{4}\min\{\sqrt{\frac{\log \frac{|\mathcal{H}'|}{\delta}}{n}}, 1\}$, it suffices to prove that given this subclass of scenarios, the maximum probability over $\mathcal{H}'$ that any learning algorithm fails to return h' in scenario $(P_{h'}, l_{h'})$ is greater than δ. To that end, it suffices to lower bound the average of such probabilities over $h' \in \mathcal{H}'$ by δ.

Note that this quantity is minimized by the Maximum Likelihood estimator, which can be express in closed-form. We state it formally in the following lemma.

Lemma 1. *Given a finite set of independent random variables $\{X_h\}_{h\in\mathcal{H}'}$ each with a distribution $X_h \sim \mathrm{Ber}(\frac{1+(-1)^{\mathbb{1}[h=h']}\epsilon}{2})$ for some parameter $h' \in \mathcal{H}'$. Let $\hat{h}_{\mathrm{ML}}$ denotes the maximum likelihood estimation of h' by any ML estimator given*

n samples of the entire set, then

$$P_{\mathrm{ML}} \triangleq \frac{1}{|\mathcal{H}'|} \sum_{h' \in \mathcal{H}'} \mathbb{P}[\hat{h}_{\mathrm{ML}} \neq h'] = \sum_{i=0}^{n} P_{\mathrm{Bi}}(i;n)$$
$$(1 - \sum_{j=0}^{|\mathcal{H}'|-1} \binom{|\mathcal{H}'|-1}{j} \frac{1}{j+1} P_{\mathrm{Bi}}^{j}(n-i;n)$$
$$(\sum_{k<i} P_{\mathrm{Bi}}(n-k;n))^{|\mathcal{H}'|-1-j})$$

where $P_{\mathrm{Bi}}(x;n)$ denotes the binomial mass function with n trials and the probability of success of each trial equals $\frac{1+\epsilon}{2}$, i.e.,

$$P_{\mathrm{Bi}}(x;n) \triangleq \binom{n}{x} (\frac{1+\epsilon}{2})^{x} (\frac{1-\epsilon}{2})^{n-x}. \tag{2}$$

Proof. Let M_h denote the sum of n samples for each X_h. One can show that an ML estimator returns a hypothesis h with the largest realization of M_h. Because the tie-breaking rules does not affect the value of average error probability, without loss of generality, it suffices to consider the ML estimator that breaks ties uniformly randomly. The resulting error probability is given as follows.

$$P_{\mathrm{ML}} = \sum_{i=0}^{n} \mathbb{P}[M_{h'} = i] \left(1 - \mathbb{P}[\hat{h}_{\mathrm{ML}} = h' | M_{h'} = i]\right)$$
$$= \sum_{i=0}^{n} \mathbb{P}[M_{h'} = i](1 - \sum_{\mathcal{S} \subseteq \mathcal{H}' \setminus \{h'\}} \frac{1}{|\mathcal{S}|+1}$$
$$\prod_{h \in \mathcal{S}} \mathbb{P}[M_h = i] \prod_{h \in \mathcal{H}' \setminus \{h'\} \setminus \mathcal{S}} P(M_h < i))$$

Note that M_h is a binomial random variable with a mass function

$$\mathbb{P}[M_h = i] = \begin{cases} P_{\mathrm{Bi}}(i;n) & h = h' \\ P_{\mathrm{Bi}}(n-i;n) & \text{otherwise.} \end{cases}$$

Lemma 1 directly follows.

□

Because the average probability of error for any learning algorithm is lower bounded by that of any ML estimator, the rest of the proof remains to show this lower bound is greater than δ, essentially the following lemma.

Lemma 2. *For any integers $n \geq 1$, $|\mathcal{H}'| > 1$ and parameter $\delta < \frac{1}{4}$, we can find $\epsilon > \frac{1}{4} \min\{\sqrt{\frac{\log \frac{|\mathcal{H}'|}{\delta}}{n}}, 1\}$ such that*

$$P_{\mathrm{ML}} > \delta. \tag{3}$$

6.1 The $|\mathcal{H}'| = 2$ case

We start from the base case where $|\mathcal{H}'| = 2$. In this case, the learning process reduces to a binary hypothesis testing problem and P_{ML} is simply characterized by binomial distribution functions.

$$\begin{aligned} P_{\mathrm{ML}} &= \sum_{i=0}^{n} P_{\mathrm{Bi}}(i;n)(1 - \sum_{k<i} P_{\mathrm{Bi}}(n-k;n) \\ &\quad - \frac{1}{2} P_{\mathrm{Bi}}(n-i;n)) \\ &= \sum_{i=0}^{n-1} P_{\mathrm{Bi}}(i;2n) + \frac{1}{2} P_{\mathrm{Bi}}(n;2n) \end{aligned}$$

When $n \leq 2$, let $\epsilon = 0.3$, which satisfies the condition $\epsilon > \frac{1}{4}\min\{\sqrt{\frac{\log\frac{|\mathcal{H}'|}{\delta}}{n}}, 1\}$, one can compute the exact values of the corresponding error probabilities and verify $P_{\mathrm{ML}} > \frac{1}{4} > \delta$. Hence, we focus on $n > 2$, and consider the following two possible cases.

Case a: $\sqrt{\frac{\log\frac{2}{\delta}}{n}} \leq 1$. We choose $\epsilon = 0.26\sqrt{\frac{\log\frac{2}{\delta}}{n}}$, which satisfies the required condition. To find a close estimate of P_{ML}, we use the following inequality on binomial coefficient.[2]

Lemma 3. *For any integer n and $0 < k < n$,*

$$\binom{n}{k} \geq \frac{\sqrt{n}}{\sqrt{8k(n-k)}} (\frac{k}{n})^{-k} (1 - \frac{k}{n})^{-(n-k)}.$$

Using this combinatorial inequality, we can lower bound P_{ML} for $n \geq 3$ as follows. For any positive integer d,

$$\begin{aligned} P_{\mathrm{ML}} &\geq \sum_{i=1}^{d} P_{\mathrm{Bi}}(n-i;2n) + \frac{1}{2} P_{\mathrm{Bi}}(n;2n) \\ &\geq \frac{1}{2\sqrt{n}} (1-\epsilon^2)^n (\sum_{i=1}^{d} (\frac{1-\epsilon}{1+\epsilon})^i \cdot \frac{\prod_{j=0}^{i-1}(n-j)}{\prod_{j=1}^{i}(n+j)} + \frac{1}{2}). \end{aligned}$$

For convenience, let $Z(\epsilon, n, i) \triangleq (\frac{1-\epsilon}{1+\epsilon})^i \cdot \frac{\prod_{j=0}^{i-1}(n-j)}{\prod_{j=1}^{i}(n+j)}$. Recall that in this case $\epsilon \leq 0.26$.

$$P_{\mathrm{ML}} \geq \frac{1}{2\sqrt{n}} e^{-n\epsilon^2 \frac{|\log(0.9324)|}{0.26^2}} (\sum_{i=1}^{d} Z(\epsilon, n, i) + \frac{1}{2}). \tag{4}$$

Now we consider the variation of the above upper bound with respect to ϵ, to show that the "hardest" case is where ϵ approaches its infimum. Firstly,

$$\frac{\partial}{\partial \epsilon} \log Z(\epsilon, n, i) = -\frac{2i}{1-\epsilon^2}.$$

[2] A proof can be found in Appendix B.

On the other hand,

$$\frac{\partial}{\partial \epsilon} \log \delta = -2\frac{\epsilon n}{0.26^2}.$$

Hence, as long as we let $d \leq 4.9\sqrt{n}$, we have the following inequality for bound (4).

$$\frac{\partial}{\partial \epsilon}(\log \mathrm{RHS} - \log \delta) \geq -\frac{2d}{1-\epsilon^2} + 2\frac{\epsilon n \log(0.9324e)}{0.26^2} > 0.$$

As a consequence, it suffices to prove the RHS of inequality (4) is greater than or equal to $\frac{1}{4}$ for $\epsilon = 0.26\sqrt{\frac{\log 8}{n}}$ and some $d \leq 4.9\sqrt{n}$. This can be directly verified for $n \leq 10^5$. To provide a proof for $n > 10^5$, we pick $d = \lceil 1.5\sqrt{n} \rceil$ and lower bound $Z(\epsilon, n, i)$ as follows.

$$\begin{aligned}
Z(\epsilon, n, i) &= \exp\left(i \log \frac{1-\epsilon}{1+\epsilon} + \sum_{j=1}^{i} \log \frac{n-j+1}{n+j}\right) \\
&\geq \exp\left(i \log \frac{1-\epsilon}{1+\epsilon} - \frac{i^2}{n-d+1}\right) \\
&\geq \exp\left(\frac{1}{2} \log \frac{1-\epsilon}{1+\epsilon} - \frac{d-\frac{1}{4}}{n-d+1}\right) \\
&\quad \cdot \exp\left((i-\frac{1}{2}) \log \frac{1-\epsilon}{1+\epsilon} - \frac{(i-\frac{1}{2})^2}{n-d+1}\right).
\end{aligned}$$

One can show that $\frac{1}{2} \log \frac{1-\epsilon}{1+\epsilon} - \frac{d-\frac{1}{4}}{n-d+1} > \log 0.99$ for $n > 10^5$, hence

$$\begin{aligned}
&\sum_{i=1}^{d} Z(\epsilon, n, i) + \frac{1}{2} \\
>&0.99 \int_0^{d+\frac{1}{2}} \exp\left(x \log \frac{1-\epsilon}{1+\epsilon} - \frac{x^2}{n-d+1}\right) dx \\
=&\frac{0.99\sqrt{\pi}}{2} \sqrt{n-d+1} \cdot \exp(\frac{n-d+1}{4} \ln^2 \frac{1-\epsilon}{1+\epsilon}) \\
&\cdot (\mathrm{erf}(\frac{d+\frac{1}{2}}{\sqrt{n-d+1}} - \frac{\sqrt{n-d+1}}{2} \log \frac{1-\epsilon}{1+\epsilon}) \\
&- \mathrm{erf}(-\frac{\sqrt{n-d+1}}{2} \log \frac{1-\epsilon}{1+\epsilon})).
\end{aligned}$$

Note that $\frac{d+\frac{1}{2}}{\sqrt{n-d+1}} > 1.5$ and $-\frac{\sqrt{n-d+1}}{2}\log\frac{1-\epsilon}{1+\epsilon} < \epsilon\sqrt{n} < 0.375$.

$$\begin{aligned}P_{\mathrm{ML}} >& \frac{0.99\sqrt{\pi}\sqrt{n-d+1}}{4\sqrt{n}}\exp(-n\epsilon^2\frac{|\log(0.9324)|}{0.26^2}\\ &+\frac{n-d+1}{4}\ln^2\frac{1-\epsilon}{1+\epsilon})\cdot(\mathrm{erf}(1.875)-\mathrm{erf}(0.375))\\ >& \frac{1.03\sqrt{n-d+1}}{4\sqrt{n}}\exp(-0.146+\frac{n-d+1}{4}\ln^2\frac{1-\epsilon}{1+\epsilon}).\end{aligned}$$

Recall that $\frac{n-d+1}{n} \geq 1-\frac{1.5}{\sqrt{n}} > 0.995$ for $n > 10^5$ and $\ln\frac{1+\epsilon}{1-\epsilon} \geq 2\epsilon$.

$$P_{\mathrm{ML}} > \frac{1.02}{4} > \delta.$$

Case b: $\sqrt{\frac{\log\frac{2}{\delta}}{n}} > 1$. We choose $\epsilon = 0.26$ to satisfy the required condition. In this case, $\delta < 2e^{-n}$. Using the result of Case a where δ is at the minimum.

$$P_{\mathrm{ML}} > 2e^{-n} > \delta.$$

6.2 Proof for $|\mathcal{H}'| > 2$

Now we generalized the above results to $|\mathcal{H}'| \geq 3$. Similar to the base cases for $n \leq 2$, note that $\mathcal{P}_{\mathrm{ML}}$ is non-decreasing with respect to $|\mathcal{H}'|$, $\epsilon = 0.3$ satisfies the requirements. In the rest of the proof we will focus on $n \geq 3$. Observe that

$$\begin{aligned}P_{\mathrm{ML}} \geq& \sum_{i=0}^{n} P_{\mathrm{Bi}}(i;n)(1-\sum_{j=0}^{|\mathcal{H}'|-1}\binom{|\mathcal{H}'|-1}{j}\\ &\frac{P_{\mathrm{Bi}}^{j}(n-i;n)}{2^j}(\sum_{k<i}P_{\mathrm{Bi}}(n-k;n))^{|\mathcal{H}'|-1-j})\\ =& \sum_{i=0}^{n} P_{\mathrm{Bi}}(i;n)(1-(\frac{P_{\mathrm{Bi}}(n-i;n)}{2}\\ &+(\sum_{k<i}P_{\mathrm{Bi}}(n-k;n)))^{|\mathcal{H}'|-1},\end{aligned}$$

and equality holds true when $|\mathcal{H}'| = 2$. Let $F_h(x) \triangleq 1-(1-x)^{h-1}$, and $P_{\mathrm{ML},2}$ denotes the corresponding error function for $|\mathcal{H}'| = 2$, i.e.,

$$\begin{aligned}P_{\mathrm{ML},2} \triangleq& \sum_{i=0}^{n} P_{\mathrm{Bi}}(i;n)(1-(\frac{P_{\mathrm{Bi}}(n-i;n)}{2}\\ &+(\sum_{k<i}P_{\mathrm{Bi}}(n-k;n))))\end{aligned}$$

F_h is concave for $x \in [0,1]$ and $h \in \mathbb{N}_+$. Using Jensen's inequality,

$$P_{\mathrm{ML}} \geq F_{|\mathcal{H}'|}(P_{\mathrm{ML},2}).$$

Now we choose $\epsilon = 0.26 \min\{\sqrt{\frac{\log \frac{|\mathcal{H}'|}{\delta}}{n}}, 1\}$. Note that earlier in the $|\mathcal{H}|' = 2$ case, we have essentially proved that for any $\epsilon \in [0.26\sqrt{\frac{\log 8}{n}}, 0.26]$,

$$P_{\mathrm{ML},2} > 2 \exp\left(-\frac{n\epsilon^2}{0.26^2}\right).$$

Hence, for the selected ϵ,

$$P_{\mathrm{ML}} > F_{|\mathcal{H}'|}\left(\frac{2\delta}{|\mathcal{H}'|}\right).$$

Recall that $\delta < \frac{1}{4}$, due to concavity,

$$\begin{aligned} P_{\mathrm{ML}} &> 2|\mathcal{H}'| \cdot F_{|\mathcal{H}'|}(\frac{1}{2|\mathcal{H}'|}) \cdot \frac{2\delta}{|\mathcal{H}'|} \\ &= 4\delta(1 - (1 - \frac{1}{2|\mathcal{H}'|})^{|\mathcal{H}'|-1}) \\ &> 1.2\delta > \delta. \end{aligned} \tag{5}$$

7 Discussion

In this paper, we formulated an theoretical framework for transfer learning to study the fundamental limits of learning algorithms. We completely characterized the achievable region of learning error performances, and provided an approach to adapt any general algorithm into a dominant version for achieving an optimal operating point. To the best of our knowledge, this is the first result in transfer learning that does not rely on explicit assumptions between the available target training dataset and the side information.

As shown in our impossibility result as well as in prior works, if there are no constraints on the loss function l defined in Section 3, then any learning algorithm can only achieve a learning error smaller than $O(1)$ for a subclass of cases where the risk is minimized by finitely many hypothesis. However, one can achieve improved learning bounds if additional information is known on the structure of the hypothesis set and the loss function.

In particular, we define a generalized agnostic transfer learning framework as follows. Similar to Section 3, the learner has access to n data points from the target domain i.i.d. sampled from a distribution P, and some prior knowledge Z. However, it is known in addition that the loss function l must be within a particular subset denoted $\mathcal{L}_{\mathrm{sub}}$. We aim to find algorithms to optimize for the learning error function ϵ_h^* defined accordingly given $\mathcal{L}_{\mathrm{sub}}$.

This generalized framework encapsules common problem scenarios studied in the machine learning literature, including formulations with explicit loss functions such as 0-1 loss, quadratic loss, as well as more general assumptions such as on Lipschitzness or strong-convexity. Moreover, one can show that improved learning bounds can be developed for certain scenarios using well known tools such as VC-dimension, pseudo dimension, and fat-shattering dimension.

We would like to conclude the discussion with an interesting example showing that, unlike the main results presented in this paper, one can not always find universally optimal transfer learning algorithms for a general set $\mathcal{L}_{\text{sub}}$. Hence, characterising the achievable performances for generalized agnostic transfer learning framework remains as a future research direction.

Theorem 4. *There is a problem setting characterized by a set $\mathcal{L}_{\text{sub}}$ with a two-hypotheses set $\mathcal{H} = \{h_1, h_2\}$ such that no algorithm can achieve a learning error function $(\epsilon^*_{h_1}, \epsilon^*_{h_2})$ that is within a constant factor of the individual compound lower bounds $(\min_{\pi_Z} \epsilon^*_{h_1}, \min_{\pi_Z} \max\{\epsilon^*_{h_1}, \epsilon^*_{h_2}\})$.*

Proof. We consider a scenario where $\mathcal{X} = \{0\}$ and $\mathcal{Y} = \{0, 1\}$. The loss function is fixed and defined as

$$l(x, y, h) = \begin{cases} 1 & h = h_1 \text{ and } y = 1 \\ \epsilon_0 & h = h_2 \text{ and } y = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Recall that P can be any distribution on domain $\mathcal{X} \times \mathcal{Y}$. We parameterize P with a quantity q, defined as the probability that $y = 0$. Assuming ϵ_0 is small, when n data points are sampled, the minimum estimation error one can achieve simultaneously for all possibly P with probability $1 - \delta$ is at most ϵ_0, i.e., $\min_{\pi_Z} \max\{\epsilon^*_{h_1}, \epsilon^*_{h_2}\} \leq \epsilon_0$.

We prove Theorem 4 by contradiction. Assume such algorithm exists, then to satisfy the requirement for $q = 1$, the algorithm has to return h_1 with a probability of at least $1 - \delta$ when all n data points gives $y = 0$. On the other hand, for $P \in \mathcal{P}_{h_2}$, the resulting estimation error is at least $(1 - q - q\epsilon_0)q^n(1 - \delta)$. If ϵ_0 is small enough, let $q = 1 - 1/n$, the resulting loss is $\Omega((1 - \delta)/n) = \omega(\epsilon_0)$, which contradicts the assumption of $\epsilon^*_{h_2} = O(\epsilon_0)$.

□

References

[1] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010.

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt,

and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

[4] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[5] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tiny transfer learning: Towards memory-efficient on-device learning. *arXiv preprint arXiv:2007.11622*, 2020.

[6] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[7] T Tony Cai, Hongji Wei, et al. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Annals of Statistics*, 49(1):100–128, 2021.

[8] Bin Cao, Sinno Jialin Pan, Yu Zhang, Dit-Yan Yeung, and Qiang Yang. Adaptive transfer learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1), Jul. 2010.

[9] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. Curran Associates, Inc., 2010.

[10] Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 308–323, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[11] Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 169–178, New York, NY, USA, 2015. ACM.

[12] Corinna Cortes, Mehryar Mohri, Ananda Theertha Suresh, and Ningshan Zhang. Multiple-source adaptation with domain classifiers. *arXiv preprint arXiv:2008.11036*, 2020.

[13] Shai Ben David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[14] Sauptik Dhar, Junyao Guo, Jiayi Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. On-device machine learning: An algorithms and learning theory perspective. *arXiv preprint arXiv:1911.00623*, 2019.

[15] Simon S Du, Jayanth Koushik, Aarti Singh, and Barnabas Poczos. Hypothesis transfer learning via transformation functions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 574–584. Curran Associates, Inc., 2017.

[16] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[17] Robin Geyer, Luca Corinzia, and Viktor Wegmayr. Transfer learning by adaptive merging of multiple models. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 185–196, London, United Kingdom, 08–10 Jul 2019. PMLR.

[18] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 9871–9881. Curran Associates, Inc., 2019.

[19] Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

[20] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8246–8256. Curran Associates, Inc., 2018.

[21] Wataru Kumagai. Learning bound for parameter transfer learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2729–2737, 2016.

[22] Wataru Kumagai and Takafumi Kanamori. Risk bound of transfer learning using parametric feature mapping and its application to sparse coding. *Machine Learning*, 108(11), 2019.

[23] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 942–950, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[24] Ilja Kuzborskij and Francesco Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2):171–195, 2017.

[25] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[26] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1041–1048. Curran Associates, Inc., 2009.

[27] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 367–374, Arlington, Virginia, USA, 2009. AUAI Press.

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

[29] Rajat Raina, Andrew Y. Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 713–720, New York, NY, USA, 2006. Association for Computing Machinery.

[30] Sashank Jakkam Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[31] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1433–1440, USA, 2007. Curran Associates Inc.

[32] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33, 2020.

[33] Heidi Williams. Applying statistical language recognition techniques in the ciphertext-only cryptanalysis of enigma. *Cryptologia*, 24(1):4–17, 2000.

[34] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.

A Proof of Theorem 3

Proof. To prove Theorem 3, it suffices to provide an algorithm π given any π_Z, such that equation (1) holds true. We construct π as follows.

Recall that Theorem 3 assumes that Z is randomly generated independent of the problem scenario. Thus, its distribution is identical for each scenario (P_i, l_i). Let Z' be another variable randomly generated with the same distribution. We define $\pi(\{x_i, y_i\}_{i=1}^n, l) = \pi_{Z'}(\{x_i, y_i\}_{i=1}^n, l)$.[3]

For any fixed (P_i, l_i), note that the hypothesis returned by π has the same distribution of that returned by the version π_Z dependent on Z. Consequently, their resulting risk functions also have the same distribution, i.e.,

$$\begin{aligned}&\mathrm{Dist}_{P_i}[R(\pi_Z(\{(x_i, y_i)\}_{i=1}^n, l))]\\ &= \mathrm{Dist}_{P_i}[R(\pi(\{(x_i, y_i)\}_{i=1}^n, l))].\end{aligned} \tag{7}$$

Recall that $\epsilon(h) - R(h)$ is a quantity that only depends on (P_i, l_i). Hence, when (P_i, l_i) is fixed, we have

$$\begin{aligned}&\mathrm{Dist}_{P_i}[\epsilon(\pi_Z(\{(x_i, y_i)\}_{i=1}^n, l))]\\ &= \mathrm{Dist}_{P_i}[\epsilon(\pi(\{(x_i, y_i)\}_{i=1}^n, l))].\end{aligned}$$

Taking the mixture over any weight function $\{w_i\}_{i=1}^k$,

$$\begin{aligned}&\sum_i w_i \mathrm{Dist}_{P_i}[\epsilon(\pi_Z(\{(x_i, y_i)\}_{i=1}^n, l))]\\ &= \sum_i w_i \mathrm{Dist}_{P_i}[\epsilon(\pi(\{(x_i, y_i)\}_{i=1}^n, l))].\end{aligned}$$

□

B Proof of Lemma 3

We prove Lemma 3 by induction. For simplicity, let

$$f(n, k) \triangleq \frac{\sqrt{n}}{\sqrt{8k(n-k)}} (\frac{k}{n})^{-k} (1 - \frac{k}{n})^{-(n-k)}. \tag{8}$$

It suffices to show that: (a). $\binom{n}{k} \geq f(n, k)$ for $k = n - k = 1$, (b). $\binom{n}{k} / \binom{n-1}{k-1} \geq f(n, k)/f(n-1, k-1)$, (c). $\binom{n}{k} / \binom{n-1}{k} \geq f(n, k)/f(n-1, k)$.

Part (a) can be proved by simple verification. Part (b) can be proved by taking the natural log, and the rest follows from convexity properties of log functions. Explicitly, the inequality needs to be proved in this induction step is

[3] If $\pi_{Z'}$ is also randomized, let π returns a variable with the same distribution.

as follows.

$$\log \frac{n}{k} \geq \log \frac{\sqrt{n}}{\sqrt{8k(n-k)}} (\frac{k}{n})^{-k} (1-\frac{k}{n})^{-(n-k)} \\ - \log \frac{\sqrt{n-1}}{\sqrt{8(k-1)(n-k)}} (\frac{k-1}{n-1})^{-(k-1)} (1-\frac{k-1}{n-1})^{-(n-k)} \tag{9}$$

The RHS of the above inequality can be simplified through the following steps.

$$\begin{aligned} \text{RHS} &= \log \frac{n^{n+\frac{1}{2}}}{k^{k+\frac{1}{2}}} - \log \frac{(n-1)^{n-\frac{1}{2}}}{(k-1)^{k-\frac{1}{2}}} \\ &= \log \frac{n}{k} + (n-\frac{1}{2}) \log \frac{n}{n-1} - (k-\frac{1}{2}) \log \frac{k}{k-1}. \end{aligned} \tag{10}$$

Consider a function $f(x) = \log \frac{1+x}{1-x}$, which is convex for $x \in [0,1)$ and is zero at $x = 0$. Hence, $\frac{f(x)}{x}$ is non-decreasing at $x \in [0,1)$. Comparing its values at $x = \frac{1}{2n-1}$ and $\frac{1}{2k-1}$, we have

$$(2n-1) \log \frac{1+\frac{1}{2n-1}}{1-\frac{1}{2n-1}} \leq (2k-1) \log \frac{1+\frac{1}{2k-1}}{1-\frac{1}{2k-1}},$$

which is equivalently

$$(n-\frac{1}{2}) \log \frac{n}{n-1} \leq (k-\frac{1}{2}) \log \frac{k}{k-1}. \tag{11}$$

Combining equation (10) and inequality (11), we have the RHS of inequality (9) is no greater than $\log \frac{n}{k}$, which completes the proof for part (b).

Finally, part (c) is proved by substituting k with $n-k$, and follows from exact proof steps of part (b).