Bounds on the Minimax Rate for Estimating a Prior over a VC Class from Independent Learning Tasks

Liu Yang **Steve Hanneke** **Jaime Carbonell**

December 2012
CMU-ML-12-112

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We study the optimal rates of convergence for estimating a prior distribution over a VC class from a sequence of independent data sets respectively labeled by independent target functions sampled from the prior. We specifically derive upper and lower bounds on the optimal rates under a smoothness condition on the correct prior, with the number of samples per data set equal the VC dimension. These results have implications for the improvements achievable via transfer learning.

This research is supported in part by NSF grant IIS-1065251 and a Google Core AI grant.

Keywords: Minimax Rates, Transfer Learning, VC Dimension, Bayesian Learning

1 Introduction

In the *transfer learning* setting, we are presented with a sequence of learning problems, each with some respective target concept we are tasked with learning. The key question in transfer learning is how to leverage our access to past learning problems in order to improve performance on learning problems we will be presented with in the future.

Among the several proposed models for transfer learning, one particularly appealing model supposes the learning problems are independent and identically distributed, with unknown distribution, and the advantage of transfer learning then comes from the ability to estimate this shared distribution based on the data from past learning problems [Baxter, 1997, Yang et al., 2011]. For instance, when customizing a speech recognition system to a particular speaker's voice, we might expect the first few people would need to speak many words or phrases in order for the system to accurately identify the nuances. However, after performing this for many different people, if the software has access to those past training sessions when customizing itself to a new user, it should have identified important properties of the speech patterns, such as the common patterns within each of the major dialects or accents, and other such information about the *distribution* of speech patterns within the user population. It should then be able to leverage this information to reduce the number of words or phrases the next user needs to speak in order to train the system, for instance by first trying to identify the individual's dialect, then presenting phrases that differentiate common subpatterns within that dialect, and so forth.

In analyzing the benefits of transfer learning in such a setting, one important question to ask is how quickly we can estimate the distribution from which the learning problems are sampled. In recent work, [Yang et al., 2011] have shown that under mild conditions on the family of possible distributions, if the target concepts reside in a known VC class, then it is possible to estimate this distribtion to arbitrary precision using only a bounded number of training samples per task: specifically, a number of samples equal the VC dimension. However, that work left open the question of quantifying the *rate* of convergence. This rate of convergence can have a direct impact on how much benefit we gain from transfer learning when we are faced with only a finite sequence of learning problems. As such, it is certainly desirable to derive tight characterizations of this rate of convergence.

The present work continues that of [Yang et al., 2011], bounding the rate of convergence for estimating this distribution, under a smoothness condition on the distribution. We derive a generic upper bound, which holds regardless of the VC class the target concepts reside in. The proof of this result builds on the earlier work of [Yang et al., 2011], but requires several interesting innovations to make the rate of convergence explicit, and to dramatically improve the upper bounds on certain quantities compared to the analogous bounds implicit in the original proofs. We further derive a nontrivial lower bound that holds for certain constructed scenarios, which illustrates a lower limit on how good of a general upper bound we might hope for in results expressed only in terms of the number of tasks, the smoothness conditions, and the VC dimension.

2 The Setting

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a Borel space [Schervish, 1995] (where $\mathcal{X}$ is called the *instance space*), and let $\mathcal{D}$ be a distribution on $\mathcal{X}$ (called the *data distribution*). Let $\mathbb{C}$ be a VC class of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$ (called the *concept space*), and denote by d the VC dimension of $\mathbb{C}$ [Vapnik, 1982]. We suppose $\mathbb{C}$ 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 our results can be formulated for general $\mathcal{D}$ (with somewhat more complicated theorem statements), to simplify the statement of results we suppose ρ is actually a *metric*, which would follow from appropriate topological conditions on $\mathbb{C}$ relative to $\mathcal{D}$. For any two probability measures μ_1, μ_2 on a measurable space $(\Omega, \mathcal{F})$, define the total variation distance

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

Let $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$ be a family of probability measures on $\mathbb{C}$ (called *priors*), where Θ is an arbitrary index set (called the *parameter space*). We additionally suppose there exists a probability measure π_0 on $\mathbb{C}$ (called the *reference measure*) such that every π_θ is absolutely continuous with respect to π_0, and therefore has a density function f_θ given by the Radon-Nikodym derivative $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_0}$ [Schervish, 1995].

We consider the following type of estimation problem. There is a collection of $\mathbb{C}$-valued random variables $\{h^*_{t\theta} : t \in \mathbb{N}, \theta \in \Theta\}$, where for any fixed $\theta \in \Theta$ the $\{h^*_{t\theta}\}_{t=1}^{\infty}$ variables are i.i.d. with distribution π_θ. 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 $t, i \in \mathbb{N}$, $Y_{ti}(\theta) = h^*_{t\theta}(X_{ti})$. We additionally denote by $\mathcal{Z}_{tk} = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$ the first k elements of $\mathcal{Z}_t(\theta)$, for any $k \in \mathbb{N}$, and similarly $\mathbb{X}_{tk} = \{X_{t1}, \ldots, X_{tk}\}$ and $\mathbb{Y}_{tk}(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$. Following the terminology used in the transfer learning literature, we refer to the collection of variables associated with each t collectively as the t^{th} *task*. We will be concerned with sequences of estimators $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1k}(\theta), \ldots, \mathcal{Z}_{Tk}(\theta))$, for $T \in \mathbb{N}$, which are based on only a bounded number k of samples per task, among the first T tasks. Our main results specifically study the case of $k = d$. For any such estimator, we measure the *risk* as $\mathbb{E}\left[\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\|\right]$, and will be particularly interested in upper-bounding the worst-case risk $\sup_{\theta_\star \in \Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\|\right]$ as a function of T, and lower-bounding the minimum possible value of this worst-case risk over all possible $\hat{\theta}_T$ estimators (called the *minimax risk*).

In previous work, [Yang et al., 2011] showed that, if Π_Θ is a totally bounded family, then even with only d number of samples per task, the minimax risk (as a function of the number of tasks T) converges to zero. In fact, that work also includes a proof that this is not necessarily the case in general for any number of samples less than d. However, the actual rates of convergence were not explicitly derived in that work, and indeed the upper bounds on the rates of convergence implicit in that analysis may often have fairly complicated dependences on $\mathbb{C}$, Π_Θ, and $\mathcal{D}$, and furthermore often provide only very slow rates of convergence.

To derive explicit bounds on the rates of convergence, in the present work we specifically focus on families of *smooth* densities. The motivation for involving a notion of smoothness in characterizing rates of convergence is clear if we consider the extreme case in which Π_Θ contains two priors π_1 and π_2, with $\pi_1(\{h\}) = \pi_2(\{g\}) = 1$, where $\rho(h, g)$ is a very small but nonzero

value; in this case, if we have only a small number of samples per task, we would require many tasks (on the order of $1/\rho(h,g)$) to observe any data points carrying any information that would distinguish between these two priors (namely, points x with $h(x) \neq g(x)$); yet $\|\pi_1 - \pi_2\| = 1$, so that we have a slow rate of convergence (at least initially). A total boundedness condition on Π_Θ would limit the number of such pairs present in Π_Θ, so that for instance we cannot have arbitrarily close h and g, but less extreme variants of this can lead to slow asymptotic rates of convergence as well.

Specifically, in the present work we consider the following notion of smoothness. For $L \in (0,\infty)$ and $\alpha \in (0,1]$, a function $f : \mathbb{C} \to \mathbb{R}$ is (L,α)-Hölder smooth if

$$\forall h, g \in \mathbb{C}, |f(h) - f(g)| \leq L\rho(h,g)^\alpha.$$

3 An Upper Bound

We now have the following theorem, holding for an arbitrary VC class $\mathbb{C}$ and data distribution $\mathcal{D}$; it is the main result of this work.

Theorem 1. *For Π_Θ any class of priors on $\mathbb{C}$ having (L,α)-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta), \ldots, \mathcal{Z}_{Td}(\theta))$ such that*

$$\sup_{\theta_\star \in \Theta} \mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

Proof. By the standard PAC analysis [Vapnik, 1982, Blumer et al., 1989], for any $\gamma > 0$, with probability greater than $1-\gamma$, a sample of $k = O((d/\gamma)\log(1/\gamma))$ random points will partition $\mathbb{C}$ into regions of width less than γ. For brevity, we omit the t subscript on quantities such as $\mathcal{Z}_{tk}(\theta)$ throughout the following analysis, since the claims hold for any arbitrary value of t.

For any $\theta \in \Theta$, let π'_θ denote a (conditional on $X_1, \ldots, X_k$) distribution defined as follows. Let f'_θ denote the (conditional on $X_1, \ldots, X_k$) density function of π'_θ with respect to π_0, and for any $g \in \mathbb{C}$, let $f'_\theta(g) = \frac{\pi_\theta(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = g(X_i)\})}{\pi_0(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = g(X_i)\})}$ (or 0 if $\pi_0(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = g(X_i)\}) = 0$). In other words, π'_θ has the same probability mass as π_θ for each of the equivalence classes induced by $X_1, \ldots, X_k$, but conditioned on the equivalence class, simply has a constant-density distribution over that equivalence class. Note that, by the smoothness condition, with probability greater than $1-\gamma$, we have *everywhere*

$$|f_\theta(h) - f'_\theta(h)| < L\gamma^\alpha.$$

So for any $\theta, \theta' \in \Theta$, with probability greater than $1-\gamma$,

$$\|\pi_\theta - \pi_{\theta'}\| = (1/2)\int |f_\theta - f_{\theta'}| \mathrm{d}\pi_0 < L\gamma^\alpha + (1/2)\int |f'_\theta - f'_{\theta'}| \mathrm{d}\pi_0.$$

Furthermore, since the regions that define f'_θ and $f'_{\theta'}$ are the same (namely, the partition induced by $X_1, \ldots, X_k$), we have $(1/2)\int |f'_\theta - f'_{\theta'}| \mathrm{d}\pi_0 = \frac{1}{2}\sum_{y_1,\ldots,y_k \in \{-1,+1\}} |\pi_\theta(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) =$

$y_i\}) - \pi_{\theta'}(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = y_i\})| = \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|$. Thus, we have that with probability at least $1-\gamma$,

$$\|\pi_\theta - \pi_{\theta'}\| < L\gamma^\alpha + \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.$$

Following analogous to the inductive argument of [Yang et al., 2011], suppose $I \subseteq \{1,\ldots,k\}$, fix $\bar{x}_I \in \mathcal{X}^{|I|}$ and $\bar{y}_I \in \{-1,+1\}^{|I|}$. Then the $\tilde{y}_I \in \{-1,+1\}^{|I|}$ for which no $h \in \mathbb{C}$ has $h(\bar{x}_I) = \tilde{y}_I$ for which $\|\bar{y}_I - \tilde{y}_I\|_1$ is minimal, has $(1/2)\|\bar{y}_I - \tilde{y}_I\|_1 \leq d+1$, and for any $i \in I$ with $\bar{y}_i \neq \tilde{y}_i$, letting $\bar{y}'_j = \bar{y}_j$ for $j \in I \setminus \{i\}$ and $\bar{y}'_i = \tilde{y}_i$, we have

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

and similarly for θ', so that

$$\begin{aligned}|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)| &\leq |\mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta')|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}})| \\ &+ |\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I)|.\end{aligned}$$

Now consider that these two terms inductively define a binary tree. Every time the tree branches left once, it arrives at a difference of probabilities for a set I of one less element than that of its parent. Every time the tree branches right once, it arrives at a difference of probabilities for a $\bar{y}_I$ one closer to an unrealized $\tilde{y}_I$ than that of its parent. Say we stop branching the tree upon reaching a set I and a $\bar{y}_I$ such that either $\bar{y}_I$ is an unrealized labeling, or $|I| = d$. Thus, we can bound the original (root node) difference of probabilities by the sum of the differences of probabilities for the leaf nodes with $|I| = d$. Any path in the tree can branch left at most $k-d$ times (total) before reaching a set I with only d elements, and can branch right at most $d+1$ times in a row before reaching a $\bar{y}_I$ such that both probabilities are zero, so that the difference is zero. So the depth of any leaf node with $|I| = d$ is at most $(k-d)d$. Furthermore, at any level of the tree, from left to right the nodes have strictly decreasing $|I|$ values, so that the maximum width of the tree is at most $k-d$. So the total number of leaf nodes with $|I| = d$ is at most $(k-d)^2 d$. Thus, for any $\bar{y} \in \{-1,+1\}^k$ and $\bar{x} \in \mathcal{X}^k$,

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

Since

$$\|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\| = (1/2) \sum_{\bar{y}^k \in \{-1,+1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,$$

and by Sauer's Lemma this is at most

$$(ek)^d \max_{\bar{y}^k \in \{-1,+1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,$$

we have that

$$\|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\| \leq (ek)^d k^2 d \max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|.$$

Thus, we have that

$$\|\pi_\theta - \pi_{\theta'}\| = \mathbb{E}\|\pi_\theta - \pi_{\theta'}\| < \gamma + L\gamma^\alpha + (ek)^d k^2 d \mathbb{E}\Big[\max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|\Big].$$

Note that

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

and by exchangeability, this last line equals

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

[Yang et al., 2011] showed that

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

so that in total we have

$$\|\pi_\theta - \pi_{\theta'}\| < (L+1)\gamma^\alpha + 4(2ek)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$$

Plugging in the value of $k = c(d/\gamma)\log(1/\gamma)$, this is

$$(L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$$

So the only remaining question is the rate of convergence of our estimate of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$. If $N(\varepsilon)$ is the ε-covering number of $\{\mathbb{P}_{\mathcal{Z}_d(\theta)} : \theta \in \Theta\}$, then taking $\hat{\theta}_{T\theta_\star}$ as the minimum distance skeleton estimate of [Yatracos, 1985, Devroye & Lugosi, 2001] achieves expected total variation distance ε from $\pi_{\theta_\star}$, for some $T = O((1/\varepsilon^2)\log N(\varepsilon/4))$. We can partition $\mathbb{C}$ into $O((L/\varepsilon)^{d/\alpha})$ cells of diameter $O((\varepsilon/L)^{1/\alpha})$, and set a constant density value within each cell, on an $O(\varepsilon)$-grid of density values, and every prior with (L,α)-Hölder smooth density will have density within ε of some density so-constructed; there are then at most $(1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$ such densities, so this bounds the covering numbers of Π_Θ. Furthermore, the covering number of Π_Θ upper bounds $N(\varepsilon)$ [Yang et al., 2011], so that $N(\varepsilon) \leq (1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$.

Solving $T = O(\varepsilon^{-2}(L/\varepsilon)^{d/\alpha}\log(1/\varepsilon))$ for ε, we have $\varepsilon = O\left(L\left(\frac{\log(TL)}{T}\right)^{\frac{\alpha}{d+2\alpha}}\right)$. So this bounds the rate of convergence for $\mathbb{E}\|\mathbb{P}_{\mathcal{Z}_d(\hat{\theta}_T)} - \mathbb{P}_{\mathcal{Z}_d(\theta_\star)}\|$, for $\hat{\theta}_T$ the minimum distance skeleton

estimate. Plugging this rate into the bound on the priors, combined with Jensen's inequality, we have

$$\mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| < (L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2} O\left(L\left(\frac{\log(TL)}{T}\right)^{\frac{\alpha}{2d+4\alpha}}\right).$$

This holds for any $\gamma > 0$, so minimizing this expression over $\gamma > 0$ yields a bound on the rate. For instance, with $\gamma = \tilde{O}\left(T^{-\frac{\alpha}{2(d+2\alpha)(\alpha+2(d+1))}}\right)$, we have

$$\mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

□

4 A Minimax Lower Bound

One natural quesiton is whether Theorem 1 can generally be improved. While we expect this to be true for some fixed VC classes (e.g., those of finite size), and in any case we expect that some of the constant factors in the exponent may be improvable, it is not at this time clear whether the general form of $T^{-\Theta(\alpha^2/(d+\alpha)^2)}$ is sometimes optimal. One way to investigate this question is to construct specific spaces $\mathbb{C}$ and distributions $\mathcal{D}$ for which a lower bound can be obtained. In particular, we are generally interested in exhibiting lower bounds that are worse than those that apply to the usual problem of density estimation based on direct access to the $h^*_{t\theta_\star}$ values (see Theorem 3 below).

Here we present a lower bound that is interesting for this reason. However, although larger than the optimal rate for methods with direct access to the target concepts, it is still far from matching the upper bound above, so that the question of tightness remains open. Specifically, we have the following result.

Theorem 2. *For any integer $d \geq 1$, any $L > 0, \alpha \in (0, 1]$, there is a value $C(d, L, \alpha) \in (0, \infty)$ such that, for any $T \in \mathbb{N}$, there exists an instance space $\mathcal{X}$, a concept space $\mathbb{C}$ of VC dimension d, a distribution $\mathcal{D}$ over $\mathcal{X}$, and a distribution π_0 over $\mathbb{C}$ such that, for Π_Θ a set of distributions over $\mathbb{C}$ with (L, α)-Hölder smooth density functions with respect to π_0, any estimator $\hat{\theta}_T = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$ has*

$$\sup_{\theta_\star \in \Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\|\right] \geq C(d, L, \alpha)T^{-\frac{\alpha}{2(d+\alpha)}}.$$

Proof. (Sketch) We proceed by a reduction from the task of determining the bias of a coin from among two given possibilities. Specifically, fix any $\gamma \in (0, 1/2)$, $n \in \mathbb{N}$, and let $B_1(p), \ldots, B_n(p)$ be i.i.d $\mathrm{Bernoulli}(p)$ random variables, for each $p \in [0, 1]$; then it is known that, for any (possibly nondeterministic) decision rule $\hat{p}_n : \{0, 1\}^n \to \{(1+\gamma)/2, (1-\gamma)/2\}$,

$$\frac{1}{2}\sum_{p \in \{(1+\gamma)/2, (1-\gamma)/2\}} \mathbb{P}(\hat{p}_n(B_1(p), \ldots, B_n(p)) \neq p) \geq (1/32) \cdot \exp\left\{-128\gamma^2 n/3\right\}. \tag{1}$$

This easily follows from the results of [Wald, 1945, Bar-Yossef, 2003], combined with a result of [Poland & Hutter, 2006] bounding the KL divergence.

To use this result, we construct a learning problem as follows. Fix some $m \in \mathbb{N}$ with $m \geq d$, let $\mathcal{X} = \{1, \ldots, m\}$, and let $\mathbb{C}$ be the space of all classifiers $h : \mathcal{X} \to \{-1, +1\}$ such that $|\{x \in \mathcal{X} : h(x) = +1\}| \leq d$. Clearly the VC dimension of $\mathbb{C}$ is d. Define the distribution $\mathcal{D}$ as uniform over $\mathcal{X}$. Finally, we specify a family of (L, α)-Hölder smooth priors, parameterized by $\Theta = \{-1, +1\}^{\binom{m}{d}}$, as follows. Let $\gamma_m = (L/2)(1/m)^\alpha$. First, enumerate the $\binom{m}{d}$ distinct d-sized subsets of $\{1, \ldots, m\}$ as $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{\binom{m}{d}}$. Define the reference distribution π_0 by the property that, for any $h \in \mathbb{C}$, letting $q = |\{x : h(x) = +1\}|$, $\pi_0(\{h\}) = (\frac{1}{2})^d \binom{m-q}{d-q} / \binom{m}{d}$. For any $\mathbf{b} = (b_1, \ldots, b_{\binom{m}{d}}) \in \{-1, 1\}^{\binom{m}{d}}$, define the prior $\pi_{\mathbf{b}}$ as the distribution of a random variable $h_{\mathbf{b}}$ specified by the following generative model. Let $i^* \sim \text{Uniform}(\{1, \ldots, \binom{m}{d}\})$, let $C_{\mathbf{b}}(i^*) \sim \text{Bernoulli}((1 + \gamma_m b_{i^*})/2)$; finally, $h_{\mathbf{b}} \sim \text{Uniform}(\{h \in \mathbb{C} : \{x : h(x) = +1\} \subseteq \mathcal{X}_{i^*}, \text{Parity}(|\{x : h(x) = +1\}|) = C_{\mathbf{b}}(i^*)\})$, where $\text{Parity}(n)$ is 1 if n is odd, or 0 if n is even. We will refer to the variables in this generative model below. For any $h \in \mathbb{C}$, letting $H = \{x : h(x) = +1\}$ and $q = |H|$, we can equivalently express $\pi_{\mathbf{b}}(\{h\}) = (\frac{1}{2})^d \binom{m}{d}^{-1} \sum_{i=1}^{\binom{m}{d}} \mathbb{1}[H \subseteq \mathcal{X}_i](1 + \gamma_m b_i)^{\text{Parity}(q)}(1 - \gamma_m b_i)^{1-\text{Parity}(q)}$. From this explicit representation, it is clear that, letting $f_{\mathbf{b}} = \frac{\mathrm{d}\pi_{\mathbf{b}}}{\mathrm{d}\pi_0}$, we have $f_{\mathbf{b}}(h) \in [1-\gamma_m, 1+\gamma_m]$ for all $h \in \mathbb{C}$. The fact that $f_{\mathbf{b}}$ is Hölder smooth follows from this, since every distinct $h, g \in \mathbb{C}$ have $\mathcal{D}(\{x : h(x) \neq g(x)\}) \geq 1/m = (2\gamma_m/L)^{1/\alpha}$.

Next we set up the reduction as follows. For any estimator $\hat{\pi}_T = \hat{\pi}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, and each $i \in \{1, \ldots, \binom{m}{d}\}$, let h_i be the classifier with $\{x : h_i(x) = +1\} = \mathcal{X}_i$; also, if $\hat{\pi}_T(\{h_i\}) > (\frac{1}{2})^d / \binom{m}{d}$, let $\hat{b}_i = 2\text{Parity}(d) - 1$, and otherwise $\hat{b}_i = 1 - 2\text{Parity}(d)$. We use these $\hat{b}_i$ values to estimate the original b_i values. Specifically, let $\hat{p}_i = (1 + \gamma_m \hat{b}_i)/2$ and $p_i = (1 + \gamma_m b_i)/2$, where $\mathbf{b} = \theta_\star$. Then

$$\begin{aligned}
\|\hat{\pi}_T - \pi_{\theta_\star}\| &\geq (1/2) \sum_{i=1}^{\binom{m}{d}} |\hat{\pi}_T(\{h_i\}) - \pi_{\theta_\star}(\{h_i\})| \\
&\geq (1/2) \sum_{i=1}^{\binom{m}{d}} \frac{\gamma_m}{2^d \binom{m}{d}} |\hat{b}_i - b_i|/2 \\
&= (1/2) \sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} |\hat{p}_i - p_i|.
\end{aligned}$$

Thus, we have reduced from the problem of deciding the biases of these $\binom{m}{d}$ independent Bernoulli random variables. To complete the proof, it suffices to lower bound the expectation of the right side for an *arbitrary* estimator.

Toward this end, we in fact study an even easier problem. Specifically, consider an estimator $\hat{q}_i = \hat{q}_i(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star), i_1^*, \ldots, i_T^*)$, where i_t^* is the i^* random variable in the generative model that defines $h^*_{t\theta_\star}$; that is, $i_t^* \sim \text{Uniform}(\{1, \ldots, \binom{m}{d}\})$, $C_t \sim \text{Bernoulli}((1 + \gamma_m b_{i_t^*})/2)$, and $h^*_{t\theta_\star} \sim \text{Uniform}(\{h \in \mathbb{C} : \{x : h(x) = +1\} \subseteq \mathcal{X}_{i_t^*}, \text{Parity}(|\{x : h(x) = +1\}|) = C_t\})$,

where the i_t^* are independent across t, as are the C_t and $h^*_{t\theta_\star}$. Clearly the $\hat{p}_i$ from above can be viewed as an estimator of this type, which simply ignores the knowledge of i_t^*. The knowledge of these i_t^* variables simplifies the analysis, since given $\{i_t^* : t \le T\}$, the data can be partitioned into $\binom{m}{d}$ disjoint sets, $\{\{\mathcal{Z}_{td}(\theta_\star) : i_t^* = i\} : i = 1, \ldots, \binom{m}{d}\}$, and we can use only the set $\{\mathcal{Z}_{td}(\theta_\star) : i_t^* = i\}$ to estimate p_i. Furthermore, we can use only the subset of these for which $\mathbb{X}_{td} = \mathcal{X}_i$, since otherwise we have zero information about the value of $\mathrm{Parity}(|\{x : h^*_{t\theta_\star}(x) = +1\}|)$. That is, given $i_t^* = i$, any $\mathcal{Z}_{td}(\theta_\star)$ is conditionally independent from every b_j for $j \neq i$, and is even conditionally independent from b_i when $\mathbb{X}_{td}$ is not completely contained in $\mathcal{X}_i$; specifically, in this case, regardless of b_i, the conditional distribution of $\mathbb{Y}_{td}(\theta_\star)$ given $i_t^* = i$ and given $\mathbb{X}_{td}$ is a product distribution, which deterministically assigns label -1 to those $Y_{tk}(\theta_\star)$ with $X_{tk} \notin \mathcal{X}_i$, and gives uniform random values to the subset of $\mathbb{Y}_{td}(\theta_\star)$ with their respective $X_{tk} \in \mathcal{X}_i$. Finally, letting $r_t = \mathrm{Parity}(|\{k \le d : Y_{tk}(\theta_\star) = +1\}|)$, we note that given $i_t^* = i$, $\mathbb{X}_{td} = \mathcal{X}_i$, and the value r_t, b_i is conditionally independent from $\mathcal{Z}_{td}(\theta_\star)$. Thus, the set of values $C_{iT}(\theta_\star) = \{r_t : i_t^* = i, \mathbb{X}_{td} = \mathcal{X}_i\}$ is a sufficient statistic for b_i (hence for p_i). Recall that, when $i_t^* = i$ and $\mathbb{X}_{td} = \mathcal{X}_i$, the value of r_t is equal to C_t, a $\mathrm{Bernoulli}(p_i)$ random variable. Thus, we neither lose nor gain anything (in terms of risk) by restricting ourselves to estimators $\hat{q}_i$ of the type $\hat{q}_i = \hat{q}_i(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star), i_1^*, \ldots, i_T^*) = \hat{q}_i'(C_{iT}(\theta_\star))$, for some $\hat{q}_i'$ [Schervish, 1995]: that is, estimators that are a function of the $N_{iT}(\theta_\star) = |C_{iT}(\theta_\star)|$ $\mathrm{Bernoulli}(p_i)$ random variables, which we should note are conditionally i.i.d. given $N_{iT}(\theta_\star)$.

Thus, by (1), for any $n \le T$,

$$\begin{aligned}
&\frac{1}{2}\sum_{b_i \in \{-1,+1\}} \mathbb{E}\left[|\hat{q}_i - p_i| \middle| N_{iT}(\theta_\star) = n\right] \\
&= \frac{1}{2}\sum_{b_i \in \{-1,+1\}} \gamma_m \mathbb{P}\left(\hat{q}_i \neq p_i \middle| N_{iT}(\theta_\star) = n\right) \\
&\ge (\gamma_m/32) \cdot \exp\left\{-128\gamma_m^2 N_i/3\right\}.
\end{aligned}$$

Also note that, for each i, $\mathbb{E}[N_i] = \frac{d!(1/m)^d}{\binom{m}{d}} T \le (d/m)^{2d} T = d^{2d}(2\gamma_m/L)^{2d/\alpha} T$, so that Jensen's inequality, linearity of expectation, and the law of total expectation imply

$$\frac{1}{2}\sum_{b_i \in \{-1,+1\}} \mathbb{E}\left[|\hat{q}_i - p_i|\right] \ge (\gamma_m/32) \cdot \exp\left\{-43(2/L)^{2d/\alpha} d^{2d} \gamma_m^{2+2d/\alpha} T\right\}.$$

Thus, by linearity of the expectation,

$$\begin{aligned}
&\left(\frac{1}{2}\right)^{\binom{m}{d}} \sum_{\mathbf{b} \in \{-1,+1\}^{\binom{m}{d}}} \mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} |\hat{q}_i - p_i|\right] \\
&= \sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} \frac{1}{2} \sum_{b_i \in \{-1,+1\}} \mathbb{E}\left[|\hat{q}_i - p_i|\right] \\
&\ge (\gamma_m/(32 \cdot 2^d)) \cdot \exp\left\{-43(2/L)^{2d/\alpha} d^{2d} \gamma_m^{2+2d/\alpha} T\right\}.
\end{aligned}$$

In particular, taking

$$m = \left\lceil (L/2)^{1/\alpha} \left(43(2/L)^{2d/\alpha} d^{2d} T\right)^{\frac{1}{2(d+\alpha)}} \right\rceil,$$

we have

$$\gamma_m = \Theta\left(\left(43(2/L)^{2d/\alpha} d^{2d} T\right)^{-\frac{\alpha}{2(d+\alpha)}}\right),$$

so that

$$\left(\frac{1}{2}\right)^{\binom{m}{d}} \sum_{\mathbf{b}\in\{-1,+1\}^{\binom{m}{d}}} \mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} |\hat{q}_i - p_i|\right] = \Omega\left(2^{-d}\left(43(2/L)^{2d/\alpha} d^{2d} T\right)^{-\frac{\alpha}{2(d+\alpha)}}\right).$$

In particular, this implies there exists some $\mathbf{b}$ for which

$$\mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} |\hat{q}_i - p_i|\right] = \Omega\left(2^{-d}\left(43(2/L)^{2d/\alpha} d^{2d} T\right)^{-\frac{\alpha}{2(d+\alpha)}}\right).$$

Applying this lower bound to the estimator $\hat{p}_i$ defined above yields the result. □

In the extreme case of allowing arbitrary dependence on the data samples, we merely recover the known results lower bounding the risk of density estimation from i.i.d. samples from a smooth density, as indicated by the following result.

Theorem 3. *For any integer $d \geq 1$, there exists an instance space $\mathcal{X}$, a concept space $\mathbb{C}$ of VC dimension d, a distribution $\mathcal{D}$ over $\mathcal{X}$, and a distribution π_0 over $\mathbb{C}$ such that, for Π_Θ the set of distributions over $\mathbb{C}$ with (L,α)-Hölder smooth density functions with respect to π_0, any sequence of estimators, $\hat{\theta}_T = \hat{\theta}_T(\mathcal{Z}_1(\theta_\star), \ldots, \mathcal{Z}_T(\theta_\star))$ ($T = 1, 2, \ldots$), has*

$$\sup_{\theta_\star \in \Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\|\right] = \Omega\left(T^{-\frac{\alpha}{d+2\alpha}}\right).$$

The proof is a simple reduction from the problem of estimating $\pi_{\theta_\star}$ based on direct access to $h^*_{1\theta_\star}, \ldots, h^*_{T\theta_\star}$, which is essentially equivalent to the standard model of density estimation, and indeed the lower bound in Theorem 3 is a well-known result for density estimation from T i.i.d. samples from a Hölder smooth density in a d-dimensional space [**?**, see e.g.,]]devroye:01.

5 Future Directions

There are several interesting questions that remain open at this time. Can either the lower bound or upper bound be improved in general? If, instead of d samples per task, we instead use $m \geq d$ samples, how does the minimax risk vary with m? Related to this, what is the optimal value of m to optimize the rate of convergence as a function of mT, the total number of samples? More generally, if an estimator is permitted to use N total samples, taken from however many tasks it wishes, what is the optimal rate of convergence as a function of N?

References

Bar-Yossef, Z. (2003). Sampling lower bounds via information theory. *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing* (pp. 335–344).

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

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, *36*, 929–965.

Devroye, L., & Lugosi, G. (2001). *Combinatorial methods in density estimation*. New York, NY, USA: Springer.

Poland, J., & Hutter, M. (2006). MDL convergence speed for Bernoulli sequences. *Statistics and Computing*, *16*, 161–175.

Schervish, M. J. (1995). *Theory of statistics*. New York, NY, USA: Springer.

Vapnik, V. (1982). *Estimation of dependencies based on empirical data*. Springer-Verlag, New York.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*, 117–186.

Yang, L., Hanneke, S., & Carbonell, J. (2011). Identifiability of priors from bounded sample sizes with applications to transfer learning. *24*$^{\text{th}}$ *Annual Conference on Learning Theory*.

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