A Theory of Transfer Learning with Applications to Active Learning

Liu Yang[1], Steve Hanneke[2], and Jaime Carbonell[3]

[1] Machine Learning Department, Carnegie Mellon University
liuy@cs.cmu.edu
[2] Department of Statistics, Carnegie Mellon University
shanneke@stat.cmu.edu
[3] Language Technologies Institute, Carnegie Mellon University
jgc@cs.cmu.edu

Abstract. We explore a transfer learning setting, in which a finite sequence of target concepts are sampled independently with an unknown distribution from a known family. We study the total number of labeled examples required to learn all targets to an arbitrary specified expected accuracy, focusing on the asymptotics in the number of tasks and the desired accuracy. Our primary interest is formally understanding the fundamental benefits of transfer learning, compared to learning each target independently from the others. Our approach to the transfer problem is general, in the sense that it can be used with a variety of learning protocols. As a particularly interesting application, we study in detail the benefits of transfer for self-verifying active learning; in this setting, we find that the number of labeled examples required for learning with transfer is often significantly smaller than that required for learning each target independently.

1 Introduction

Transfer learning reuses knowledge from past related tasks to ease the process of learning to perform a new task. The goal of transfer learning is to leverage previous learning and experience to more efficiently learn novel, but related, concepts, compared to what would be possible without this prior experience. The utility of transfer learning is typically measured by a reduction in the number of training examples required to achieve a target performance on a sequence of related learning problems, compared to the number required for unrelated problems: i.e., reduced sample complexity. In many real-life scenarios, just a few training examples of a new concept or process is often sufficient for a human learner to grasp the new concept given knowledge of related ones. For example, learning to drive a van becomes much easier a task if we have already learned how to drive a car. Learning French is somewhat easier if we have already learned English (vs Chinese), and learning Spanish is easier if we know Portuguese (vs German). We are therefore interested in understanding the conditions that enable a learning machine to leverage abstract knowledge obtained as a by-product of learning past concepts, to improve its performance on future learning problems. Furthermore, we are interested in how the magnitude of these improvements grows as the learning system gains more experience from learning multiple related concepts.

The ability to transfer knowledge gained from previous tasks to make it easier to learn a new task can potentially benefit a wide range of real-world applications, including computer vision, natural language processing, cognitive science (e.g., fMRI brain state classification), and speech recognition, to name a few. As an example, consider training a speech recognizer. After training on a number of individuals, a learning system can identify common patterns of speech, such as accents or dialects, each of which requires a slightly different speech recognizer; then, given a new person to train a recognizer for, it can quickly determine the particular dialect from only a few well-chosen examples, and use the previously-learned recognizer for that particular dialect. In this case, we can think of the transferred knowledge as consisting of the common aspects of each recognizer variant and more generally the *distribution* of speech patterns existing in the population these subjects are from. This same type of distribution-related knowledge transfer can be helpful in a host of applications, including all those mentioned above.

Supposing these target concepts (e.g., speech patterns) are sampled independently from a fixed population, having knowledge of the distribution of concepts in the population may often be quite valuable. More generally, we may consider a general scenario in which the target concepts are sampled i.i.d. according to a fixed distribution. As we show below, the number of labeled examples required to learn a target concept sampled according to this distribution may be dramatically reduced if we have direct knowledge of the distribution. However, since in many real-world learning scenarios, we do not have direct access to this distribution, it is desirable to be able to somehow *learn* the distribution, based on observations from a sequence of learning problems with target concepts sampled according to that distribution. The hope is that an estimate of the distribution so-obtained might be almost as useful as direct access to the true distribution in reducing the number of labeled examples required to learn subsequent target concepts. The focus of this paper is an approach to transfer learning based on estimating the distribution of the target concepts. Whereas we acknowledge that there are other important challenges in transfer learning, such as exploring improvements obtainable from transfer under various alternative notions of task relatedness [EP04, BDS03], or alternative reuses of knowledge obtained from previous tasks [Thr96], we believe that learning the distribution of target concepts is a central and crucial component in many transfer learning scenarios, and can reduce the total sample complexity across tasks.

Note that it is not immediately obvious that the distribution of targets can even be learned in this context, since we do not have direct access to the target concepts sampled according to it, but rather have only indirect access via a finite number of labeled examples for each task; a significant part of the present work focuses on establishing that as long as these finite labeled samples are larger than a certain size, they hold sufficient information about the distribution over concepts for estimation to be possible. In particular, in contrast to standard results on consistent density estimation, our estimators are not directly based on the target concepts, but rather are only indirectly dependent on these via the labels of a finite number of data points from each task. One desideratum we pay particular attention to is minimizing the number of *extra* labeled examples needed for each task, beyond what is needed for learning that particular target, so that the benefits of transfer learning are obtained almost as a *by-product* of learning the targets. Our technique is general, in that it applies to any concept space with finite VC dimension; also, the process of learning the target concepts is (in some sense) decoupled from the mechanism of learning the concept distribution, so that we may apply our technique to a variety of learning protocols, including passive supervised learning, active supervised learning, semi-supervised learning, and learning with certain general data-dependent forms of interaction [Han09]. For simplicity, we choose to formulate our transfer learning algorithms in the language of active learning; as we show, this problem can benefit significantly from transfer. Formulations for other learning protocols would follow along similar lines, with analogous theorems; only the results in Section 5 are specific to active learning.

Transfer learning is related at least in spirit to much earlier work on case-based and analogical learning [Car83, Car86, VC93, Kol93, Thr96], although that body of work predated modern machine learning, and focused on symbolic reuse of past problem solving solutions rather than on current machine learning problems such as classification, regression or structured learning. More recently, transfer learning (and the closely related problem of *multitask* learning) has been studied in specific cases with interesting (though sometimes hueristic) approaches [Car97, Sil00, MP04, Bax97, BDS03]. This paper considers a general theoretical framework for transfer learning, based on an Empirical Bayes perspective, and derives rigorous theoretical results on the benefits of transfer. We discuss the relation of this analysis to existing theoretical work on transfer learning below.

1.1 Active Learning

Active learning is a powerful form of supervised machine learning characterized by interaction between the learning algorithm and data source during the learning process [CAL94,MN98,CCS00,TK01,NS04,BEYL04,DCB07,DC08,HC08]. In this work, we consider a variant known as *pool-based* active learning, in which a learning algorithm is given access to a (typically very large) collection of unlabeled examples, and is able to select any of those examples, request the supervisor to label it (in agreement with the target concept), then after receiving the label, select another example from the pool, etc. This sequential label-requesting process continues until some halting criterion is reached, at which point the algorithm outputs a classifier, and the objective is for this classifier to closely approximate the (unknown) target concept in the future. The primary motivation behind pool-based active learning is that, often, unlabeled examples are inexpensive and available in abundance, while annotating those examples can be costly or time-consuming; as such, we often wish to select only the informative examples to be labeled, thus reducing information-redundancy to some extent, compared to the baseline of selecting the examples to be labeled uniformly at random from the pool (passive learning).

There has recently been an explosion of fascinating theoretical results on the advantages of this type of active learning, compared to passive learning, in terms of the number of labels required to obtain a prescribed accuracy (called the *sample complexity*): e.g., [FSST97, Das04, DKM09, Das05, Han07b, BHV10, BBL09, Wan09, Kää06, Han07a, DHM08, Fri09, CN08, Now08, BBZ07, Han11, Kol10, Han09, BDL09]. In particular, [BHV10] show that in noise-free binary classifier learning, for any passive learning algorithm for a concept space of finite VC dimension, there exists an active learning algorithm with asymptotically much smaller sample complexity for any nontrivial target concept. Thus, it appears there are profound advantages to active learning compared to passive learning. In later work, [Han09] strengthens this result by removing a certain dependence on the underlying distribution of the data in the learning algorithm.

However, the ability to rapidly converge to a good classifier using only a small number of labels is only one desirable quality of a machine learning method, and there are other qualities that may also be important in certain scenarios. In particular, the ability to *verify* the performance of a learning method is often a crucial part of machine learning applications, as (among other things) it helps us determine whether we have enough data to achieve a desired level of accuracy with the given method. In passive learning, one common practice for this verification is to hold out a random sample of labeled examples as a *validation sample* to evaluate the trained classifier (e.g., to determine when training is complete). It turns out this technique is not feasiable in active learning, since in order to be really useful as an indicator of whether we have seen enough labels to guarantee the desired accuracy, the number of labeled examples in the random validation sample would need to be much larger than the number of labels requested by the active learning algorithm itself, thus (to some extent) canceling the savings obtained by performing active rather than passive learning. Another common practice in passive learning is to examine the training error rate of the returned classifier, which can serve as a reasonable indicator of performance (after adjusting for model complexity). However, again this measure of performance is not necessarily reasonable for active learning, since the set of examples the algorithm requests the labels of is typically distributed very differently from the test examples the classifier will be applied to after training.

This reasoning seems to indicate that performance verification is (at best) a far more subtle issue in active learning than in passive learning. Indeed, [BHV10] note that although the number of labels required to achieve good accuracy in active learning is significantly smaller than passive learning, it is sometimes the case that the number of labels required to *verify* that the accuracy is good is not significantly improved. In particular, this phenomenon can increase significantly the sample complexity of active learning algorithms that adaptively determine how many labels to request before terminating. In short, if we require the algorithm both to *learn* an accurate concept and to *know* that its concept is accurate, then the number of labels required by active learning may sometimes not be significantly smaller than the number required by passive learning.

In the present work, we are interested in the question of whether a form of transfer learning can help to bridge this gap, enabling self-verifying active learning algorithms to obtain the same types of dramatic improvements over passive learning as can be achieved by their non-self-verifying counterparts.

1.2 Outline of the paper

The remainder of the paper is organized as follows. In Section 2 we introduce basic notation used throughout, and survey some related work from the existing literature. In Section 3, we describe and analyze our proposed method for estimating the distribution of target concepts, the key ingrediant in our approach to transfer learning, which we then present in Section 4. Finally, in Section 5, we investigate the benefits of this type of transfer learning for self-verifying active learning.

2 Definitions and Related Work

First, we state a few basic notational conventions. We denote $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any random variable X, we generally denote by $\mathbb{P}_X$ the distribution of X (the induced probability measure on the range of X), and by $\mathbb{P}_{X|Y}$ the regular conditional distribution of X given Y. For any pair of probability measures μ_1, μ_2 on a measurable space $(\Omega, \mathcal{F})$, we define

$$\|\mu_1 - \mu_2\| = \sup_{A \in \mathcal{F}} |\mu_1(A) - \mu_2(A)|.$$

Next we define the particular objects of interest to our present discussion. Let Θ be an arbitrary set (called the *parameter space*), $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a Borel space [Sch95] (where $\mathcal{X}$ is called the *instance space*), and $\mathcal{D}$ be a fixed distribution on $\mathcal{X}$ (called the *data distribution*). For instance, Θ could be $\mathbb{R}^n$ and $\mathcal{X}$ could be $\mathbb{R}^m$, for some $n, m \in \mathbb{N}$, though more general scenarios are certainly possible as well, including infinite-dimensional parameter spaces. Let $\mathbb{C}$ be a set of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$ (called the *concept space*), and suppose $\mathbb{C}$ has VC dimension $d < \infty$ [Vap82] (such a space is called a *VC class*). $\mathbb{C}$ is equipped with its Borel σ-algebra $\mathcal{B}$, induced by the pseudo-metric $\rho(h, g) = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq g(x)\})$. Though all of our results can be formulated for general $\mathcal{D}$ in slightly more complex terms, for simplicity throughout the discussion below we suppose ρ is actually a *metric*, in that any $h, g \in \mathbb{C}$ with $h \neq g$ have $\rho(h, g) > 0$; this amounts to a topological assumption on $\mathbb{C}$ relative to $\mathcal{D}$.

For each $\theta \in \Theta$, π_θ is a distribution on $\mathbb{C}$ (called a *prior*). Our only (rather mild) assumption on this family of prior distributions is that $\{\pi_\theta : \theta \in \Theta\}$ be totally bounded, in the sense that $\forall \varepsilon > 0$, $\exists$ *finite* $\Theta_\varepsilon \subseteq \Theta$ s.t. $\forall \theta \in \Theta, \exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\pi_\theta - \pi_{\theta_\varepsilon}\| < \varepsilon$. See [DL01] for examples of categories of classes that satisfy this.

The general setup for the learning problem is that we have a *true* parameter value $\theta_\star \in \Theta$, and a collection of $\mathbb{C}$-valued random variables $\{h^*_{t\theta}\}_{t \in \mathbb{N}, \theta \in \Theta}$, where for a fixed $\theta \in \Theta$ the $\{h^*_{t\theta}\}_{t \in \mathbb{N}}$ variables are i.i.d. with distribution π_θ.

The learning problem is the following. For each $\theta \in \Theta$, there is a sequence

$$\mathcal{Z}_t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \ldots\},$$

where $\{X_{ti}\}_{t,i \in \mathbb{N}}$ are i.i.d. $\mathcal{D}$, and for each $t, i \in \mathbb{N}$, $Y_{ti}(\theta) = h^*_{t\theta}(X_{ti})$. For $k \in \mathbb{N}$ we denote by $\mathcal{Z}_{tk}(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$.

The algorithm receives values ε and T as input, and for each $t \in \{1, 2, \ldots, T\}$ in increasing order, it observes the sequence $X_{t1}, X_{t2}, \ldots$, and may then select an index i_1, receive label $Y_{ti_1}(\theta_\star)$, select another index i_2, receive label $Y_{ti_2}(\theta_\star)$, etc. The algorithm proceeds in this fashion, sequentially requesting labels, until eventually it produces a classifier $\hat{h}_t$. It then increments t and repeats this process until it produces a sequence $\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_T$, at which time it halts. To be called *correct*, the algorithm must have a guarantee that $\forall \theta_\star \in \Theta, \forall t \leq T, \mathbb{E}\left[\rho\left(\hat{h}, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$. We will be interested in the expected number of label requests necessary for a correct learning algorithm, averaged over the T tasks, and in particular in how shared information between tasks can help to reduce this quantity when direct access to $\theta_\star$ is not available to the algorithm.

2.1 Relation to Existing Theoretical Work on Transfer Learning

Although we know of no existing work on the theoretical advantages of transfer learning for active learning, the existing literature contains several analyses of the advantages of transfer learning for passive learning. In his classic work, Baxter [Bax97] explores a similar setup for a general form of passive learning, except in a *full* Bayesian setting (in contrast to our setting, often referred to as "empirical Bayes," which includes a constant parameter $\theta_\star$ to be estimated from data). Essentially, [Bax97] sets up a hierarchical Bayesian model, in which (in our notation) $\theta_\star$ is a random variable with known distribution (hyper-prior), but otherwise the specialization of Baxter's setting to the pattern recognition problem is essentially identical to our setup above. This hyper-prior does make the problem slightly easier, but generally the results of [Bax97] are of a different nature than our objectives here. Specifically, Baxter's results on learning from labeled examples can be interpretted as indicating that transfer learning can improve certain *constant factors* in the asymptotic rate of convergence of the average of expected error rates across the learning problems. That is, certain constant complexity terms (for instance, related to the concept space) can be reduced to (potentially much smaller) values related to $\pi_{\theta_\star}$ by transfer learning. Baxter argues that, as the number of tasks grows large, this effectively achieves close to the known results on the sample complexity of passive learning with direct access to $\theta_\star$. A similar claim is discussed by Ando and Zhang [AZ04] (though in less detail and formality) for a setting closer to that studied here, where $\theta_\star$ is an unknown parameter to be estimated.

There are also several results on transfer learning of a slightly different variety, in which, rather than having a prior distribution for the target concept, the learner initially has several potential concept spaces to choose from, and the role of transfer is to help the learner select from among these concept spaces [Bax00, AZ04]. In this case, the idea is that one of these concept spaces has the best average minimum achievable error rate per learning problem, and the objective of transfer learning is to perform nearly as well as if we knew which of the spaces has this property. In particular, if we assume the target functions for each task all reside in one of the concept spaces, then the objective of transfer learning is to perform nearly as well as if we knew which of the spaces contains the targets. Thus, transfer learning results in a sample complexity related to the number of learning problems, a complexity term for this best concept space, and a complexity term related to the diversity of concept spaces we have to choose from. In particular, as with [Bax97], these results can typically be interpretted as giving constant factor improvements from transfer in a passive learning context, at best reducing the complexity constants, from those for the union over the given concept spaces, down to the complexity constants of the single best concept space.

In addition to the above works, there are several analyses of transfer learning and multitask learning of an entirely different nature than our present discussion, in that the objectives of the analysis are somewhat different. Specifically, there is a branch of the literature concerned with task *relatedness*, not in terms of the underlying process that generates the target concepts, but rather directly in terms of relations between the target concepts themselves. In this sense, several tasks with related target concepts should be much easier to learn than tasks with unrelated target concepts. This is studied in the context of kernel methods by [MP04, EP04, EMP05], and in a more general theoretical framework by [BDS03]. As mentioned, our approach to transfer learning is based on the idea of estimating the distribution of target concepts. As such, though interesting and important, these notions of direct relatedness of target concepts are not as relevant to our present discussion.

As with [Bax97], the present work is interested in showing that as the number of tasks grows large, we can effectively achieve a sample complexity close to that achieveable with direct access to $\theta_\star$. However, in contrast, we are interested in a general approach to transfer learning and the analysis thereof, leading to concrete results for a variety of learning protocols such as active learning and semi-supervised learning. In particular, our analysis of active learning reveals the interesting phenomenon that transfer learning can sometimes improve the asymptotic dependence on ε, rather than merely the constant factors as in the analysis of [Bax97].

Additionally, unlike [Bax97], we study the benefits of transfer learning in terms of the asymptotics as the number of learning problems grows large, *without* necessarily requiring the number of labeled examples per learning problem to also grow large. That is, our analysis reveals benefits from transfer learning even if the number of labeled examples per learning problem is *bounded*. This is desirable for the following practical reasons. In many settings where transfer learning may be useful, it is desirable that the number of labeled examples we need to collect from each particular learning problem never be significantly larger than the number of such examples required to solve that particular problem (i.e., to learn that target concept to the desired accuracy). For instance, this is the case when the learning problems are not all solved by the same individual (or company, etc.), but rather a coalition of cooperating individuals (e.g., hospitals sharing data on clinical trials); each individual may be willing to share the data they used to learn their problem, in the interest of making others' learning problems easier; however, they may not be willing to collect significantly *more* data to advance this cause than they themselves need for their own learning problem. Given a desired error rate ε for each learning problem, the number of labeled examples required to learn each particular target concept to this desired error rate is always bounded by an ε-dependent value. Therefore, an analysis that requires a growing number of examples per learning problem seems undesirable in these scenarios, since for some of the problems we would need to label a number of examples far beyond what is needed to learn a good classifier for that particular problem. We should therefore be particularly interested in studying transfer as a *by-product* of the usual learning process; failing this, we are interested in the minimum possible number of *extra* labeled examples per task to gain the benefits of transfer learning. To our knowledge, no result of this type (bounded sample size per learning problem) has yet been established at the level of generality studied here.

3 Estimating the Prior

The advantage of transfer learning in this setting is that each learning problem provides some information about $\theta_\star$, so that after solving several of the learning problems, we might hope to be able to *estimate* $\theta_\star$. Then, with this estimate in hand, we can use the corresponding estimated prior distribution in the learning algorithm for subsequent learning problems, to help inform the learning process similarly to how direct knowledge of $\theta_\star$ might be helpful. However, the difficulty in approaching this is how to define such an estimator. Since we do not have direct access to the h_t^* values, but rather only indirect observations via a finite number of example labels, the standard results for density estimation from i.i.d. samples cannot be applied.

The idea we pursue below is to consider the distributions on $\mathcal{Z}_{tk}(\theta_\star)$. These variables *are* directly observable, by requesting the labels of those examples. Thus, for any finite $k \in \mathbb{N}$, this distribution *is* estimable from observable data. That is, using the i.i.d. values $\mathcal{Z}_{1k}(\theta_\star), \ldots, \mathcal{Z}_{tk}(\theta_\star)$, we can apply standard techniques for density estimation to arrive at an estimator of $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$. Then the question is whether the distribution $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ uniquely characterizes the prior distribution $\pi_{\theta_\star}$: that is, whether $\pi_{\theta_\star}$ is *identifiable* from $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$.

As an example, consider the space of *half-open interval* classifiers on $[0,1]$: $\mathbb{C} = \{\mathbb{1}^{\pm}_{[a,b)} : 0 \leq a \leq b \leq 1\}$, where $\mathbb{1}^{\pm}_{[a,b)}(x) = +1$ if $a \leq x < b$ and -1 otherwise. In this case, $\pi_{\theta_\star}$ is *not* necessarily identifiable from $\mathbb{P}_{\mathcal{Z}_{t1}(\theta_\star)}$; for instance, the distributions π_{θ_1} and π_{θ_2} characterized by $\pi_{\theta_1}(\{\mathbb{1}^{\pm}_{[0,1)}\}) = \pi_{\theta_1}(\{\mathbb{1}^{\pm}_{\emptyset}\}) = 1/2$ and $\pi_{\theta_2}(\{\mathbb{1}^{\pm}_{[0,1/2)}\}) = \pi_{\theta_2}(\{\mathbb{1}^{\pm}_{[1/2,1)}\}) = 1/2$ are not distinguished by these one-dimensional distributions. However, it turns out that for this half-open intervals problem, $\pi_{\theta_\star}$ *is* uniquely identifiable from $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_\star)}$; for instance, in the θ_1 vs θ_2 scenario, the conditional probability $\mathbb{P}_{(Y_{t1}(\theta_i),Y_{t2}(\theta_i))|(X_{t1},X_{t2})}((+1,+1)|(1/4,3/4))$ will distinguish π_{θ_1} from π_{θ_2}, and this can be calculated from $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_i)}$. The crucial element of the analysis below is determining the appropriate value of k to uniquely identify $\pi_{\theta_\star}$ from $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ *in general*. As we will see, $k = d$ is *always* sufficient, a key insight for the results that follow.

To be specific, in order to transfer knowledge from one task to the next, we use a few labeled data points from each task to gain information about $\theta_\star$. For this, for each task t, we simply take the first d data points in the $\mathcal{Z}_t(\theta_\star)$ sequence. That is, we request the labels

$$Y_{t1}(\theta_\star), Y_{t2}(\theta_\star), \ldots, Y_{td}(\theta_\star)$$

and use the points $\mathcal{Z}_{td}(\theta_\star)$ to update an estimate of $\theta_\star$.

The following result shows that this technique does provide a consistent estimator of $\pi_{\theta_\star}$. Again, note that this result is not a straightforward application of the standard approach to consistent estimation, since the observations here are not the $h^*_{t\theta_\star}$ variables themselves, but rather a number of the $Y_{ti}(\theta_\star)$ values. The key insight in this result is that $\pi_{\theta_\star}$ is *uniquely identified* by the joint distribution $\mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}$ over the first d labeled examples; later, we prove this is *not* necessarily true for $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ for values $k < d$.

Theorem 1. *There exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, and functions $R : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ and $\delta : \mathbb{N}_0 \times (0,1] \to [0,1]$, such that for any $\alpha > 0$,*

$\lim_{T \to \infty} R(T, \alpha) = \lim_{T \to \infty} \delta(T, \alpha) = 0$ *and for any* $T \in \mathbb{N}_0$ *and* $\theta_\star \in \Theta$*,*

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T, \alpha) \right) \leq \delta(T, \alpha) \leq \alpha.$$

One important detail to note, for our purposes, is that $R(T, \alpha)$ is independent from $\theta_\star$, so that the value of $R(T, \alpha)$ can be calculated and used within a learning algorithm. The proof of Theorem 1 will be established via the following sequence of lemmas. Lemma 1 relates distances in the space of priors to distances in the space of distributions on the full data sets. In turn, Lemma 2 relates these distances to distances in the space of distributions on a finite number of examples from the data sets. Lemma 3 then relates the distances between distributions on any finite number of examples to distances between distributions on d examples. Finally, Lemma 4 presents a standard result on the existence of a converging estimator, in this case for the distribution on d examples, for totally bounded families of distributions. Tracing these relations back, they relate convergence of the estimator for the distribution of d examples to convergence of the corresponding estimator for the prior itself.

Lemma 1. *For any* $\theta, \theta' \in \Theta$ *and* $t \in \mathbb{N}$*,*

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.$$

Proof. Fix $\theta, \theta' \in \Theta$, $t \in \mathbb{N}$. Let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$, $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$. and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$. For $h \in \mathbb{C}$, let $c_{\mathbb{X}}(h) = \{(X_{t1}, h(X_{t1})), (X_{t2}, h(X_{t2})), \ldots\}$.

For $h, g \in \mathbb{C}$, define $\rho_{\mathbb{X}}(h, g) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^m \mathbb{1}[h(X_{ti}) \neq g(X_{ti})]$ (if the limit exists), and $\rho_{\mathbb{X}_k}(h, g) = \frac{1}{k} \sum_{i=1}^k \mathbb{1}[h(X_{ti}) \neq g(X_{ti})]$. Note that since $\mathbb{C}$ has finite VC dimension, so does the collection of sets $\{\{x : h(x) \neq g(x)\} : h, g \in \mathbb{C}\}$, so that the uniform strong law of large numbers implies that with probability one, $\forall h, g \in \mathbb{C}$, $\rho_{\mathbb{X}}(h, g)$ exists and has $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ [Vap82].

Consider any $\theta, \theta' \in \Theta$, and any $A \in \mathcal{B}$. Then since $\mathcal{B}$ is the Borel σ-algebra induced by ρ, any $h \notin A$ has $\forall g \in A$, $\rho(h, g) > 0$. Thus, if $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ for all $h, g \in \mathbb{C}$, then $\forall h \notin A$,

$$\forall g \in A, \rho_{\mathbb{X}}(h, g) = \rho(h, g) > 0 \implies \forall g \in A, c_{\mathbb{X}}(h) \neq c_{\mathbb{X}}(g) \implies c_{\mathbb{X}}(h) \notin c_{\mathbb{X}}(A).$$

This implies $c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A)) = A$. Under these conditions,

$$\mathbb{P}_{\mathcal{Z}_t(\theta)|\mathbb{X}}(c_{\mathbb{X}}(A)) = \pi_\theta(c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A))) = \pi_\theta(A),$$

and similarly for θ'.

Any measurable set C for the range of $\mathcal{Z}_t(\theta)$ can be expressed as $C = \{c_{\bar{x}}(h) : (h, \bar{x}) \in C'\}$ for some appropriate $C' \in \mathcal{B} \otimes \mathcal{B}_{\mathcal{X}}^\infty$. Letting $C'_{\bar{x}} = \{h : (h, \bar{x}) \in C'\}$, we have

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(C) = \int \pi_\theta(c_{\bar{x}}^{-1}(c_{\bar{x}}(C'_{\bar{x}})))\mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \int \pi_\theta(C'_{\bar{x}})\mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}). = \mathbb{P}_{(h^*_{t\theta}, \mathbb{X})}(C').$$

Likewise, this reasoning holds for θ'. Then

$$\begin{aligned}\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| &= \|\mathbb{P}_{(h^*_{t\theta},\mathbb{X})} - \mathbb{P}_{(h^*_{t\theta'},\mathbb{X})}\| \\ &= \sup_{C' \in \mathcal{B}\otimes\mathcal{B}^\infty_{\mathcal{X}}} \left| \int (\pi_\theta(C'_{\bar{x}}) - \pi_{\theta'}(C'_{\bar{x}}))\mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) \right| \\ &\leq \int \sup_{A\in\mathcal{B}} |\pi_\theta(A) - \pi_{\theta'}(A)| \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \|\pi_\theta - \pi_{\theta'}\|.\end{aligned}$$

Since we also have

$$\begin{aligned}\|\pi_\theta - \pi_{\theta'}\| &= \|\mathbb{P}_{(h^*_{t\theta},\mathbb{X})}(\cdot\times\mathcal{X}^\infty) - \mathbb{P}_{(h^*_{t\theta'},\mathbb{X})}(\cdot\times\mathcal{X}^\infty)\| \\ &\leq \|\mathbb{P}_{(h^*_{t\theta},\mathbb{X})} - \mathbb{P}_{(h^*_{t\theta'},\mathbb{X})}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|,\end{aligned}$$

this means $\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|$. □

Lemma 2. *There exists a sequence $r_k = o(1)$ such that $\forall t, k \in \mathbb{N}$, $\forall \theta, \theta' \in \Theta$,*

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k.$$

Proof. The left inequality follows from Lemma 1 and the basic definition of $\|\cdot\|$, since $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_t(\theta)}(\cdot\times(\mathcal{X}\times\{-1,+1\})^\infty)$, so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|.$$

The remainder of this proof focuses on the right inequality. Fix $\theta, \theta' \in \Theta$, let $\gamma > 0$, and let $B \subseteq (\mathcal{X}\times\{-1,+1\})^\infty$ be a measurable set such that

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| < \mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) + \gamma.$$

Let $\mathcal{A}$ be the collection of all measurable subsets of $(\mathcal{X}\times\{-1,+1\})^\infty$ representable in the form $A'\times(\mathcal{X}\times\{-1,+1\})^\infty$, for some measurable $A' \subseteq (\mathcal{X}\times\{-1,+1\})^k$ and some $k \in \mathbb{N}$. In particular, since $\mathcal{A}$ is an algebra that generates the product σ-algebra, Carathéodory's extention theorem [Sch95] implies that there exist disjoint sets $\{A_i\}_{i\in\mathbb{N}}$ in $\mathcal{A}$ such that $B \subseteq \bigcup_{i\in\mathbb{N}} A_i$ and

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) < \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) + \gamma.$$

Additionally, as these sums are bounded, there must exist $n \in \mathbb{N}$ such that

$$\sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) < \gamma + \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i),$$

so that

$$\begin{aligned}\sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) &< \gamma + \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) \\ &= \gamma + \mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^n A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^n A_i\right).\end{aligned}$$

As $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, there exists $k' \in \mathbb{N}$ and measurable $A' \subseteq (\mathcal{X} \times \{-1,+1\})^{k'}$ such that $\bigcup_{i=1}^{n} A_i = A' \times (\mathcal{X} \times \{-1,+1\})^{\infty}$, and therefore

$$\begin{aligned}\mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}_{\mathcal{Z}_{tk'}(\theta)}(A') - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}(A') \\ &\leq \|\mathbb{P}_{\mathcal{Z}_{tk'}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.\end{aligned}$$

In summary, we have $\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + 3\gamma$. Since this is true for an arbitrary $\gamma > 0$, taking the limit as $\gamma \to 0$ implies

$$\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

In particular, this implies there exists a sequence $r_k(\theta, \theta') = o(1)$ such that

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k(\theta, \theta').$$

This would suffice to establish the upper bound if we were allowing r_k to depend on the particular θ and θ'. However, to guarantee the same rates of convergence for all pairs of parameters requires an additional argument. Specifically, let $\gamma > 0$ and let Θ_γ denote a minimal subset of Θ such that, $\forall \theta \in \Theta$, $\exists \theta_\gamma \in \Theta_\gamma$ s.t. $\|\pi_\theta - \pi_{\theta_\gamma}\| < \gamma$: that is, a minimal γ-cover. Since $|\Theta_\gamma| < \infty$ by assumption, defining $r_k(\gamma) = \max_{\theta,\theta' \in \Theta_\gamma} r_k(\theta, \theta')$, we have $r_k(\gamma) = o(1)$. Furthermore, for any $\theta, \theta' \in \Theta$, letting $\theta_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_\theta - \pi_{\theta''}\|$ and $\theta'_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_{\theta'} - \pi_{\theta''}\|$, we have (by triangle inequalities)

$$\begin{aligned}\|\pi_\theta - \pi_{\theta'}\| &\leq \|\pi_\theta - \pi_{\theta_\gamma}\| + \|\pi_{\theta_\gamma} - \pi_{\theta'_\gamma}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\| \\ &< 2\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|.\end{aligned}$$

By triangle inequalities and the left inequality from the lemma statement (established above), we also have

$$\begin{aligned}&\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\| \\ &\leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta')} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\| \\ &\leq \|\pi_{\theta_\gamma} - \pi_\theta\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\pi_{\theta'} - \pi_{\theta'_\gamma}\| \\ &< 2\gamma + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.\end{aligned}$$

Defining $r_k = \inf_{\gamma>0} (4\gamma + r_k(\gamma))$, we have the right inequality of the lemma statement, and since $r_k(\gamma) = o(1)$ for each $\gamma > 0$, we have $r_k = o(1)$. □

Lemma 3. *$\forall t, k \in \mathbb{N}$, $\forall \theta, \theta' \in \Theta$,*

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq 4 \cdot 2^{2k+d} k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|}.$$

Proof. Fix any $t \in \mathbb{N}$, and let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$ and $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$ and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$.

If $k \leq d$, then $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_{td}(\theta)}(\cdot \times (\mathcal{X} \times \{-1,+1\})^{d-k})$, so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|,$$

and therefore the result trivially holds.

Now suppose $k > d$. For a sequence $\bar{z}$ and $I \subseteq \mathbb{N}$, we will use the notation $\bar{z}_I = \{\bar{z}_i : i \in I\}$. Note that, for any $k > d$ and $\bar{x}^k \in \mathcal{X}^k$, there is a sequence $\bar{y}(\bar{x}^k) \in \{-1,+1\}^k$ such that no $h \in \mathbb{C}$ has $h(\bar{x}^k) = \bar{y}(\bar{x}^k)$ (i.e., $\forall h \in \mathbb{C}$, $\exists i \leq k$ s.t. $h(\bar{x}^k_i) \neq \bar{y}_i(\bar{x}^k)$). Now suppose $k > d$ and take as an inductive hypothesis that there is a measurable set $A^* \subseteq \mathcal{X}^\infty$ of probability one with the property that $\forall \bar{x} \in A^*$, for every finite $I \subset \mathbb{N}$ with $|I| > d$, for every $\bar{y} \in \{-1,+1\}^\infty$ with $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 \leq k-1$,

$$\begin{aligned}
&\left|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)\right| \\
&\qquad \leq 2^{k-1} \cdot \max_{\tilde{y}^d \in \{-1,+1\}^d, D \in I^d} \left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D)\right|.
\end{aligned}$$

This clearly holds for $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = 0$, since $\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = 0$ in this case, so this will serve as our base case in the inductive proof. Next we inductively extend this to the value $k > 0$. Specifically, let A^*_{k-1} be the A^* guaranteed to exist by the inductive hypothesis, and fix any $\bar{x} \in A^*$, $\bar{y} \in \{-1,+1\}^\infty$, and finite $I \subset \mathbb{N}$ with $|I| > d$ and $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = k$. Let $i \in I$ be such that $\bar{y}_i \neq \bar{y}_i(\bar{x}_I)$, and let $\bar{y}' \in \{-1,+1\}$ have $\bar{y}'_j = \bar{y}_j$ for every $j \neq i$, and $\bar{y}'_i = -\bar{y}_i$. Then

$$\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I),$$

and similarly for θ'. By the inductive hypothesis, this means

$$\begin{aligned}
&\left|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)\right| \\
&\qquad \leq \left|\mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta')|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}})\right| \\
&\qquad\quad + \left|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I)\right| \\
&\qquad \leq 2^k \cdot \max_{\tilde{y}^d \in \{-1,+1\}^d, D \in I^d} \left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D)\right|.
\end{aligned}$$

Therefore, by the principle of induction, this inequality holds for all $k > d$, for every $\bar{x} \in A^*$, $\bar{y} \in \{-1,+1\}^\infty$, and finite $I \subset \mathbb{N}$, where A^* has $\mathcal{D}^\infty$-probability one.

In particular, we have that for $\theta, \theta' \in \Theta$,

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \\
&\leq 2^k \mathbb{E}\left[\max_{\bar{y}^k \in \{-1,+1\}^k} \left|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k)\right|\right] \\
&\leq 2^{2k} \mathbb{E}\left[\max_{\tilde{y}^d \in \{-1,+1\}^d, D \in \{1,\ldots,k\}^d} \left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D)\right|\right] \\
&\leq 2^{2k} \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D)\right|\right].
\end{aligned}$$

Exchangeability implies this is at most

$$2^{2k} \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right]$$
$$\leq 2^{2k+d} k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right].$$

To complete the proof, we need only bound this value by an appropriate function of $\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|$. Toward this end, suppose

$$\mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right] \geq \varepsilon,$$

for some $\tilde{y}^d$. Then either

$$\mathbb{P}\left(\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4\right) \geq \varepsilon/4,$$

or

$$\mathbb{P}\left(\mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4\right) \geq \varepsilon/4.$$

For which ever is the case, let A_ε denote the corresponding measurable subset of $\mathcal{X}^d$, of probability at least $\varepsilon/4$. Then

$$\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\| \geq \left|\mathbb{P}_{\mathcal{Z}_{td}(\theta)}(A_\varepsilon \times \{\tilde{y}^d\}) - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}(A_\varepsilon \times \{\tilde{y}^d\})\right|$$
$$\geq (\varepsilon/4)\mathbb{P}_{\mathbb{X}_d}(A_\varepsilon) \geq \varepsilon^2/16.$$

Therefore,

$$\mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|},$$

which means

$$2^{2k+d} k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right]$$
$$\leq 4 \cdot 2^{2k+d} k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|}.$$

□

The following lemma is a standard result on the existence of converging density estimators for totally bounded families of distributions. For instance, the *skeleton* estimates described by [Yat85, DL01] satisfy this; in fact, in many contexts (though certainly not all), even a simple maximum likelihood estimator would suffice. The reader is referred to [Yat85, DL01] for a proof of this lemma.

Lemma 4. *[Yat85, DL01] Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a totally bounded family of probability measures on a measurable space $(\Omega, \mathcal{F})$, and let $\{W_t(\theta)\}_{t\in\mathbb{N},\theta\in\Theta}$ be Ω-valued random variables such that $\{W_t(\theta)\}_{t\in\mathbb{N}}$ are i.i.d. p_θ for each $\theta \in \Theta$. Then there exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(W_1(\theta_\star), \ldots, W_T(\theta_\star))$ and functions $R_\mathcal{P} : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ and $\delta_\mathcal{P} : \mathbb{N}_0 \times (0,1] \to [0,1]$ such that $\forall \alpha > 0$, $\lim_{T\to\infty} R_\mathcal{P}(T,\alpha) = \lim_{T\to\infty} \delta_\mathcal{P}(T,\alpha) = 0$, and $\forall \theta_\star \in \Theta$ and $T \in \mathbb{N}_0$,*

$$\mathbb{P}\left(\|p_{\hat{\theta}_{T\theta_\star}} - p_{\theta_\star}\| > R_\mathcal{P}(T,\alpha)\right) \leq \delta_\mathcal{P}(T,\alpha) \leq \alpha.$$

We are now ready for the proof of Theorem 1

Proof (Theorem 1). For $\varepsilon > 0$, let $\Theta_\varepsilon \subseteq \Theta$ be any finite subset such that $\forall \theta \in \Theta$, $\exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$; this exists by the assumption that $\{\pi_\theta : \theta \in \Theta\}$ is totally bounded. Then Lemma 2 implies that $\forall \theta \in \Theta$, $\exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\mathbb{P}_{\mathcal{Z}_{td}(\theta_\varepsilon)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta)}\| \leq \|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$, so that $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta_\varepsilon)} : \theta_\varepsilon \in \Theta_\varepsilon\}$ is a finite ε-cover of $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta)} : \theta \in \Theta\}$. Therefore, $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta)} : \theta \in \Theta\}$ is totally bounded. Lemma 4 then implies that there exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$ and functions $R_d : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ and $\delta_d : \mathbb{N}_0 \times (0,1] \to [0,1]$ such that $\forall \alpha > 0$, $\lim_{T\to\infty} R_d(T,\alpha) = \lim_{T\to\infty} \delta_d(T,\alpha) = 0$, and $\forall \theta_\star \in \Theta$ and $T \in \mathbb{N}_0$,

$$\mathbb{P}\left(\|\mathbb{P}_{\mathcal{Z}_{(T+1)d}(\hat{\theta}_{T\theta_\star})|\hat{\theta}_{T\theta_\star}} - \mathbb{P}_{\mathcal{Z}_{(T+1)d}(\theta_\star)}\| > R_d(T,\alpha)\right) \leq \delta_d(T,\alpha) \leq \alpha. \tag{1}$$

Defining

$$R(T,\alpha) = \min_{k\in\mathbb{N}} \left(r_k + 4 \cdot 2^{2k+d} k^d \sqrt{R_d(T,\alpha)}\right),$$

and $\delta(T,\alpha) = \delta_d(T,\alpha)$, and combining (1) with Lemmas 3 and 2, we have

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T,\alpha)\right) \leq \delta(T,\alpha) \leq \alpha.$$

Finally, note that $\lim\limits_{k\to\infty} r_k = 0$ and $\lim\limits_{T\to\infty} R_d(T,\alpha) = 0$ imply that $\lim\limits_{T\to\infty} R(T,\alpha) = 0$. □

3.1 Identifiability from d Points

Inspection of the above proof reveals that the assumption that the family of priors is totally bounded is required only to establish the estimability and bounded rate guarantees. In particular, the implied identifiability condition is, in fact, *always* satisfied, as stated formally in the following corollary.

Corollary 1. *For any priors π_1, π_2 on $\mathbb{C}$, if $h_i^* \sim \pi_i$, $X_1, \ldots, X_d$ are i.i.d. $\mathcal{D}$ independent from h_i^*, and $Z_d(i) = \{(X_1, h_i^*(X_1)), \ldots, (X_d, h_i^*(X_d))\}$ for $i \in \{1,2\}$, then $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)} \implies \pi_1 = \pi_2$.*

Proof. The described scenario is a special case of our general setting, with $\Theta = \{1,2\}$, in which case $\mathbb{P}_{Z_d(i)} = \mathbb{P}_{\mathcal{Z}_{1d}(i)}$. Thus, if $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)}$, then Lemma 3 and Lemma 2 combine to imply that $\|\pi_1 - \pi_2\| \leq \inf_{k\in\mathbb{N}} r_k = 0$. □

It is natural to wonder whether this identifiability remains true for some smaller number of points $k < d$, so that we might hope to create an estimator for $\pi_{\theta_\star}$ based on an estimator for $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$. However, one can show that d is actually the *minimum* possible value for which this remains true for all $\mathcal{D}$ and all families of priors. Formally, we have the following result, holding for every VC class $\mathbb{C}$.

Theorem 2. *There exists a data distribution $\mathcal{D}$ and priors π_1, π_2 on $\mathbb{C}$ such that, for any positive integer $k < d$, if $h_i^* \sim \pi_i$, $X_1, \ldots, X_k$ are i.i.d. $\mathcal{D}$ independent from h_i^*, and $Z_k(i) = \{(X_1, h_i^*(X_1)), \ldots, (X_k, h_i^*(X_k))\}$ for $i \in \{1,2\}$, then $\mathbb{P}_{Z_k(1)} = \mathbb{P}_{Z_k(2)}$ but $\pi_1 \neq \pi_2$.*

Proof. Note that it suffices to show this is the case for $k = d-1$, since any smaller k is a marginal of this case. Consider a shatterable set of points $S_d = \{x_1, x_2, \ldots, x_d\} \subseteq \mathcal{X}$, and let $\mathcal{D}$ be uniform on S_d. Let $\mathbb{C}[S_d]$ be any 2^d classifiers in $\mathbb{C}$ that shatter S_d. Let π_1 be the uniform distribution on $\mathbb{C}[S]$. Now let $S_{d-1} = \{x_1, \ldots, x_{d-1}\}$ and $\mathbb{C}[S_{d-1}] \subseteq \mathbb{C}[S_d]$ shatter S_{d-1} with the property that $\forall h \in \mathbb{C}[S_{d-1}]$, $h(x_d) = \prod_{j=1}^{d-1} h(x_j)$. Let π_2 be uniform on $\mathbb{C}[S_{d-1}]$. Now for any $k < d$ and distinct indices $t_1, \ldots, t_k \in \{1, \ldots, d\}$, $\{h_i^*(x_{t_1}), \ldots, h_i^*(x_{t_k})\}$ is distributed uniformly in $\{-1, +1\}^k$ for both $i \in \{1, 2\}$. This implies $\mathbb{P}_{Z_{d-1}(1)|X_1,\ldots,X_{d-1}} = \mathbb{P}_{Z_{d-1}(2)|X_1,\ldots,X_{d-1}}$, which implies $\mathbb{P}_{Z_{d-1}(1)} = \mathbb{P}_{Z_{d-1}(2)}$. However, π_1 is clearly different from π_2, since even the sizes of the supports are different. □

4 Transfer Learning

In this section, we look at an application of the techniques from the previous section to transfer learning. Like the previous section, the results in this section are general, in that they are applicable to a variety of learning protocols, including passive supervised learning, passive semi-supervised learning, active learning, and learning with certain general types of data-dependent interaction (see [Han09]). For simplicity, we restrict our discussion to the active learning formulation; the analogous results for these other learning protocols follow by similar reasoning.

The result of the previous section implies that an estimator for $\theta_\star$ based on d-dimensional joint distributions is consistent with a bounded rate of convergence R. Therefore, for certain prior-dependent learning algorithms, their behavior should be similar under $\pi_{\hat{\theta}_{T\theta_\star}}$ to their behavior under $\pi_{\theta_\star}$.

To make this concrete, we formalize this in the active learning protocol as follows. A *prior-dependent* active learning algorithm $\mathcal{A}$ takes as inputs $\varepsilon > 0$, $\mathcal{D}$, and a distribution π on $\mathbb{C}$. It initially has access to $X_1, X_2, \ldots$ i.i.d. $\mathcal{D}$; it then selects an index i_1 to request the label for, receives $Y_{i_1} = h^*(X_{i_1})$, then selects another index i_2, etc., until it eventually terminates and returns a classifier. Denote by $\mathcal{Z} = \{(X_1, h^*(X_1)), (X_2, h^*(X_2)), \ldots\}$. To be *correct*, $\mathcal{A}$ must guarantee that for $h^* \sim \pi$, $\forall \varepsilon > 0$, $\mathbb{E}[\rho(\mathcal{A}(\varepsilon, \mathcal{D}, \pi), h^*)] \leq \varepsilon$. We define the random variable $N(\mathcal{A}, f, \varepsilon, \mathcal{D}, \pi)$ as the number of label requests $\mathcal{A}$ makes before terminating, when given ε, $\mathcal{D}$, and π as inputs, and when $h^* = f$ is the value of the target function; we make the particular data sequence $\mathcal{Z}$ the algorithm is run with implicit in this notation. We will be interested in the *expected sample complexity* $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = \mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)]$.

We propose the following algorithm $\mathcal{A}_\tau$ for transfer learning, defined in terms of a given correct prior-dependent active learning algorithm $\mathcal{A}_a$. We discuss interesting specifications for $\mathcal{A}_a$ in the next section, but for now the only assumption we require is that for any $\varepsilon > 0$ and $\mathcal{D}$, there is a value $s_\varepsilon < \infty$ such that for every π and $f \in \mathbb{C}$, $N(\mathcal{A}_a, f, \varepsilon, \mathcal{D}, \pi) \leq s_\varepsilon$; this is a very mild requirement, and any active learning algorithm can be converted into one that satisfies this without significantly increasing its sample complexities for the priors it is already good for [BHV10]. We denote by $m_\varepsilon = \frac{16d}{\varepsilon} \ln\left(\frac{24}{\varepsilon}\right)$, and $\mathrm{B}(\theta, \gamma) = \{\theta' \in \Theta : \|\pi_\theta - \pi_{\theta'}\| \leq \gamma\}$.

Algorithm 1 $\mathcal{A}_\tau(T,\varepsilon)$: an algorithm for transfer learning, specified in terms of a generic subroutine $\mathcal{A}_a$.

for $t = 1, 2, \ldots, T$ **do**
  Request labels $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$
  if $R(t-1, \varepsilon/2) > \varepsilon/8$ **then**
    Request labels $Y_{t(d+1)}(\theta_\star), \ldots, Y_{tm_\varepsilon}(\theta_\star)$
    Take $\hat{h}_t$ as any $h \in \mathbb{C}$ s.t. $\forall i \leq m_\varepsilon$, $h(X_{ti}) = Y_{ti}(\theta_\star)$
  else
    Let $\check{\theta}_{t\theta_\star} \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1,\varepsilon/2)\right)$ be such that
    $SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \leq \min\limits_{\theta \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1,\varepsilon/2)\right)} SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_\theta) + 1/t$
    Run $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})$ with data sequence $\mathcal{Z}_t(\theta_\star)$ and let $\hat{h}_t$ be the classifier it returns
  end if
end for

Theorem 3. *The algorithm $\mathcal{A}_\tau$ is correct. Furthermore, if $S_T(\varepsilon)$ is the total number of label requests made by $\mathcal{A}_\tau(T,\varepsilon)$, then $\limsup\limits_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}) + d$.*

The remarkable implication of Theorem 3 is that, via transfer learning, it is possible to achieve almost the *same* long-run average sample complexity as would be achievable if the target's prior distribution were *known* to the learner. We will see in the next section that this is sometimes significantly better than the single-task sample complexity.

The algorithm $\mathcal{A}_\tau$ is stated in a simple way here, but Theorem 3 can be improved with some obvious modifications to $\mathcal{A}_\tau$. The extra "$+d$" in Theorem 3 is not actually necessary, since we could stop updating the estimator $\check{\theta}_{t\theta_\star}$ (and the corresponding R value) after some $o(T)$ number of rounds (e.g., $\sqrt{T}$), in which case we would not need to request $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$ for t larger than this, and the extra $d \cdot o(T)$ number of labeled examples vanishes in the average as $T \to \infty$. Additionally, the $\varepsilon/4$ term can easily be improved to any value arbitrarily close to ε (even $(1-o(1))\varepsilon$) by running $\mathcal{A}_a$ with argument $\varepsilon - 2R(t-1,\varepsilon/2) - \delta(t-1,\varepsilon/2)$ instead of $\varepsilon/4$, and using this value in the SC calculations in the definition of $\check{\theta}_{t\theta_\star}$ as well. In fact, for many algorithms $\mathcal{A}_a$ (e.g., with $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$ continuous in ε), combining the above two tricks yields $\limsup\limits_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$.

Returning to our motivational remarks from Subsection 2.1, we can ask how many *extra* labeled examples are required from each learning problem to gain the benefits of transfer learning. This question essentially concerns the initial step of requesting the labels $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$. Clearly this indicates that from each learning problem, we need at most d extra labeled examples to gain the benefits of transfer. Whether these d label requests are indeed *extra* depends on the particular learning algorithm $\mathcal{A}_a$; that is, in some cases (e.g., certain passive learning algorithms), $\mathcal{A}_a$ may itself use these initial d labels for learning, so that in these cases the benefits of transfer learning are essentially gained as a *by-product* of the learning processes, and essentially no additional labeling effort need be expended to gain these benefits. On the other hand, for some active learning algorithms, we may expect that at least some of these initial

d labels would not be requested by the algorithm, so that some extra labeling effort is expended to gain the benefits of transfer in these cases.

4.1 Proof of Theorem 3

Recall that, to establish correctness, we must show that $\forall t \leq T$, $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$, regardless of the value of $\theta_\star \in \Theta$. Fix any $\theta_\star \in \Theta$ and $t \leq T$. If $R(t-1, \varepsilon/2) > \varepsilon/8$, then classic results from passive learning indicate that $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$ [Vap82]. Otherwise, by Theorem 1, with probability at least $1-\varepsilon/2$, we have $\|\pi_{\theta_\star} - \pi_{\hat{\theta}_{(t-1)\theta_\star}}\| \leq R(t-1, \varepsilon/2)$. On this event, if $R(t-1, \varepsilon/2) \leq \varepsilon/8$, then by a triangle inequality $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq 2R(t-1, \varepsilon/2) \leq \varepsilon/4$. Thus,

$$\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \mathbb{E}\left[\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\middle|\check{\theta}_{t\theta_\star}\right]\mathbb{1}\left[\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4\right]\right] + \varepsilon/2. \quad (2)$$

For $\theta \in \Theta$, let $\hat{h}_{t\theta}$ denote the classifier that would be returned by $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})$ when run with data sequence $\{(X_{t1}, h^*_{t\theta}(X_{t1})), (X_{t2}, h^*_{t\theta}(X_{t2})), \ldots\}$. Note that for any $\theta \in \Theta$, any measurable function $F : \mathbb{C} \to [0,1]$ has

$$\mathbb{E}\left[F(h^*_{t\theta_\star})\right] \leq \mathbb{E}\left[F(h^*_{t\theta})\right] + \|\pi_\theta - \pi_{\theta_\star}\|. \quad (3)$$

In particular, supposing $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4$, we have

$$\begin{aligned}
&\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\middle|\check{\theta}_{t\theta_\star}\right] = \mathbb{E}\left[\rho\left(\hat{h}_{t\theta_\star}, h^*_{t\theta_\star}\right)\middle|\check{\theta}_{t\theta_\star}\right] \\
&\leq \mathbb{E}\left[\rho\left(\hat{h}_{t\check{\theta}_{t\theta_\star}}, h^*_{t\check{\theta}_{t\theta_\star}}\right)\middle|\check{\theta}_{t\theta_\star}\right] + \|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2.
\end{aligned}$$

Combined with (2), this implies $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$.

We establish the sample complexity claim as follows. First note that convergence of $R(t-1, \varepsilon/2)$ implies that $\lim_{T\to\infty} \sum_{t=1}^{T} \mathbb{1}\left[R(t, \varepsilon/2) > \varepsilon/8\right]/T = 0$, and that the number of labels used for a value of t with $R(t-1, \varepsilon/2) > \varepsilon/8$ is bounded by a finite function m_ε of ε. Therefore,

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \\
&\leq d + \limsup_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\right]\mathbb{1}[R(t-1, \varepsilon/2) \leq \varepsilon/8]/T \\
&\leq d + \limsup_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\right]/T. \quad (4)
\end{aligned}$$

By the definition of R, δ from Theorem 1, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \mathbb{1}\left[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)\right]\right]$$

$$\leq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} s_{\varepsilon/4} \mathbb{P}\left(\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)\right)$$

$$\leq s_{\varepsilon/4} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \delta(t-1, \varepsilon/2) = 0.$$

Combined with (4), this implies

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq d+$$

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \mathbb{1}\left[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)\right]\right].$$

For any $t \leq T$, on the event $\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)$, we have (by the property (3) and a triangle inequality)

$$\begin{aligned}
&\mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \middle| \check{\theta}_{t\theta_\star}\right] \\
&\qquad \leq \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\check{\theta}_{t\theta_\star}}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \middle| \check{\theta}_{t\theta_\star}\right] + 2R(t-1, \varepsilon/2) \\
&\qquad = SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}\right) + 2R(t-1, \varepsilon/2) \\
&\qquad \leq SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right) + 1/t + 2R(t-1, \varepsilon/2),
\end{aligned}$$

where the last inequality follows by definition of $\check{\theta}_{t\theta_\star}$. Therefore,

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \\
&\leq d + \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right) + 1/t + 2R(t-1, \varepsilon/2) \\
&= d + SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right).
\end{aligned}$$

□

5 Application to Self-Verifying Active Learning

In this section, we examine a specific sample complexity guarantee achievable by active learning, when combined with the above transfer learning procedure. As mentioned, [BHV10] found that there is often a significant gap between the sample complexities achievable by good active learning algorithms in general and the sample complexities achievable by active learning algorithms that themselves adaptively determine how many label requests to make for any given problem (referred to as *self-verifying* active learning algorithms). Specifically, while the former can *always* be strictly superior to the sample complexities achievable by passive learning, there are simple examples where this is not the case for self-verifying active learning algorithms.

We should note, however, that all of the above considerations were proven for a learning scenario in which the target concept is considered a constant, and no information about the process that generates this concept is available to the learner. Thus, there is a natural question of whether, in the context of the transfer learning setting described above, we might be able to close this gap, so that self-verifying active learning algorithms are able to achieve the same type of guaranteed strict improvements over passive learning that are achievable by their non-self-verifying counterparts.

The considerations of the previous section indicate that this question is in some sense reducible to an analogous question for *prior-dependent* self-verifying active learning algorithms. The quantity $SC(\mathcal{A}_a, \varepsilon, /4, \mathcal{D}, \pi_{\theta_\star})$ then essentially characterizes the achievable average sample complexity among the sequence of tasks. We will therefore focus in this section on characterizing this quantity, for a particularly effective active learning algorithm $\mathcal{A}_a$.

Throughout this section, we suppose the active learning algorithm $\mathcal{A}_a$ is run with the sequence of examples $\mathcal{Z} = \{(X_1, h^*(X_1)), (X_2, h^*(X_2)), \ldots\}$, where $X_1, X_2, \ldots$ are i.i.d. $\mathcal{D}$, and $h^* \sim \pi$.

5.1 Related Work on Prior-dependent Learning

Prior-dependent learning algorithms have been studied in depth in the context of passive learning. In particular, [HKS92] found that for any concept space of finite VC dimension d, for any prior and data distribution, $O(d/\varepsilon)$ random labeled examples are sufficient for the expected error rate of the Bayes classifier produced under the posterior distribution to be at most ε. Furthermore, it is easy to construct learning problems for which there is an $\Omega(1/\varepsilon)$ lower bound on the number of random labeled examples required to achieve expected error rate at most ε, by any passive learning algorithm; for instance, the problem of learning threshold classifiers on $[0, 1]$ under a uniform data distribution and uniform prior is one such scenario.

In contrast, a relatively small amount is known about prior-dependent *active* learning. [FSST97] analyze the *Query By Committee* algorithm in this context, and find that if a certain information gain quantity for the points requested by the algorithm is lower-bounded by a value g, then the algorithm requires only $O((d/g)\log(1/\varepsilon))$ labels to achieve expected error rate at most ε. In particular, they show that this is satisfied for *constant* g for linear separators under a near-uniform prior, and a near-uniform data distribution over the unit sphere. This represents a marked improvement over the results of [HKS92] for passive learning, and since the Query by Committee algorithm is self-verifying, this result is highly relevant to the present discussion. However, the condition that the information gains be lower-bounded by a constant is quite restrictive, and many interesting learning problems are precluded by this requirement. Furthermore, there exist learning problems (with finite VC dimension) for which the Query by Committee algorithm makes an expected number of label requests exceeding $\Omega(1/\varepsilon)$. To date, there has not been a general analysis of how the value of g can behave as a function of ε, though such an analysis would likely be quite interesting.

In the present section, we take a more general approach to the question of prior-dependent active learning. We are interested in the broad question of whether access to the prior bridges the gap between the sample complexity of *learning* and the sample complexity of learning *with verification*. Specifically, we ask the following question.

Can a prior-dependent self-terminating active learning algorithm for a concept class of finite VC dimension always achieve expected error rate at most ε using $o(1/\varepsilon)$ label requests?

5.2 Prior-Independent Learning Algorithms

One may initially wonder whether we could achieve this $o(1/\varepsilon)$ result merely by calculating the expected sample complexity of some prior-independent method, thus precluding the need for novel algorithms. Indeed, for some $\mathbb{C}$ and $\mathcal{D}$, it is known that there *are* π-independent active learning algorithms (meaning the algorithm's behavior is independent of the π argument) $\mathcal{A}$ that have $\mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)|h^*] = o(1/\varepsilon)$ (always); for instance, threshold classifiers have this property under any $\mathcal{D}$, homogeneous linear separators have this property under a uniform $\mathcal{D}$ on the unit sphere in k dimensions, and intervals with positive width on $\mathcal{X} = [0, 1]$ have this property under $\mathcal{D} = \mathrm{Uniform}([0, 1])$ (see e.g., [Das05]). It is straightforward to show that any such $\mathcal{A}$ will also have $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon)$ for every π. In particular, the law of total expectation and the dominated convergence theorem imply

$$\begin{aligned}\lim_{\varepsilon \to 0} \varepsilon SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) &= \lim_{\varepsilon \to 0} \varepsilon \mathbb{E}[\mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)|h^*]] \\ &= \mathbb{E}\left[\lim_{\varepsilon \to 0} \varepsilon \mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)|h^*]\right] = 0.\end{aligned}$$

In these cases, we can think of SC as a kind of *average-case* analysis of these algorithms. However, there are also many well-known $\mathbb{C}$ and $\mathcal{D}$ for which there is no π-independent algorithm achieving $o(1/\varepsilon)$ sample complexity for *all* priors. For instance, this is the case for $\mathbb{C}$ as the space of interval classifiers (including the empty interval) on $\mathcal{X} = [0, 1]$ under $\mathcal{D} = \mathrm{Uniform}([0, 1])$; this essentially follows from a proof of [BHV10]. Thus, any general result on $o(1/\varepsilon)$ expected sample complexity for π-dependent algorithms would indicate that there is a real advantage to having access to the prior, beyond the apparent *smoothing* effects of an average-case analysis.

5.3 An Example: Intervals

We begin our exploration of π-dependent active learning with a concrete example, namely interval classifiers under a uniform data density but arbitrary prior, to illustrate how access to the prior can make a difference in the sample complexity. Specifically, consider $\mathcal{X} = [0, 1]$, $\mathcal{D}$ uniform on $[0, 1]$, and the concept space of *interval classifiers*: $\mathbb{C} = \{\mathbb{1}^{\pm}_{(a,b)} : 0 \le a \le b \le 1\}$, where $\mathbb{1}^{\pm}_{(a,b)}(x) = +1$ if $x \in (a, b)$ and -1 otherwise. For each classifier $h \in \mathbb{C}$, let $w(h) = \mathcal{D}(x : h(x) = +1)$ (the width of the interval h). Note that because we allow $a = b$, there is a classifier $h_- \in \mathbb{C}$ with $w(h_-) = 0$.

For simplicity, in this example (only) we will suppose the algorithm may request the label of *any* point in $\mathcal{X}$, not just those in the sequence $\{X_i\}$; the same ideas can easily be adapted to the setting where queries are restricted to $\{X_i\}$. Consider an active learning algorithm that sequentially requests the labels $h^*(x)$ for points x at $1/2$, $1/4$, $3/4, 1/8, 3/8, 5/8, 7/8, 1/16, 3/16$, etc., until (case 1) it encounters an example x with $h^*(x) = +1$ or until (case 2) the set of classifiers $V \subseteq \mathbb{C}$ consistent with all observed labels so far satisfies $\mathbb{E}[w(h^*)|V] \leq \varepsilon$ (which ever comes first). In case 2, the algorithm simply halts and returns the constant classifier that always predicts -1: call it h_-; note that $\rho(h_-, h^*) = w(h^*)$. In case 1, the algorithm enters a second phase, in which it performs a binary search (repeatedly querying the midpoint between the closest two -1 and $+1$ points, taking 0 and 1 as known negative points) to the left and right of the observed positive point, halting after $\log_2(4/\varepsilon)$ label requests on each side; this results in estimates of the target's endpoints up to $\pm\varepsilon/4$, so that returning any classifier among the set $V \subseteq \mathbb{C}$ consistent with these labels results in error rate at most ε; in particular, if $\tilde{h}$ is the classifer in V returned, then $\mathbb{E}[\rho(\tilde{h}, h^*)|V] \leq \varepsilon$.

Denoting this algorithm by $\mathcal{A}_{[]}$, and $\hat{h}$ the classifier it returns, we have

$$\mathbb{E}\left[\rho\left(\hat{h}, h^*\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\rho\left(\hat{h}, h^*\right)\middle|V\right]\right] \leq \varepsilon,$$

so that the algorithm is definitely correct.

Note that case 2 will definitely be satisfied after at most $\frac{2}{\varepsilon}$ label requests, and if $w(h^*) > \varepsilon$, then case 1 will definitely be satisfied after at most $\frac{2}{w(h^*)}$ label requests, so that the algorithm never makes more than $\frac{2}{\max\{w(h^*),\varepsilon\}}$ label requests before satisfying one of the two cases. Abbreviating $N(h^*) = N(\mathcal{A}_{[]}, h^*, \varepsilon, \mathcal{D}, \pi)$, we have

$$\begin{aligned}
&\mathbb{E}\left[N(h^*)\right] \\
&= \mathbb{E}\left[N(h^*)\middle|w(h^*) = 0\right]\mathbb{P}\left(w(h^*) = 0\right) \\
&\quad + \mathbb{E}\left[N(h^*)\middle|0 < w(h^*) \leq \sqrt{\varepsilon}\right]\mathbb{P}\left(0 < w(h^*) \leq \sqrt{\varepsilon}\right) \\
&\quad + \mathbb{E}\left[N(h^*)\middle|w(h^*) > \sqrt{\varepsilon}\right]\mathbb{P}\left(w(h^*) > \sqrt{\varepsilon}\right) \\
&\leq \mathbb{E}\left[N(h^*)\middle|w(h^*) = 0\right]\mathbb{P}\left(w(h^*) = 0\right) \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[N(h^*)\middle|h^*\right]\middle|0 < w(h^*) \leq \sqrt{\varepsilon}\right]\mathbb{P}\left(0 < w(h^*) \leq \sqrt{\varepsilon}\right) \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[N(h^*)\middle|h^*\right]\middle|w(h^*) > \sqrt{\varepsilon}\right]\mathbb{P}\left(w(h^*) > \sqrt{\varepsilon}\right) \\
&\leq \mathbb{E}\left[N(h^*)\middle|w(h^*) = 0\right]\mathbb{P}\left(w(h^*) = 0\right) + \frac{2}{\varepsilon}\mathbb{P}\left(0 < w(h^*) \leq \sqrt{\varepsilon}\right) + \frac{2}{\sqrt{\varepsilon}} + 2\log_2\frac{4}{\varepsilon}.
\end{aligned} \tag{5}$$

The third and fourth terms in (5) are $o(1/\varepsilon)$. Since $\mathbb{P}(0 < w(h^*) \leq \sqrt{\varepsilon}) \to 0$ as $\varepsilon \to 0$, the second term in (5) is $o(1/\varepsilon)$ as well. If $\mathbb{P}(\lambda(h^*) = 0) = 0$, this completes the proof. We focus the rest of the proof on the first term in (5), in the case that $\mathbb{P}(w(h^*) = 0) > 0$: i.e., there is nonzero probability that the target h^* labels the space almost all negative.

Letting V denote the subset of $\mathbb{C}$ consistent with all requested labels, note that on the event $w(h^*) = 0$, after n label requests (for n a power of 2) we have $\max_{h \in V} w(h) \leq 1/n$. Thus, for any value $w \in (0,1)$, after at most $\frac{2}{w}$ label requests, on the event that $w(h^*) = 0$,

$$\mathbb{E}\left[w(h^*)\middle|V\right] \leq \frac{\mathbb{E}\left[w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]\right]}{\pi(V)} \leq \frac{\mathbb{E}[w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]]}{\mathbb{P}(w(h^*) = 0)}. \tag{6}$$

Now note that, by the dominated convergence theorem,

$$\lim_{w \to 0} \mathbb{E}\left[\frac{w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]}{w}\right] = \mathbb{E}\left[\lim_{w \to 0} \frac{w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]}{w}\right] = 0.$$

Therefore, $\mathbb{E}\left[w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]\right] = o(w)$. If we define w_ε as the largest value of w for which $\mathbb{E}\left[w(h^*)\mathbb{1}\left[w(h^*) \leq w\right]\right] \leq \varepsilon\mathbb{P}(w(h^*) = 0)$ (or, say, half the supremum if the maximum is not achieved), then we have $w_\varepsilon \gg \varepsilon$. Combined with (6), this implies

$$\mathbb{E}\left[N(h^*)\middle|w(h^*) = 0\right] \leq \frac{2}{w_\varepsilon} = o(1/\varepsilon).$$

Thus, all of the terms in (5) are $o(1/\varepsilon)$, so that in total $\mathbb{E}[N(h^*)] = o(1/\varepsilon)$.

In conclusion, for this concept space $\mathbb{C}$ and data distribution $\mathcal{D}$, we have a correct active learning algorithm $\mathcal{A}$ achieving a sample complexity $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon)$ for all priors π on $\mathbb{C}$.

5.4 A General Result for Self-Verifying Bayesian Active Learning

In this subsection, we present our main result for improvements achievable by prior-dependent self-verifying active learning: a general result stating that $o(1/\varepsilon)$ expected sample complexity is always achievable for some appropriate prior-dependent active learning algorithm, for *any* $(\mathcal{X}, \mathbb{C}, \mathcal{D}, \pi)$ for which $\mathbb{C}$ has finite VC dimension. Since the known results for the sample complexity of passive learning with access to the prior are typically $\Theta(1/\varepsilon)$ [HKS92], and since there are known learning problems $(\mathcal{X}, \mathbb{C}, \mathcal{D}, \pi)$ for which every passive learning algorithm requires $\Omega(1/\varepsilon)$ samples, this $o(1/\varepsilon)$ result for active learning represents an improvement over passive learning.

The proof is simple and accessible, yet represents an important step in understanding the problem of self-termination in active learning algorithms, and the general issue of the complexity of verification. Also, as there are well-known problems $(\mathcal{X}, \mathbb{C}, \mathcal{D})$ for which no (single-task) prior-independent correct active learning algorithm (of the self-terminating type studied here) can achieve $o(1/\varepsilon)$ expected sample complexity for every π, this also represents a significant step toward understanding the inherent value of having access to the prior in active learning [BHV10]. Additionally, via Theorem 3, this result implies that active *transfer* learning (of the type discussed above) can provide strictly superior sample complexities compared to the known results for passive learning (even compared to passive learning algorithms having *direct* access to the prior $\pi_{\theta_\star}$), and often strictly superior to the sample complexities achievable by (prior-independent) active learning without transfer.

First, we have a small lemma.

Lemma 5. *For any sequence of functions $\phi_n : \mathbb{C} \to [0,\infty)$ such that, $\forall f \in \mathbb{C}$, $\phi_n(f) = o(1/n)$ and $\forall n \in \mathbb{N}$, $\phi_n(f) \leq c/n$ (for an f-independent constant $c \in (0,\infty)$), there exists a sequence $\bar{\phi}_n$ in $[0,\infty)$ such that*

$$\bar{\phi}_n = o(1/n) \text{ and } \lim_{n\to\infty} \mathbb{P}\left(\phi_n(h^*) > \bar{\phi}_n\right) = 0.$$

Proof. For any constant $\gamma \in (0,\infty)$, we have (by Markov's inequality and the dominated convergence theorem)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(n\phi_n(h^*) > \gamma\right) &\leq \frac{1}{\gamma} \lim_{n\to\infty} \mathbb{E}\left[n\phi_n(h^*)\right] \\ &= \frac{1}{\gamma}\mathbb{E}\left[\lim_{n\to\infty} n\phi_n(h^*)\right] = 0.\end{aligned}$$

Therefore (by induction), there exists a diverging sequence n_i in $\mathbb{N}$ such that

$$\lim_{i\to\infty} \sup_{n \geq n_i} \mathbb{P}\left(n\phi_n(h^*) > 2^{-i}\right) = 0.$$

Inverting this, let $i_n = \max\{i \in \mathbb{N} : n_i \leq n\}$, and define $\bar{\phi}_n(h^*) = (1/n) \cdot 2^{-i_n}$. By construction, $\mathbb{P}\left(\phi_n(h^*) > \bar{\phi}_n\right) \to 0$. Furthermore, $n_i \to \infty \implies i_n \to \infty$, so that we have

$$\lim_{n\to\infty} n\bar{\phi}_n = \lim_{n\to\infty} 2^{-i_n} = 0,$$

implying $\bar{\phi}_n = o(1/n)$. □

Theorem 4. *For any VC class $\mathbb{C}$, there is a correct active learning algorithm $\mathcal{A}_a$ that, for every data distribution $\mathcal{D}$ and prior π, achieves expected sample complexity*

$$SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon).$$

Our approach to proving Theorem 4 is via a reduction to established results about (prior-independent) active learning algorithms that are *not* self-verifying. Specifically, consider a slightly different type of active learning algorithm than that defined above: namely, an algorithm $\mathcal{A}_b$ that takes as input a *budget* $n \in \mathbb{N}$ on the number of label requests it is allowed to make, and that after making at most n label requests returns as output a classifier $\hat{h}_n$. Let us refer to any such algorithm as a *budget-based* active learning algorithm. Note that budget-based active learning algorithms are prior-independent (have no direct access to the prior). The following result was proven by [Han09] (see also the related earlier work of [BHV10]).

Lemma 6. *[Han09] For any VC class $\mathbb{C}$, there exists a constant $c \in (0,\infty)$, a function $\mathcal{E}(n; f, \mathcal{D})$, and a budget-based active learning algorithm $\mathcal{A}_b$ such that*

$$\forall \mathcal{D}, \forall f \in \mathbb{C}, \mathcal{E}(n; f, \mathcal{D}) \leq c/n \text{ and } \mathcal{E}(n; f, \mathcal{D}) = o(1/n),$$

and $\mathbb{E}\left[\rho\left(\mathcal{A}_b(n), h^*\right) \middle| h^*\right] \leq \mathcal{E}(n; h^*, \mathcal{D})$ *(always).*[4]

[4] Furthermore, it is not difficult to see that we can take this $\mathcal{E}$ to be measurable in the h^* argument.

That is, equivalently, for any fixed value for the target function, the expected error rate is $o(1/n)$, where the random variable in the expectation is only the data sequence $X_1, X_2, \ldots$. Our task in the proof of Theorem 4 is to convert such a budget-based algorithm into one that is correct, self-terminating, and prior-dependent, taking ε as input.

Proof (Theorem 4). Consider $\mathcal{A}_b$, $\mathcal{E}$, and c as in Lemma 6, let $\hat{h}_n$ denote the classifier returned by $\mathcal{A}_b(n)$, and define

$$n_{\pi,\varepsilon} = \min\left\{n \in \mathbb{N} : \mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right] \leq \varepsilon\right\}.$$

This value is accessible based purely on access to π and $\mathcal{D}$. Furthermore, we clearly have (by construction) $\mathbb{E}\left[\rho\left(\hat{h}_{n_{\pi,\varepsilon}}, h^*\right)\right] \leq \varepsilon$. Thus, letting $\mathcal{A}_a$ denote the active learning algorithm taking $(\mathcal{D}, \pi, \varepsilon)$ as input, which runs $\mathcal{A}_b(n_{\pi,\varepsilon})$ and then returns $\hat{h}_{n_{\pi,\varepsilon}}$, we have that $\mathcal{A}_a$ is a *correct* learning algorithm (i.e., its expected error rate is at most ε).

As for the expected sample complexity $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi)$ achieved by $\mathcal{A}_a$, we have $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi) \leq n_{\pi,\varepsilon}$, so that it remains only to bound $n_{\pi,\varepsilon}$. By Lemma 5, there is a π-dependent function $\mathcal{E}(n; \pi, \mathcal{D})$ such that

$$\pi\left(\{f \in \mathbb{C} : \mathcal{E}(n; f, \mathcal{D}) > \mathcal{E}(n; \pi, \mathcal{D})\}\right) \to 0$$
$$\text{and } \mathcal{E}(n; \pi, \mathcal{D}) = o(1/n).$$

Therefore, by the law of total expectation,

$$\begin{aligned}
&\mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\Big| h^*\right]\right] \leq \mathbb{E}\left[\mathcal{E}(n; h^*, \mathcal{D})\right] \\
&\leq \frac{c}{n}\pi\left(\{f \in \mathbb{C} : \mathcal{E}(n; f, \mathcal{D}) > \mathcal{E}(n; \pi, \mathcal{D})\}\right) + \mathcal{E}(n; \pi, \mathcal{D}) \\
&= o(1/n).
\end{aligned}$$

If $n_{\pi,\varepsilon} = O(1)$, then clearly $n_{\pi,\varepsilon} = o(1/\varepsilon)$ as needed. Otherwise, since $n_{\pi,\varepsilon}$ is monotonic in ε, we must have $n_{\pi,\varepsilon} \uparrow \infty$ as $\varepsilon \downarrow 0$. In particular, in this latter case we have

$$\begin{aligned}
&\lim_{\varepsilon \to 0} \varepsilon \cdot n_{\pi,\varepsilon} \\
&\leq \lim_{\varepsilon \to 0} \varepsilon \cdot \left(1 + \max\left\{n \geq n_{\pi,\varepsilon} - 1 : \mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right] > \varepsilon\right\}\right) \\
&= \lim_{\varepsilon \to 0} \varepsilon \cdot \max_{n \geq n_{\pi,\varepsilon}-1} n \mathbb{1}\left[\mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right]/\varepsilon > 1\right] \\
&\leq \lim_{\varepsilon \to 0} \varepsilon \cdot \max_{n \geq n_{\pi,\varepsilon}-1} n \mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right]/\varepsilon \\
&= \lim_{\varepsilon \to 0} \max_{n \geq n_{\pi,\varepsilon}-1} n \mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right] = \limsup_{n \to \infty} n \mathbb{E}\left[\rho\left(\hat{h}_n, h^*\right)\right] = 0,
\end{aligned}$$

so that $n_{\pi,\varepsilon} = o(1/\varepsilon)$, as required. □

Theorem 4 implies that, if we have *direct* access to the prior distribution of h^*, regardless of what that prior distribution π is, we can always construct a *self-verifying* active learning algorithm $\mathcal{A}_a$ that has a guarantee of $\mathbb{E}\left[\rho\left(\mathcal{A}_a(\varepsilon,\mathcal{D},\pi),h^*\right)\right] \leq \varepsilon$ and its expected number of label requests is $o(1/\varepsilon)$. This guarantee is *not* possible for prior-independent (single-task) self-verifying active learning algorithms.

Additionally, when combined with Theorem 3, Theorem 4 implies that $\mathcal{A}_\tau$, with this particular algorithm $\mathcal{A}_a$ as its subroutine, has $\limsup_{T\to\infty}\mathbb{E}[S_T(\varepsilon)]/T = o(1/\varepsilon)$. Again, since there are known cases where there is *no* prior-independent self-verifying active learning algorithm with sample complexity $o(1/\varepsilon)$ [BHV10], this sometimes represents a significant improvement over the results provable for learning the tasks independently (i.e., without transfer).

5.5 Dependence on $\mathcal{D}$ in the Learning Algorithm

The dependence on $\mathcal{D}$ in the algorithm described in the proof of Theorem 4 is fairly weak, and we can eliminate any direct dependence on $\mathcal{D}$ by replacing $\rho\left(\hat{h}_n,h^*\right)$ by a $1-\varepsilon/2$ confidence upper bound based on $M_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ i.i.d. unlabeled examples $X_1',X_2',\ldots,X_{M_\varepsilon}'$ independent from the examples used by the algorithm (e.g., set aside in a pre-processing step, where the bound is calculated via Hoeffding's inequality and a union bound over the values of n that we check, of which there are at most $O(1/\varepsilon)$). Then we simply increase the value of n (starting at some constant, such as 1) until

$$\frac{1}{M_\varepsilon}\sum_{i=1}^{M_\varepsilon}\pi\left(\left\{f\in\mathbb{C}: f\left(X_i'\right)\neq\hat{h}_n\left(X_i'\right)\right\}\right)\leq\varepsilon/2.$$

The expected value of the smallest value of n for which this occurs is $o(1/\varepsilon)$. Note that this only requires access to the prior π, not the data distribution $\mathcal{D}$ (the budget-based algorithm $\mathcal{A}_b$ of [Han09] has no direct dependence on $\mathcal{D}$); if desired for computational efficiency, this dependence may also be estimated by a $1-\varepsilon/4$ confidence upper bound based on $\Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ independent samples of h^* values with distribution π, where for each sample we simulate the execution of $\mathcal{A}_b(n)$ for that (simulated) target function in order to obtain the returned classifier. In particular, note that no actual label requests to the oracle are required during this process of estimating the appropriate label budget $n_{\pi,\varepsilon}$, as all executions of $\mathcal{A}_b$ are *simulated*.

5.6 Inherent Dependence on π in the Sample Complexity

We have shown that for every prior π, the sample complexity is bounded by a $o(1/\varepsilon)$ function. One might wonder whether it is possible that the asymptotic dependence on ε in the sample complexity can be prior-independent, while still being $o(1/\varepsilon)$. That is, we can ask whether there exists a (π-independent) function $s(\varepsilon) = o(1/\varepsilon)$ such that, for every π, there is a correct π-dependent algorithm $\mathcal{A}$ achieving a sample complexity $SC(\mathcal{A},\varepsilon,\mathcal{D},\pi) = O(s(\varepsilon))$, possibly involving π-dependent constants. Certainly in some cases, such as threshold classifiers, this is true. However, it seems this is not generally the case, and in particular it fails to hold for the space of interval classifiers.

For instance, consider a prior π on the space $\mathbb{C}$ of interval classifiers, constructed as follows. We are given an arbitrary monotonic $g(\varepsilon) = o(1/\varepsilon)$; since $g(\varepsilon) = o(1/\varepsilon)$, there must exist (nonzero) functions $q_1(i)$ and $q_2(i)$ such that $\lim_{i\to\infty} q_1(i) = 0$, $\lim_{i\to\infty} q_2(i) = 0$, and $\forall i \in \mathbb{N}, g(q_1(i)/2^{i+1}) \leq q_2(i) \cdot 2^i$; futhermore, letting $q(i) = \max\{q_1(i), q_2(i)\}$, by monotonicity of g we also have $\forall i \in \mathbb{N}, g(q(i)/2^{i+1}) \leq q(i) \cdot 2^i$, and $\lim_{i\to\infty} q(i) = 0$. Then define a function $p(i)$ with $\sum_{i\in\mathbb{N}} p(i) = 1$ such that $p(i) \geq q(i)$ for infinitely many $i \in \mathbb{N}$; for instance, this can be done inductively as follows. Let $\alpha_0 = 1/2$; for each $i \in \mathbb{N}$, if $q(i) > \alpha_{i-1}$, set $p(i) = 0$ and $\alpha_i = \alpha_{i-1}$; otherwise, set $p(i) = \alpha_{i-1}$ and $\alpha_i = \alpha_{i-1}/2$. Finally, for each $i \in \mathbb{N}$, and each $j \in \{0, 1, \ldots, 2^i - 1\}$, define $\pi\left(\left\{\mathbb{1}^{\pm}_{(j\cdot 2^{-i},(j+1)\cdot 2^{-i})}\right\}\right) = p(i)/2^i$.

We let $\mathcal{D}$ be uniform on $\mathcal{X} = [0, 1]$. Then for each $i \in \mathbb{N}$ s.t. $p(i) \geq q(i)$, there is a $p(i)$ probability the target interval has width 2^{-i}, and given this any algorithm requires $\propto 2^i$ expected number of requests to determine which of these 2^i intervals is the target, failing which the error rate is at least 2^{-i}. In particular, letting $\varepsilon_i = p(i)/2^{i+1}$, any correct algorithm has sample complexity at least $\propto p(i)\cdot 2^i$ for $\varepsilon = \varepsilon_i$. Noting $p(i)\cdot 2^i \geq q(i) \cdot 2^i \geq g(q(i)/2^{i+1}) \geq g(\varepsilon_i)$, this implies there exist arbitrarily small values of $\varepsilon > 0$ for which the optimal sample complexity is at least $\propto g(\varepsilon)$, so that the sample complexity is *not* $o(g(\varepsilon))$.

For any $s(\varepsilon) = o(1/\varepsilon)$, there exists a monotonic $g(\varepsilon) = o(1/\varepsilon)$ such that $s(\varepsilon) = o(g(\varepsilon))$. Thus, constructing π as above for this g, we have that the sample complexity is not $o(g(\varepsilon))$, and therefore not $O(s(\varepsilon))$. So at least for the space of interval classifiers, the specific $o(1/\varepsilon)$ asymptotic dependence on ε is inherently π-dependent.

6 Conclusions

We have shown that when learning a sequence of i.i.d. target concepts from a known VC class, with an unknown distribution from a known totally bounded family, transfer learning can lead to amortized average sample complexity close to that achievable by an algorithm with direct knowledge of the the targets' distribution. Furthermore, for the problem of active learning (with self-verification), we have shown that this latter quantity is always $o(1/\varepsilon)$, where ε is the desired expected error rate. This represents an improvement in the asymptotic dependence on ε compared to the general results provable for active learning without transfer [BHV10].

Acknowledgments

We would like to extend our sincere thanks to Avrim Blum for several thought-provoking discussions on this topic.

References

[AZ04] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Technical Report RC23462, IBM T.J. Watson Research Center, 2004.

[Bax97] J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39, 1997.

[Bax00] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[BBL09] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[BBZ07] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007.

[BDL09] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the International Conference on Machine Learning*, 2009.

[BDS03] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Conference on Learning Theory*, 2003.

[BEYL04] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *The Journal of Machine Learning Research*, 5:255–291, 2004.

[BHV10] M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, September 2010.

[CAL94] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201221, 1994.

[Car83] J. G. Carbonell. Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G.Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach*. Tioga Press, Palo Alto, CA, 1983.

[Car86] J. G. Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach, Volume II*. Morgan Kaufmann, 1986.

[Car97] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[CCS00] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *International Conference on Machine Learning*, 2000.

[CN08] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008.

[Das04] S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, pages 337–344. MIT Press, 2004.

[Das05] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Proc. of Neural Information Processing Systems (NIPS)*, 2005.

[DC08] P. Donmez and J. Carbonell. Paired sampling in density-sensitive active learning. In *Proceedings of the 10th International Symposium on Artificial Intelligence and Mathematics (SIAM)*, 2008.

[DCB07] P. Donmez, J. Carbonell, and P. Bennett. Dual strategy active learning. In *Proceedings of the 18th European Conference on Machine Learning*, 2007.

[DHM08] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, 2008.

[DKM09] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, NY, USA, 2001.

[EMP05] T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[EP04] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[Fri09] E. Friedman. Active learning for smooth problems. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

[FSST97] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. In *Machine Learning*, pages 133–168, 1997.

[Han07a] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proc. of the 24th international conference on Machine learning*, 2007.

[Han07b] S. Hanneke. Teaching dimension and the complexity of active learning. In *Proc. of the 20th Annual Conference on Learning Theory (COLT)*, 2007.

[Han09] S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009.

[Han11] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

[HC08] J. He and J. Carbonell. Rare class discovery based on active learning. In *Proceedings of the 10th International Symposium on Artificial Intelligence and Mathematics*, 2008.

[HKS92] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. In *Machine Learning*, pages 61–74. Morgan Kaufmann, 1992.

[Kää06] M. Kääriäinen. Active learning in the non-realizable case. In *Proc. of the 17th International Conference on Algorithmic Learning Theory*, 2006.

[Kol93] J. Kolodner (Ed). *Case-Based Learning*. Kluwer Academic Publishers, The Netherlands, 1993.

[Kol10] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, To Appear, 2010.

[MN98] A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. In *International Conference on Machine Learning*, 1998.

[MP04] C. Micchelli and M. Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing 18*, 2004.

[Now08] R. D. Nowak. Generalized binary search. In *Proceedings of the 46*$^{\mathrm{th}}$ *Annual Allerton Conference on Communication, Control, and Computing*, 2008.

[NS04] H. Nguyen and A. Smeulders. Active learning using pre-clustering. In *International Conference on Machine Learning*, 2004.

[Sch95] M. J. Schervish. *Theory of Statistics*. Springer, New York, NY, USA, 1995.

[Sil00] D. L. Silver. *Selective Transfer of Neural Network Task Knowledge*. PhD thesis, Computer Science, University of Western Ontario, 2000.

[Thr96] S. Thrun. Is learning the n-th thing any easier than learning the first? In *In Advances in Neural Information Processing Systems 8*, 1996.

[TK01] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(1):45–66, 2001.

[Vap82] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[VC93] M. M. Veloso and J. G. Carbonell. Derivational analogy in prodigy: Automating case acquisition, storage and utilization. *Machine Learning*, 10:249–278, 1993.

[Wan09] L. Wang. Sufficient conditions for agnostic active learnable. In *Advances in Neural Information Processing Systems 22*, 2009.

[Yat85] Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13:768–774, 1985.