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

Liu Yang[a], Steve Hanneke[b], Jaime Carbonell[c]

[a] *Yale University, New Haven, CT USA*
[b] *Princeton, NJ USA*
[c] *Carnegie Mellon University, Pittsburgh, PA USA*

**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. We additionally extend this setting to real-valued function, where we establish consistency of an estimator for the prior, and discuss an additional application to a preference elicitation problem in algorithmic economics.

## 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

*Email addresses:* `liu.yang@yale.edu` (Liu Yang), `steve.hanneke@gmail.com` (Steve Hanneke), `jgc@cs.cmu.edu` (Jaime Carbonell)

*Preprint submitted to Theoretical Computer Science* *April 3, 2017*

based on the data from past learning problems [1, 2]. 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, [2] 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 distribution 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 of the estimated distribution to the true one, in terms of the number of learning tasks. 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 [2], 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 that earlier work, but requires several interesting innovations to make the rate of convergence explicit, and to dramatically improve the upper bound implicit in the proofs of those earlier results. 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.

We additionally include an extension of the results of [2] to the setting of real-valued functions, establishing consistency (at a uniform rate) for an estimator of a prior over any VC subgraph class. In addition to the application to transfer learning, analogous to the original work of [2], we also discuss an application of this result to a preference elicitation problem in algorithmic economics, in which we are tasked with allocating items to a sequence of customers to approximately maximize the customers' satisfaction, while permitted access to the customer valuation functions only via value queries.

## 2. The Setting

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space [3] (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 $\mathbb{C}$ [4]. We suppose $\mathbb{C}$ is equipped with its Borel $\sigma$-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 $\rho$ is actually a *metric*, which would follow from appropriate topological conditions on $\mathbb{C}$ relative to $\mathcal{D}$.

For any two probability measures $\mu_1, \mu_2$ on a measurable space $(\Omega, \Sigma)$, define the total variation distance

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

For a set function $\mu$ on a *finite* measurable space $(\Omega, \Sigma)$, we abbreviate $\mu(\omega) = \mu(\{\omega\})$, $\forall \omega \in \Omega$. Let $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$ be a family of probability measures on $\mathbb{C}$ (called *priors*), where $\Theta$ is an arbitrary index set (called the *parameter space*). We suppose there exists a probability measure $\pi_0$ on $\mathbb{C}$ (the *reference measure*) such that every $\pi_\theta$ is absolutely continuous with respect to $\pi_0$, and therefore has a density function $f_\theta$ given by the Radon-Nikodym derivative $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_0}$ [3].

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 $\pi_\theta$. 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 $\mathcal{Z}^t_k(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$ 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 $t^{\text{th}}$ *task*. We will be concerned with sequences of estimators $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_k^1(\theta), \ldots, \mathcal{Z}_k^T(\theta))$, 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 $d$ samples per task. 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, [2] showed that, if $\Pi_\Theta$ 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 proved 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 $\mathbb{C}$, $\Pi_\Theta$, 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 $\Pi_\Theta$ contains two priors $\pi_1$ and $\pi_2$, 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 $\Pi_\Theta$ would limit the number of such pairs present in $\Pi_\Theta$, 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 $(L, \alpha)$-Hölder smooth if

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

**An Example:** The main result (Theorem 1) below applies to *any* choices of $\mathbb{C}$, $\mathcal{D}$, and $\Pi_\Theta$ with densities of the priors in $\Pi_\Theta$ satisfying the above smoothness condition. As a concrete example of this, consider $\mathcal{X} = \mathbb{R}^n$ for some integer $n \geq 2$, let $S_n = \{w \in \mathbb{R}^n : \|w\| = 1\}$, let $h_w(x) = \mathrm{sign}(w \cdot x)$ for any $x \in \mathbb{R}^n$ and $w \in S_n$, and consider the case of $\mathbb{C} = \{h_w : w \in S_n\}$: the class of *homogeneous linear separators*, for which the VC dimension $d$ equals $n$. Then suppose $\mathcal{D}$ is the uniform distribution on $\{x \in \mathbb{R}^n : \|x\| = 1\}$. In this case, one can show that $\rho(h_w, h_{w'}) = \frac{1}{\pi}\cos^{-1}(w \cdot w')$, where (in this context) $\pi$ is the well-known numerical constant representing the ratio of a circle's circumference to its diameter. In particular, for any probability density function $f : S_n \to [0, \infty)$ on $S_n$ with $|f(w) - f(w')| \leq L\|w - w'\|^\alpha$ (where $\|\cdot\|$ denotes the Euclidean norm in this context), there is an induced density $f' : \mathbb{C} \to [0, \infty)$ on $\mathbb{C}$ defined by $f'(h_w) = f(w)$ for each $w \in S_n$, satisfying $|f'(h_w) - f'(h_{w'})| \leq L\|w - w'\|^\alpha = L(2 - 2\cos(\pi\rho(h_w, h_{w'})))^{\alpha/2} \leq L(\pi\rho(h_w, h_{w'}))^\alpha$. Thus, the family of Hölder smooth densities on weight vectors $w$ induces a family of priors over $\mathbb{C}$ satisfying the Hölder smoothness condition expressed above (with just a slightly larger $L$ constant). For instance, this would include such distributions on $w$ as the projection to $S_n$ of any multivariate normal distribution with bounded mean and with well-conditioned covariance matrix (in which case $\alpha = 1$), as well as many other commonly-studied prior distributions.

## 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.[1]

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

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

Before stating the proof of this result, we first present a general technique for estimating probability measures in a given family from i.i.d. samples,

[1]To simplify the theorem statements, throughout this article, we adopt a common variant of $O$-notation, where $g(T) = \tilde{O}(f(T))$ indicates that there exist constants $c$ and $k$ such that $g(T) \leq f(T) c \log^k(T)$, where we count quantities such as $d$, $L$, and $\alpha$ as constants (focusing on the dependence on $T$). The interested reader is referred to the proofs for more-precise bounds.

due to [5, 6]. It will be useful as a core component of the estimator $\hat{\theta}_T$ constructed in the proof of Theorem 1 below, as well as the estimator in Theorem 5 of Section 6. Specifically, fix any nonempty set $\mathcal{M}$ of probability measures $\mu$ on a measurable space $(\Omega, \Sigma)$, and suppose that $\mathcal{M}$ is totally bounded in the total variation distance. Fix any $\varepsilon > 0$, and let $\mathcal{M}_\varepsilon$ denote a minimal subset of $\mathcal{M}$ such that $\sup_{\mu \in \mathcal{M}} \inf_{\mu_\varepsilon \in \mathcal{M}_\varepsilon} \|\mu_\varepsilon - \mu\| \leq \varepsilon$; note that $\mathcal{M}_\varepsilon$ is a finite set, due to the total boundedness assumption for $\mathcal{M}$. For any distinct $\mu, \mu' \in \mathcal{M}$, one can show that $\exists A_{\mu,\mu'} \in \Sigma$ with $\|\mu - \mu'\| = |\mu(A_{\mu,\mu'}) - \mu'(A_{\mu,\mu'})|$: that is, either $A_{\mu,\mu'}$ or its complement achieves the supremum in the definition of $\|\mu - \mu'\|$ (e.g., this follows immediately from the Hahn decomposition theorem; see [7] Theorem 32.1). Without loss of generality, suppose $A_{\mu,\mu'} = A_{\mu',\mu}$ for any distinct $\mu, \mu' \in \mathcal{M}$. Denote by $\mathcal{A}_\varepsilon = \{A_{\mu,\mu'} : \mu, \mu' \in \mathcal{M}_\varepsilon, \mu \neq \mu'\}$. Now fix any $\mu^* \in \mathcal{M}$ and $n \in \mathbb{N}$, let $W_1, W_2, \ldots, W_n$ be independent $\mu^*$-distributed random variables, and for each $A \in \Sigma$, denote $\hat{\mu}(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[W_i \in A]$: that is, $\hat{\mu}$ is the empirical measure. Finally, define $\hat{\mu}_\varepsilon = \operatorname{argmin}_{\mu \in \mathcal{M}_\varepsilon} \max_{A \in \mathcal{A}_\varepsilon} |\mu(A) - \hat{\mu}(A)|$ (breaking ties arbitrarily). This estimator, proposed by [5], may be referred to as a *minimum-distance $\varepsilon$-skeleton estimator* [6]. A guarantee on its performance of the following form is entailed by the analysis of [5]; the variant stated here is taken from [6] with only minor changes. For completeness, we provide a brief proof in Appendix A.

**Lemma 1.** $\mathbb{E}\|\hat{\mu}_\varepsilon - \mu^*\| \leq 3\varepsilon + 2\sqrt{\frac{\ln(|\mathcal{M}_\varepsilon|)}{n}}$.

We are now ready for the proof of Theorem 1.

*Proof of Theorem 1.* By the standard PAC analysis [8, 9], 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 $\gamma$ (under $L_1(\mathcal{D})$). For brevity, we omit the $t$ subscripts and superscripts on quantities such as $\mathcal{Z}_k^t(\theta)$ throughout the following analysis, since the claims hold for any arbitrary value of $t$.

For any $\theta \in \Theta$, let $\pi'_\theta$ denote a (conditional on $X_1, \ldots, X_k$) distribution defined as follows. Let $f'_\theta$ denote the (conditional on $X_1, \ldots, X_k$) density function of $\pi'_\theta$ with respect to $\pi_0$, 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, $\pi'_\theta$ has the same probability mass as $\pi_\theta$ 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 every $h \in \mathbb{C}$ has $f'_\theta(h)$ between the smallest and largest values of $f_\theta(g)$ among $g \in \mathbb{C}$ with $\forall i \leq k, g(X_i) = h(X_i)$; therefore, by the smoothness condition, on the event (of probability greater than $1-\gamma$) that each of these regions has diameter less than $\gamma$, we have $\forall h \in \mathbb{C}, |f_\theta(h) - f'_\theta(h)| < L\gamma^\alpha$. On this event, for any $\theta, \theta' \in \Theta$,

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

Furthermore, since the regions that define $f'_\theta$ and $f'_{\theta'}$ are the same (namely, the partition induced by $X_1, \ldots, X_k$), we have

$$\begin{aligned}(1/2)\int |f'_\theta - f'_{\theta'}|\mathrm{d}\pi_0 = (1/2)\sum_{y_1,\ldots,y_k \in \{-1,+1\}} &|\pi_\theta(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = y_i\}) \\ &- \pi_{\theta'}(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = y_i\})| \\ = \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.&\end{aligned}$$

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 [2], consider any $I \subseteq \{1,\ldots,k\}$, 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 $\|\bar{y}_I - \tilde{y}_I\|_1$ is minimal, subject to the constraint that no $h \in \mathbb{C}$ has $h(\bar{x}_I) = \tilde{y}_I$, has $(1/2)\|\bar{y}_I - \tilde{y}_I\|_1 \leq d+1$; also, 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 $\theta'$, so that

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

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

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

Thus, we have that

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

Note that

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

and by exchangeability, this last line equals

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

[2] 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

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

Thus, it suffices to bound the rate of convergence (in total variation distance) of some estimator of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$. If $N(\varepsilon)$ is the $\varepsilon$-covering number of $\{\mathbb{P}_{\mathcal{Z}_d(\theta)}:\theta\in\Theta\}$, then taking $\hat{\theta}_{T\theta_\star}$ such that $\mathbb{P}_{\mathcal{Z}_d(\hat{\theta}_{T\theta_\star})}$ is the minimum-distance $(\varepsilon/4)$-skeleton estimator of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$ [5, 6], Lemma 1 implies that $\mathbb{E}\|\mathbb{P}_{\mathcal{Z}_d(\hat{\theta}_{T\theta_\star})}-\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}\|\leq\varepsilon$ for any $T\geq(64/\varepsilon^2)\ln(N(\varepsilon/4))$. 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 $(L,\alpha)$-Hölder smooth density will have density within $\varepsilon$ 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 $\Pi_\Theta$. Furthermore, the covering number of $\Pi_\Theta$ upper bounds $N(\varepsilon)$ [2], so that $N(\varepsilon)\leq(1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$.

Therefore, solving $T\geq(64/\varepsilon^2)\ln(N(\varepsilon/4))$ for a smallest satisfying $\varepsilon$, 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 of $\mathbb{E}\|\mathbb{P}_{\mathcal{Z}_d(\hat{\theta}_{T\theta_\star})}-\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}\|$, for $\hat{\theta}_{T\theta_\star}$ defined as above with this particular value of $\varepsilon$. Plugging this rate into the bound on the priors, combined with Jensen's inequality, we have

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

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

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

□

*Remark.* Although the above proof makes use of the minimum-distance skeleton estimator, which is typically not computationally efficient, it is often possible to achieve this same result (for certain families of distributions) using a simpler estimator in its place, such as the maximum likelihood estimator. All we require is that the risk of the estimator converges to 0 at a rate as in Lemma 1. For instance, see [10] for conditions on the family of distributions sufficient for a rate of this type to be true for the maximum likelihood estimator. The same can be said for the uses of the minimum-distance skeleton estimator in the proofs of Theorems 4 and 5 below, and indeed in the latter case, any estimator whose risk converges to 0 at a known rate that is independent of $\theta_\star$ would suffice.

## 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 $\pi_0$ over $\mathbb{C}$ such that, for $\Pi_\Theta$ a set of distributions over $\mathbb{C}$ with $(L, \alpha)$-Hölder smooth density functions with respect to $\pi_0$, any estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_d^1(\theta), \ldots, \mathcal{Z}_d^T(\theta))$ has*

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

*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 Bernoulli($p$) random variables, for each $p \in [0,1]$; then it is known that, for any (possibly nondeterministic) decision rule $\hat{p}_n : \{0,1\}^n \to \{(1+\gamma)/2, (1-\gamma)/2\}$,

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

This easily follows from the results of [11], combined with a result of [12] bounding the KL divergence (see also [13])

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 $(L, \alpha)$-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 $\pi_0$ 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 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 $\pi_{\mathbf{b}}(\{h\}) = (\frac{1}{2})^d \binom{m}{d}^{-1} \sum_{i=1}^{\binom{m}{d}} \mathbb{1}[H \subseteq \mathcal{X}_i](1+\gamma_m b_i)^{\mathrm{Parity}(q)}(1-\gamma_m b_i)^{1-\mathrm{Parity}(q)}$. 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 $\mathcal{D}(\{x : h(x) \neq g(x)\}) \geq 1/m = (2\gamma_m/L)^{1/\alpha}$.

Now we set up the reduction as follows. For any estimator $\hat{\pi}_{T\theta} = \hat{\pi}_T[\mathcal{Z}_d^1(\theta), \ldots, \mathcal{Z}_d^T(\theta)]$, 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 $\hat{\pi}_{T\theta_\star}(\{h_i\}) > (\frac{1}{2})^d / \binom{m}{d}$, let $\hat{b}_i = 2\mathrm{Parity}(d) - 1$, and otherwise $\hat{b}_i = 1 - 2\mathrm{Parity}(d)$. 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

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

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 $\hat{q}_i = \hat{q}_i(\mathcal{Z}_d^1(\theta_\star), \ldots, \mathcal{Z}_d^T(\theta_\star), i_1^*, \ldots, i_T^*)$, where $i_t^*$ is the $i^*$ random variable in the generative model that defines $h_{t\theta_\star}^*$; that is, $i_t^* \sim \mathrm{Uniform}(\{1, \ldots, \binom{m}{d}\})$, $C_t \sim \mathrm{Bernoulli}((1+\gamma_m b_{i_t^*})/2)$, and $h_{t\theta_\star}^* \sim \mathrm{Uniform}(\{h \in \mathbb{C} : \{x : h(x) = +1\} \subseteq \mathcal{X}_{i_t^*}, \mathrm{Parity}(|\{x : h(x) = +1\}|) = C_t\})$, 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 $\{i_t^* : t \leq T\}$, the data can be partitioned into $\binom{m}{d}$ disjoint sets, $\{\{\mathcal{Z}_d^t(\theta_\star) : i_t^* = i\} : i = 1, \ldots, \binom{m}{d}\}$, and we can use only the set $\{\mathcal{Z}_d^t(\theta_\star) : i_t^* = i\}$ 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 $\mathcal{Z}_d^t(\theta_\star)$ 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 $r_t = \mathrm{Parity}(|\{k \leq d : Y_{tk}(\theta_\star) = +1\}|)$, we note that given $i_t^* = i$, $\mathbb{X}_{td} = \mathcal{X}_i$, and the value $r_t$, $b_i$ is conditionally independent from $\mathcal{Z}_d^t(\theta_\star)$. 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 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 $\hat{q}_i = \hat{q}_i(\mathcal{Z}_d^1(\theta_\star), \ldots, \mathcal{Z}_d^T(\theta_\star), i_1^*, \ldots, i_T^*) = \hat{q}_i'(C_{iT}(\theta_\star))$, for some $\hat{q}_i'$ [3]: that is, estimators that are a function of the $N_{iT}(\theta_\star) = |C_{iT}(\theta_\star)|$ Bernoulli$(p_i)$ random variables, which we should note are conditionally i.i.d. given $N_{iT}(\theta_\star)$.

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

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

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

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

Thus, by linearity of the expectation,

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

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 $\hat{p}_i$ above yields the result. $\square$

## 5. Using More Than $d$ Samples Per Task

It is natural to wonder how the above rates might potentially improve if we allow $\hat{\theta}_T$ to depend on more than $d$ samples per data set. To establish limits on such improvements, we note that in the extreme case of allowing the estimator to depend on the full $\mathcal{Z}^t(\theta_\star)$ data sets, we may 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 $\pi_0$ over $\mathbb{C}$ such that, for $\Pi_\Theta$ the set of distributions over $\mathbb{C}$ with $(L,\alpha)$-Hölder smooth density functions with respect to $\pi_0$, any sequence of estimators, $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}^1(\theta),\ldots,\mathcal{Z}^T(\theta))$ $(T = 1, 2, \ldots)$, has*

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

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 [6].

Interestingly, it turns out a general upper bound, matching the above special-case lower bound up to a log factor, follows readily from the details of the proof of Theorem 1. Specifically, we have the following theorem, holding for general $\mathcal{X}$, $\mathbb{C}$, $\mathcal{D}$ as in Section 2.

**Theorem 4.** *For $\Pi_\Theta$ any class of priors on $\mathbb{C}$ having $(L,\alpha)$-Hölder smooth densities, $\forall T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}^1(\theta), \ldots, \mathcal{Z}^T(\theta))$ such that*

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

*Proof.* Lemma 1 of [2] (or Lemma 2 in Appendix B) implies that, $\forall \theta, \theta' \in \Theta$, $\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}^t(\theta)} - \mathbb{P}_{\mathcal{Z}^t(\theta')}\|$. Furthermore, as established in the proof of Theorem 1, the $\varepsilon$-covering number of $\Pi_\Theta$ is at most $(1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$. Denoting by $N_\infty(\varepsilon)$ the $\varepsilon$-covering number of $\{\mathbb{P}_{\mathcal{Z}^t(\theta)} : \theta \in \Theta\}$, these two facts imply that $N_\infty(\varepsilon) \leq (1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$. In particular, this implies that there is a value $\varepsilon = O\left(\left(\frac{L^{d/\alpha}\log(T/L^{d/\alpha})}{T}\right)^{\frac{\alpha}{d+2\alpha}}\right)$ such that $T \geq (64/\varepsilon^2)\ln(N_\infty(\varepsilon/4))$. Therefore, taking $\hat{\theta}_{T\theta_\star}$ such that $\mathbb{P}_{\mathcal{Z}^t(\hat{\theta}_{T\theta_\star})}$ is the minimum-distance $(\varepsilon/4)$-skeleton estimator of $\mathbb{P}_{\mathcal{Z}^t(\theta_\star)}$ based on the i.i.d. samples $\mathcal{Z}^1(\theta_\star), \ldots, \mathcal{Z}^T(\theta_\star)$, Lemma 1 implies $\mathbb{E}\|\mathbb{P}_{\mathcal{Z}^t(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}^t(\theta_\star)}\| \leq \varepsilon$. Using again the fact that $\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}^t(\theta)} - \mathbb{P}_{\mathcal{Z}^t(\theta')}\|$ for all $\theta, \theta' \in \Theta$, the theorem follows. □

## 6. Real-Valued Functions and an Application in Algorithmic Economics

In this section, we present results generalizing the analysis of [2] to classes of real-valued functions. We also present an application of this generalization to a preference elicitation problem.

### *6.1. Consistent Estimation of Priors Over Real-Valued Functions at a Bounded Rate*

In this section, we let $\mathcal{B}$ denote a $\sigma$-algebra on $\mathcal{X} \times \mathbb{R}$, and again let $\mathcal{B}_\mathcal{X}$ denote the corresponding $\sigma$-algebra on $\mathcal{X}$. Also, for measurable functions $h, g : \mathcal{X} \to \mathbb{R}$, let $\rho(h,g) = \int |h - g| \mathrm{d}\mathcal{D}$, where $\mathcal{D}$ is a probability measure on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Let $\mathcal{F}$ be a class of measurable functions $\mathcal{X} \to \mathbb{R}$ with Borel $\sigma$-algebra $\mathcal{B}_\mathcal{F}$ induced by $\rho$. Let $\Theta$ be a nonempty set, and for each $\theta \in \Theta$, let $\pi_\theta$ denote a probability measure on $(\mathcal{F}, \mathcal{B}_\mathcal{F})$. We suppose $\{\pi_\theta : \theta \in \Theta\}$

is totally bounded in total variation distance, and that $\mathcal{F}$ is a uniformly bounded VC subgraph class with pseudodimension $d$. We also suppose $\rho$ is a *metric* when restricted to $\mathcal{F}$.

As above, let $\{X_{ti}\}_{t,i\in\mathbb{N}}$ be independent $\mathcal{D}$-distributed random variables. For each $\theta \in \Theta$, let $\{h^*_{t\theta}\}_{t\in\mathbb{N}}$ be independent $\pi_\theta$-distributed random variables, independent from $\{X_{ti}\}_{t,i\in\mathbb{N}}$. For each $t \in \mathbb{N}$ and $\theta \in \Theta$, let $Y_{ti}(\theta) = h^*_{t\theta}(X_{ti})$ for $i \in \mathbb{N}$, and let $\mathcal{Z}^t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \ldots\}$; for each $k \in \mathbb{N}$, define $\mathcal{Z}^t_k(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$, $\mathbb{X}_{tk} = \{X_{t1}, \ldots, X_{tk}\}$, and define $\mathbb{Y}_{tk}(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$.

We have the following result. The proof parallels that of [2] (who studied the special case of binary functions), with a few important twists (in particular, a significantly different approach in the analogue of their Lemma 3). The details are included in Appendix B.

**Theorem 5.** *There exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}^1_d(\theta_\star), \ldots, \mathcal{Z}^T_d(\theta_\star))$, 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 $\theta_\star \in \Theta$,*

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

Here $R(T, \alpha)$ plays the role of a bound on the loss $\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\|$, holding with high probability $1-\delta(T, \alpha)$, where both $R(T, \alpha)$ and $\delta(T, \alpha)$ shrink to 0 as $T \to \infty$. Since $R$ and $\delta$ are functionally independent of $\theta_\star$, this represents a *uniform consistency* result. To provide the convenience of being able to control the failure probability, for all $T$, we also guarantee that $\delta(T, \alpha)$ is bounded by the arbitrary given value $\alpha$.

### *6.2. Maximizing Customer Satisfaction in Combinatorial Auctions*

Theorem 5 has a clear application in the context of transfer learning, following analogous arguments to those given in the special case of binary classification by [2]. Indeed, the transfer learning algorithm, theorem, and proof from [2] (Theorem 3 there) all remain valid (almost verbatim) with the generalization to VC subgraph classes of real-valued functions $\mathcal{X} \to [0, 1]$ under the absolute loss (though it could easily be generalized to most other bounded losses), simply by using our more-general Theorem 5 in place of the analogous result for binary classifiers established by [2] (Theorem 1 there), and using the standard analysis of sample-consistent functions from VC subgraph classes for the initial phase of the algorithm there. The interested reader is referred to the original paper for the details of that algorithm, theorem, and proof. In addition to the application to transfer learning, in the present work we develop a new application of Theorem 5 in the context of the following problem in algorithmic economics, where the objective is to serve a sequence of customers so as to maximize their satisfaction.

Consider an online travel agency, where customers go to the site with some idea of what type of travel they are interested in; the site then poses a series of questions to each customer, and identifies a travel package that best suits their desires, budget, and dates. There are many options of travel packages, with options on location, site-seeing tours, hotel and room quality, etc. Because of this, serving the needs of an *arbitrary* customer might be a lengthy process, requiring many detailed questions. Fortunately, the stream of customers is typically not a worst-case sequence, and in particular obeys many statistical regularities: in particular, it is not too far from reality to think of the customers as being independent and identically distributed samples. With this assumption in mind, it becomes desirable to identify some of these statistical regularities so that we can pose the questions that are typically most relevant, and thereby more quickly identify the travel package that best suits the needs of the typical customer. One straightforward way to do this is to directly *estimate* the distribution of customer value functions, and optimize the questioning system to minimize the expected number of questions needed to find a suitable travel package.

One can model this problem in the style of Bayesian combinatorial auctions, in which each customer has a value function for each possible bundle of items. However, it is slightly different, in that we do not assume the distribution of customers is known, but rather are interested in estimating this distribution; the obtained estimate can then be used in combination with methods based on Bayesian decision theory. In contrast to the literature on Bayesian auctions (and subjectivist Bayesian decision theory in general), this technique is able to maintain general guarantees on performance that hold under an objective interpretation of the problem, rather than merely guarantees holding under an arbitrary assumed prior belief. This general idea is sometimes referred to as *Empirical Bayesian* decision theory in the machine learning and statistics literatures. The ideal result for an Empirical Bayesian algorithm is to be competitive with the corresponding Bayesian methods based on the *actual* distribution of the data (assuming the data are random, with an unknown distribution); that is, although the Empirical Bayesian methods only operate with a data-based estimate of the distribution, the aim is to perform nearly as well as methods based on the true (unobservable) distribution. In this work, we present results of this type, in the context of an abstraction of the aforementioned online travel agency problem, where the measure of performance is the expected number of questions to find a suitable package.

The specific application we are interested in here may be expressed abstractly as a kind of combinatorial auction with preference elicitation. Specifically, we suppose there is a collection of items on a menu, and each possible bundle of items has an associated fixed price. There is a stream of customers, each with a valuation function that provides a value for each possible bundle of items. The objective is to serve each customer a bundle of items that nearly-maximizes his or her surplus value (value minus price). However, we are not permitted direct observation of the customer valuation functions; rather, we may query for the value of any given bundle of items; this is referred to as a *value query* in the literature on preference elicitation in combinatorial auctions (see Chapter 14 of [14], [15]). The objective is to achieve this near-maximal surplus guarantee, while making only a small number of queries per customer. We suppose the customer valuation function are sampled i.i.d. according to an unknown distribution over a known (but arbitrary) class of real-valued functions having finite pseudo-dimension. Reasoning that knowledge of this distribution should allow one to make a smaller number of value queries per customer, we are interested in estimating this unknown distribution, so that as we serve more and more customers, the number of queries per customer required to identify a near-optimal bundle should decrease. In this context, we in fact prove that in the limit, the expected number of queries per customer converges to the number required of a method having direct knowledge of the true distribution of valuation functions.

Formally, suppose there is a menu of $n$ items $[n] = \{1, \ldots, n\}$, and each bundle $B \subseteq [n]$ has an associated price $p(B) \geq 0$. Suppose also there is a sequence of customers, each with a valuation function $v_t : 2^{[n]} \to \mathbb{R}$. We suppose these $v_t$ functions are i.i.d. samples. We can then calculate the satisfaction function for each customer as $s_t(x)$, where $x \in \{0,1\}^n$, and $s_t(x) = v_t(B_x) - p(B_x)$, where $B_x \subseteq [n]$ contains element $i \in [n]$ iff $x_i = 1$.

Now suppose we are able to ask each customer a number of questions before serving up a bundle $B_{\hat{x}_t}$ to that customer. More specifically, we are able to ask for the value $s_t(x)$ for any $x \in \{0,1\}^n$. This is referred to as a *value query* in the literature on preference elicitation in combinatorial auctions (see

Chapter 14 of [14], [15]). We are interested in asking as few questions as possible, while satisfying the guarantee that $\mathbb{E}[s_t(\hat{x}_t) - \max_x s_t(x)] \leq \varepsilon$.

Now suppose, for every $\pi$ and $\varepsilon$, we have a method $A(\pi, \varepsilon)$ such that, given that $\pi$ is the actual distribution of the $s_t$ functions, $A(\pi, \varepsilon)$ guarantees that the $\hat{x}_t$ value it selects has $\mathbb{E}[\max_x s_t(x) - s_t(\hat{x}_t)] \leq \varepsilon$; also let $\hat{N}_t(\pi, \varepsilon)$ denote the actual (random) number of queries the method $A(\pi, \varepsilon)$ would ask for the $s_t$ function, and let $Q(\pi, \varepsilon) = \mathbb{E}[\hat{N}_t(\pi, \varepsilon)]$. We suppose the method never queries any $s_t(x)$ value twice for a given $t$, so that its number of queries for any given $t$ is bounded.

Also suppose $\mathcal{F}$ is a VC subgraph class of functions mapping $\mathcal{X} = \{0,1\}^n$ into $[-1, 1]$ with pseudodimension $d$, and that $\{\pi_\theta : \theta \in \Theta\}$ is a known totally bounded family of distributions over $\mathcal{F}$ such that the $s_t$ functions have distribution $\pi_{\theta_\star}$ for some unknown $\theta_\star \in \Theta$. For any $\theta \in \Theta$ and $\gamma > 0$, let $\mathrm{B}(\theta, \gamma) = \{\theta' \in \Theta : \|\pi_\theta - \pi_{\theta'}\| \leq \gamma\}$.

Suppose, in addition to $A$, we have another method $A'(\varepsilon)$ that is not $\pi$-dependent, but still provides the $\varepsilon$-correctness guarantee, and makes a bounded number of queries (e.g., in the worst case, we could consider querying all $2^n$ points, but in most cases there are more clever $\pi$-independent methods that use far fewer queries, such as $O(1/\varepsilon^2)$). Consider the method described in Algorithm 1; the quantities $\hat{\theta}_{T\theta_\star}$, $R(T, \alpha)$, and $\delta(T, \alpha)$ from Theorem 5 are here considered with respect to $\mathcal{D}$ taken as the uniform distribution on $\{0,1\}^n$.

---

**Algorithm 1** An algorithm for sequentially maximizing expected customer satisfaction.

---

**for** $t = 1, 2, \ldots, T$ **do**
  Pick points $X_{t1}, X_{t2}, \ldots, X_{td}$ uniformly at random from $\{0,1\}^n$
  **if** $R(t-1, \varepsilon/2) > \varepsilon/8$ **then**
    Run $A'(\varepsilon)$
    Take $\hat{x}_t$ as the returned value
  **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
    $Q(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4) \leq \min_{\theta \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1,\varepsilon/2)\right)} Q(\pi_\theta, \varepsilon/4) + \frac{1}{t}$
    Run $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$ and let $\hat{x}_t$ be its return value
  **end if**
**end for**

---

The following theorem indicates that this method is correct, and furthermore that the long-run average number of queries is not much worse than that of a method that has direct knowledge of $\pi_{\theta_\star}$. The proof of this result parallels that of [2] for the transfer learning setting, but is included here for completeness.

**Theorem 6.** *In Algorithm 1,* $\forall t \leq T, \mathbb{E}[\max_x s_t(x) - s_t(\hat{x}_t)] \leq \varepsilon$. *Furthermore, if* $S_T(\varepsilon)$ *is the total number of queries made by the method, then*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq Q(\pi_{\theta_\star}, \varepsilon/4) + d.$$

*Proof.* By Theorem 5, for any $t \leq T$, if $R(t-1, \varepsilon/2) \leq \varepsilon/8$, then with probability at least $1-\varepsilon/2$, $\|\pi_{\theta_\star} - \pi_{\hat{\theta}_{(t-1)\theta_\star}}\| \leq R(t-1, \varepsilon/2)$, so that a triangle inequality implies $\|\pi_{\theta_\star} - \pi_{\check{\theta}_{t\theta_\star}}\| \leq 2R(t-1, \varepsilon/2) \leq \varepsilon/4$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\right] \\
&\qquad \leq \varepsilon/2 + \mathbb{E}\left[\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\middle|\check{\theta}_{t\theta_\star}\right] \mathbb{1}\left[\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/2\right]\right].
\end{aligned}$$

For $\theta \in \Theta$, let $\hat{x}_{t\theta}$ denote the point $x$ that would be returned by $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$ when queries are answered by some $s_{t\theta} \sim \pi_\theta$ instead of $s_t$ (and supposing $s_t = s_{t\theta_\star}$). If $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4$, then

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

Plugging into the above bound, we have $\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\right] \leq \varepsilon$.

For the result on $S_T(\varepsilon)$, first note that $R(t-1, \varepsilon/2) > \varepsilon/8$ only finitely many times (due to $R(t, \alpha) = o(1)$), so that we can ignore those values of $t$ in the asymptotic calculation (as the number of queries is always bounded), and rely on the correctness guarantee of $A'$. For the remaining values $t$, let $N_t$ denote the number of queries made by $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$. Then

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq d + \limsup_{T\to\infty} \sum_{t=1}^{T} \frac{\mathbb{E}\left[N_t\right]}{T}.$$

Since

$$
\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1,\varepsilon/2)]\right] \\
&\le \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} 2^n \mathbb{P}\left(\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1,\varepsilon/2)\right) \\
&\le 2^n \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \delta(t-1,\varepsilon/2) = 0,
\end{aligned}
$$

we have

$$
\limsup_{T\to\infty} \sum_{t=1}^{T} \frac{\mathbb{E}\left[N_t\right]}{T} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \le R(t-1,\varepsilon/2)]\Big].
$$

For $t \le T$, let $N_t(\check{\theta}_{t\theta_\star})$ denote the number of queries $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$ would make if queries were answered with $s_{t\check{\theta}_{t\theta_\star}}$ instead of $s_t$. On the event $\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \le R(t-1,\varepsilon/2)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[N_t \middle| \check{\theta}_{t\theta_\star}\right] \le \mathbb{E}\left[N_t(\check{\theta}_{t\theta_\star}) \middle| \check{\theta}_{t\theta_\star}\right] + 2R(t-1,\varepsilon/2) \\
&\quad = Q(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4) + 2R(t-1,\varepsilon/2) \le Q(\pi_{\theta_\star}, \varepsilon/4) + 2R(t-1,\varepsilon/2) + 1/t.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \le R(t-1,\varepsilon/2)]\right] \\
&\le Q(\pi_{\theta_\star}, \varepsilon/4) + \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} 2R(t-1,\varepsilon/2) + 1/t = Q(\pi_{\theta_\star}, \varepsilon/4).
\end{aligned}
$$

$\square$

In many cases, this result will even continue to hold with an infinite number of goods ($n = \infty$), since Theorem 5 has no dependence on the cardinality of the space $\mathcal{X}$.

## 7. Open Problems

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 $N$?

## Appendix A. Proof of Lemma 1

For completeness, we include a brief proof of Lemma 1. The essential steps of this proof are taken from [6].

*Proof of Lemma 1.* The result is trivially true if $|\mathcal{M}_\varepsilon| = 1$, so let us suppose $|\mathcal{M}_\varepsilon| \geq 2$ for the remainder of the proof. Denote $\mu^*_\varepsilon = \operatorname{argmin}_{\mu \in \mathcal{M}_\varepsilon} \|\mu - \mu^*\|$ (breaking ties arbitrarily). By the triangle inequality,

$$\|\hat{\mu}_\varepsilon - \mu^*\| \leq \|\mu^* - \mu^*_\varepsilon\| + \|\hat{\mu}_\varepsilon - \mu^*_\varepsilon\| \leq \varepsilon + \|\hat{\mu}_\varepsilon - \mu^*_\varepsilon\|.$$

To focus on the nontrivial case, suppose $\hat{\mu}_\varepsilon \neq \mu^*_\varepsilon$. Then we have

$$\|\hat{\mu}_\varepsilon - \mu^*_\varepsilon\| = \left|\hat{\mu}_\varepsilon(A_{\hat{\mu}_\varepsilon, \mu^*_\varepsilon}) - \mu^*_\varepsilon(A_{\hat{\mu}_\varepsilon, \mu^*_\varepsilon})\right| \leq \max_{A \in \mathcal{A}_\varepsilon} \left|\hat{\mu}_\varepsilon(A) - \mu^*_\varepsilon(A)\right|.$$

By the triangle inequality, this is at most

$$\max_{A \in \mathcal{A}_\varepsilon} \left|\hat{\mu}_\varepsilon(A) - \hat{\mu}(A)\right| + \max_{A \in \mathcal{A}_\varepsilon} \left|\mu^*_\varepsilon(A) - \hat{\mu}(A)\right|,$$

and the definition of $\hat{\mu}_\varepsilon$ implies this is at most $2\max_{A \in \mathcal{A}_\varepsilon} |\mu^*_\varepsilon(A) - \hat{\mu}(A)|$. Again by the triangle inequality, this is not larger than

$$2\max_{A \in \mathcal{A}_\varepsilon} \left|\mu^*_\varepsilon(A) - \mu^*(A)\right| + 2\max_{A \in \mathcal{A}_\varepsilon} \left|\hat{\mu}(A) - \mu^*(A)\right| \leq 2\varepsilon + 2\max_{A \in \mathcal{A}_\varepsilon} \left|\hat{\mu}(A) - \mu^*(A)\right|.$$

Altogether, we have that

$$\|\hat{\mu}_\varepsilon - \mu^*\| \leq 3\varepsilon + 2\max_{A \in \mathcal{A}_\varepsilon} \left|\hat{\mu}(A) - \mu^*(A)\right|.$$

This inequality trivially also holds in the case that $\hat{\mu}_\varepsilon = \mu^*_\varepsilon$. To complete the proof, we take the expectations of both sides, and recall that (by Hoeffding's lemma and the standard maximal inequality for sub-Gaussian random variables; see [6] Lemma 2.2)

$$\mathbb{E} \max_{A \in \mathcal{A}_\varepsilon} |\hat{\mu}(A) - \mu^*(A)| \leq \sqrt{\frac{\ln(2|\mathcal{A}_\varepsilon|)}{2n}} \leq \sqrt{\frac{\ln(|\mathcal{M}_\varepsilon|)}{n}},$$

where we have used the fact that, since $A_{\mu,\mu'} = A_{\mu',\mu}$ for any distinct $\mu, \mu'$, we have $|\mathcal{A}_\varepsilon| \leq \binom{|\mathcal{M}_\varepsilon|}{2} \leq \frac{1}{2}|\mathcal{M}_\varepsilon|^2$. □

## Appendix B. Proofs for Section 6

The proof of Theorem 5 is based on the following sequence of lemmas, which parallel those used by [2] for establishing the analogous result for consistent estimation of priors over binary functions. The last of these lemmas (namely, Lemma 4) requires substantial modifications to the original argument of [2]; the others use arguments more-directly based on those of [2].

**Lemma 2.** *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 $h \in \mathcal{F}$, let $c_{\mathbb{X}}(h) = \{(X_{t1}, h(X_{t1})), (X_{t2}, h(X_{t2})), \ldots\}$.

For $h, g \in \mathcal{F}$, define $\rho_{\mathbb{X}}(h, g) = \lim\limits_{m \to \infty} \frac{1}{m} \sum_{i=1}^m |h(X_{ti}) - g(X_{ti})|$ (if the limit exists), and $\rho_{\mathbb{X}_k}(h, g) = \frac{1}{k} \sum_{i=1}^k |h(X_{ti}) - g(X_{ti})|$. Note that since $\mathcal{F}$ is a uniformly bounded VC subgraph class, so is the collection of functions $\{|h - g| : h, g \in \mathcal{F}\}$, so that the uniform strong law of large numbers implies that with probability one, $\forall h, g \in \mathcal{F}$, $\rho_{\mathbb{X}}(h, g)$ exists and has $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ [4].

Consider any $\theta, \theta' \in \Theta$, and any $A \in \mathcal{B}_{\mathcal{F}}$. Then any $h \notin A$ has $\forall g \in A$, $\rho(h, g) > 0$ (by the metric assumption). Thus, if $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ for all $h, g \in \mathcal{F}$, then $\forall h \notin A$,

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

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 $\theta'$.

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 $C' \in \mathcal{B}_{\mathcal{F}} \otimes \mathcal{B}_{\mathcal{X}}^{\infty}$. 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 $\theta'$. Then

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

Since $h^*_{t\theta}$ and $\mathbb{X}$ are independent, $\forall A \in \mathcal{B}_{\mathcal{F}}$, $\pi_\theta(A) = \mathbb{P}_{h^*_{t\theta}}(A) = \mathbb{P}_{h^*_{t\theta}}(A)\mathbb{P}_{\mathbb{X}}(\mathcal{X}^\infty)$ $= \mathbb{P}_{(h^*_{t\theta}, \mathbb{X})}(A \times \mathcal{X}^\infty)$. Analogous reasoning holds for $h^*_{t\theta'}$. Thus, we have

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

Altogether, we have $\|\mathbb{P}_{\mathcal{Z}^t(\theta)} - \mathbb{P}_{\mathcal{Z}^t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|$. ☐

**Lemma 3.** *There exists a sequence $r_k = o(1)$ such that, $\forall t, k \in \mathbb{N}$, $\forall \theta, \theta' \in \Theta$,*

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

*Proof.* This proof follows identically to a proof of [2], but is included here for completeness. Since $\mathbb{P}_{\mathcal{Z}^t_k(\theta)}(A) = \mathbb{P}_{\mathcal{Z}^t(\theta)}(A \times (\mathcal{X} \times \mathbb{R})^\infty)$ for all measurable $A \subseteq (\mathcal{X} \times \mathbb{R})^k$, and similarly for $\theta'$, we have

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{Z}^t_k(\theta)} - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}\| = \sup_{A \in \mathcal{B}^k} \mathbb{P}_{\mathcal{Z}^t_k(\theta)}(A) - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}(A) \\
&= \sup_{A \in \mathcal{B}^k} \mathbb{P}_{\mathcal{Z}^t(\theta)}(A \times (\mathcal{X} \times \mathbb{R})^\infty) - \mathbb{P}_{\mathcal{Z}^t(\theta')}(A \times (\mathcal{X} \times \mathbb{R})^\infty) \\
&\leq \sup_{A \in \mathcal{B}^\infty} \mathbb{P}_{\mathcal{Z}^t(\theta)}(A) - \mathbb{P}_{\mathcal{Z}^t(\theta')}(A) = \|\mathbb{P}_{\mathcal{Z}^t(\theta)} - \mathbb{P}_{\mathcal{Z}^t(\theta')}\|,
\end{aligned}$$

which implies the left inequality when combined with Lemma 2.

Next, we focus on the right inequality. Fix $\theta, \theta' \in \Theta$ and $\gamma > 0$, and let $B \in \mathcal{B}^\infty$ be 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 $\mathcal{A} = \{A \times (\mathcal{X} \times \mathbb{R})^\infty : A \in \mathcal{B}^k, k \in \mathbb{N}\}$. Note that $\mathcal{A}$ is an algebra that generates $\mathcal{B}^\infty$. Thus, Carathéodory's extension theorem (specifically, the version presented by [3]) 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

$$\mathbb{P}_{\mathcal{Z}^t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}^t(\theta')}(B) < \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}^t(\theta)}(A_i) - \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}^t(\theta')}(A_i) + \gamma.$$

Since these $A_i$ sets are disjoint, each of these sums is bounded by a probability value, which implies that there exists some $n \in \mathbb{N}$ such that

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

which implies

$$\begin{aligned}\sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}^t(\theta)}(A_i) - \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}^t(\theta')}(A_i) &< \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}^t(\theta)}(A_i) - \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}^t(\theta')}(A_i) \\ &= \gamma + \mathbb{P}_{\mathcal{Z}^t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}^t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right).\end{aligned}$$

As $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, there exists $m \in \mathbb{N}$ and measurable $B_m \in \mathcal{B}^m$ such that $\bigcup_{i=1}^{n} A_i = B_m \times (\mathcal{X} \times \mathbb{R})^\infty$, and therefore

$$\begin{aligned}\mathbb{P}_{\mathcal{Z}^t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}^t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}_{\mathcal{Z}^t_m(\theta)}(B_m) - \mathbb{P}_{\mathcal{Z}^t_m(\theta')}(B_m) \\ &\le \|\mathbb{P}_{\mathcal{Z}^t_m(\theta)} - \mathbb{P}_{\mathcal{Z}^t_m(\theta')}\| \le \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}^t_k(\theta)} - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}\|.\end{aligned}$$

Combining the above, we have $\|\pi_\theta - \pi_{\theta'}\| \le \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}^t_k(\theta)} - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}\| + 3\gamma$. By letting $\gamma$ approach 0, we have

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

So there exists a sequence $r_k(\theta, \theta') = o(1)$ such that

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| + r_k(\theta, \theta').$$

Now let $\gamma > 0$ and let $\Theta_\gamma$ be a minimal $\gamma$-cover of $\Theta$. Define the quantity $r_k(\gamma) = \max_{\theta,\theta' \in \Theta_\gamma} r_k(\theta, \theta')$. Then for any $\theta, \theta' \in \Theta$, let $\theta_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_\theta - \pi_{\theta''}\|$ and $\theta'_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_{\theta'} - \pi_{\theta''}\|$. Then a triangle inequality implies that $\forall k \in \mathbb{N}$,

$$\begin{aligned}
&\|\pi_\theta - \pi_{\theta'}\| \leq \|\pi_\theta - \pi_{\theta_\gamma}\| + \|\pi_{\theta_\gamma} - \pi_{\theta'_\gamma}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\| \\
&< 2\gamma + r_k(\theta_\gamma, \theta'_\gamma) + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\| \leq 2\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\|.
\end{aligned}$$

Triangle inequalities and the left inequality from the lemma statement (already established) imply

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\| \leq \|\mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta)}\| + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| \\
&\leq \|\pi_{\theta_\gamma} - \pi_\theta\| + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\| < 2\gamma + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\|.
\end{aligned}$$

So in total we have

$$\|\pi_\theta - \pi_{\theta'}\| \leq 4\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\|.$$

Since this holds for all $\gamma > 0$, defining $r_k = \inf_{\gamma > 0}(4\gamma + r_k(\gamma))$, we have the right inequality of the lemma statement. Furthermore, since each $r_k(\theta, \theta') = o(1)$, and $|\Theta_\gamma| < \infty$, we have $r_k(\gamma) = o(1)$ for each $\gamma > 0$, and thus we also have $r_k = o(1)$. □

**Lemma 4.** $\forall t, k \in \mathbb{N}$, *there exists a monotone function* $M_k(x) = o(1)$ *such that,* $\forall \theta, \theta' \in \Theta$,

$$\|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| \leq M_k\left(\|\mathbb{P}_{\mathcal{Z}_d^t(\theta)} - \mathbb{P}_{\mathcal{Z}_d^t(\theta')}\|\right).$$

*Proof.* Fix any $t \in \mathbb{N}$, 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 $\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_d^t(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^{d-k})$, so that

$$\|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_d^t(\theta)} - \mathbb{P}_{\mathcal{Z}_d^t(\theta')}\|,$$

and therefore the result trivially holds.

Now suppose $k > d$. Fix any $\gamma > 0$, and let $B_{\theta,\theta'} \subseteq (\mathcal{X} \times \mathbb{R})^k$ be a measurable set such that

$$\begin{aligned}\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_{\theta,\theta'}) &\leq \|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| \\ &\leq \mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_{\theta,\theta'}) + \gamma.\end{aligned}$$

By Carathéodory's extension theorem (specifically, the version presented by [3]), there exists a disjoint sequence of sets $\{B_i(\theta,\theta')\}_{i=1}^{\infty}$ such that

$$\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_{\theta,\theta'}) < \gamma + \sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta,\theta')) - \sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta,\theta')),$$

and such that each $B_i(\theta,\theta')$ is representable as follows; for some $\ell_i(\theta,\theta') \in \mathbb{N}$, and sets $C_{ij} = (A_{ij1} \times (-\infty, t_{ij1}]) \times \cdots \times (A_{ijk} \times (-\infty, t_{ijk}])$, for $j \leq \ell_i(\theta,\theta')$, where each $A_{ijp} \in \mathcal{B}_{\mathcal{X}}$, the set $B_i(\theta,\theta')$ is representable as $\bigcup_{s \in S_i} \bigcap_{j=1}^{\ell_i(\theta,\theta')} D_{ijs}$, where $S_i \subseteq \{0, \ldots, 2^{\ell_i(\theta,\theta')} - 1\}$, each $D_{ijs} \in \{C_{ij}, C_{ij}^c\}$, and $s \neq s' \Rightarrow \bigcap_{j=1}^{\ell_i(\theta,\theta')} D_{ijs} \cap \bigcap_{j=1}^{\ell_i(\theta,\theta')} D_{ijs'} = \emptyset$. Since the $B_i(\theta,\theta')$ are disjoint, the above sums are bounded, so that there exists $m_k(\theta,\theta',\gamma) \in \mathbb{N}$ such that every $m \geq m_k(\theta,\theta',\gamma)$ has

$$\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_{\theta,\theta'}) < 2\gamma + \sum_{i=1}^{m} \mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta,\theta')) - \sum_{i=1}^{m} \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta,\theta')),$$

Now define $\tilde{M}_k(\gamma) = \max_{\theta,\theta' \in \Theta_\gamma} m_k(\theta,\theta',\gamma)$. Then for any $\theta,\theta' \in \Theta$, let $\theta_\gamma, \theta'_\gamma \in \Theta_\gamma$ be such that $\|\pi_\theta - \pi_{\theta_\gamma}\| < \gamma$ and $\|\pi_{\theta'} - \pi_{\theta'_\gamma}\| < \gamma$, which implies $\|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)}\| < \gamma$ and $\|\mathbb{P}_{\mathcal{Z}_k^t(\theta')} - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\| < \gamma$ by Lemma 3. Then

$$\begin{aligned}\|\mathbb{P}_{\mathcal{Z}_k^t(\theta)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\| &< \|\mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\| + 2\gamma \\ &\leq \mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)}(B_{\theta_\gamma,\theta'_\gamma}) - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}(B_{\theta_\gamma,\theta'_\gamma}) + 3\gamma \\ &\leq \sum_{i=1}^{\tilde{M}_k(\gamma)} \mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)}(B_i(\theta_\gamma,\theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}(B_i(\theta_\gamma,\theta'_\gamma)) + 5\gamma.\end{aligned}$$

Again, since the $B_i(\theta_\gamma, \theta'_\gamma)$ are disjoint, this equals

$$
\begin{aligned}
&5\gamma + \mathbb{P}_{\mathcal{Z}_k^t(\theta_\gamma)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) - \mathbb{P}_{\mathcal{Z}_k^t(\theta'_\gamma)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) \\
&\leq 7\gamma + \mathbb{P}_{\mathcal{Z}_k^t(\theta)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) \\
&= 7\gamma + \sum_{i=1}^{\tilde{M}_k(\gamma)} \mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta_\gamma, \theta'_\gamma)) \\
&\leq 7\gamma + \tilde{M}_k(\gamma) \max_{i \leq \tilde{M}_k(\gamma)} \left|\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))\right|.
\end{aligned}
$$

Thus, if we can show that each term $\left|\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))\right|$ is bounded by a $o(1)$ function of $\|\mathbb{P}_{\mathcal{Z}_d^t(\theta)} - \mathbb{P}_{\mathcal{Z}_d^t(\theta')}\|$, then the result will follow by substituting this relaxation into the above expression and defining $M_k$ by minimizing the resulting expression over $\gamma > 0$.

Toward this end, let $C_{ij}$ be as above from the definition of $B_i(\theta_\gamma, \theta'_\gamma)$, and note that $I_{B_i(\theta_\gamma, \theta'_\gamma)}$ is representable as a function of the $I_{C_{ij}}$ indicators, so that

$$
\begin{aligned}
&\left|\mathbb{P}_{\mathcal{Z}_k^t(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_k^t(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))\right| = \|\mathbb{P}_{I_{B_i(\theta_\gamma, \theta'_\gamma)}(\mathcal{Z}_k^t(\theta))} - \mathbb{P}_{I_{B_i(\theta_\gamma, \theta'_\gamma)}(\mathcal{Z}_k^t(\theta'))}\| \\
&\leq \|\mathbb{P}_{(I_{C_{i1}}(\mathcal{Z}_k^t(\theta)), \ldots, I_{C_{i\ell_i(\theta_\gamma, \theta'_\gamma)}}(\mathcal{Z}_k^t(\theta)))} - \mathbb{P}_{(I_{C_{i1}}(\mathcal{Z}_k^t(\theta')), \ldots, I_{C_{i\ell_i(\theta_\gamma, \theta'_\gamma)}}(\mathcal{Z}_k^t(\theta')))}\| \\
&\leq 2^{\ell_i(\theta_\gamma, \theta'_\gamma)} \max_{J \subseteq \{1, \ldots, \ell_i(\theta_\gamma, \theta'_\gamma)\}} \mathbb{E}\left[\left(\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta))\right) \prod_{j \notin J}\left(1 - I_{C_{ij}}(\mathcal{Z}_k^t(\theta))\right)\right. \\
&\qquad\qquad\qquad \left. - \left(\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta'))\right) \prod_{j \notin J}\left(1 - I_{C_{ij}}(\mathcal{Z}_k^t(\theta'))\right)\right] \\
&\leq 2^{\ell_i(\theta_\gamma, \theta'_\gamma)} \sum_{J \subseteq \left\{1, \ldots, 2^{\ell_i(\theta_\gamma, \theta'_\gamma)}\right\}} \left|\mathbb{E}\left[\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta)) - \prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta'))\right]\right| \\
&\leq 4^{\ell_i(\theta_\gamma, \theta'_\gamma)} \max_{J \subseteq \left\{1, \ldots, 2^{\ell_i(\theta_\gamma, \theta'_\gamma)}\right\}} \left|\mathbb{E}\left[\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta)) - \prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_k^t(\theta'))\right]\right|,
\end{aligned}
$$

which is equal

$$4^{\ell_i(\theta_\gamma,\theta'_\gamma)} \max_{J\subseteq\left\{1,\ldots,2^{\ell_i(\theta_\gamma,\theta'_\gamma)}\right\}} \left|\mathbb{P}_{\mathcal{Z}^t_k(\theta)}\left(\bigcap_{j\in J} C_{ij}\right) - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}\left(\bigcap_{j\in J} C_{ij}\right)\right|. \tag{B.1}$$

Note that $\bigcap_{j\in J} C_{ij}$ can be expressed as some $(A_1\times(-\infty,t_1])\times\cdots\times(A_k\times(-\infty,t_k])$, where each $A_p\in\mathcal{B}_{\mathcal{X}}$ and $t_p\in\mathbb{R}$, so that, for

$$\mathcal{C}_k = \{(A_1\times(-\infty,t_1])\times\cdots\times(A_k\times(-\infty,t_k]) : \forall j\leq k, A_j\in\mathcal{B}_{\mathcal{X}}, t_k\in\mathbb{R}\}$$

and $\hat{\ell} = \max_{\theta,\theta'\in\Theta_\gamma}\max_{i\leq\tilde{M}_k(\gamma)}\ell_i(\theta,\theta')$, the expression (B.1) is at most

$$4^{\hat{\ell}}\sup_{C\in\mathcal{C}_k}\left|\mathbb{P}_{\mathcal{Z}^t_k(\theta)}(C) - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}(C)\right|.$$

Next note that for any $C=(A_1\times(-\infty,t_1])\times\cdots\times(A_k\times(-\infty,t_k])\in\mathcal{C}_k$, letting $C_1 = A_1\times\cdots\times A_k$ and $C_2 = (-\infty,t_1]\times\cdots\times(-\infty,t_k]$,

$$\begin{aligned}\mathbb{P}_{\mathcal{Z}^t_k(\theta)}(C) - \mathbb{P}_{\mathcal{Z}^t_k(\theta')}(C) &= \mathbb{E}\left[\left(\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2)\right) I_{C_1}(\mathbb{X}_{tk})\right]\\ &\leq \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2)\right|\right].\end{aligned}$$

For $p\in\{1,\ldots,k\}$, let $C_{2p} = (-\infty,t_p]$. Then note that, by definition of $d$, for any given $x=(x_1,\ldots,x_k)$, the class $\mathcal{H}_x = \{x_p\mapsto I_{C_{2p}}(h(x_p)) : h\in\mathcal{F}\}$ is a VC class over $\{x_1,\ldots,x_k\}$ with VC dimension at most $d$. Furthremore, we have

$$\begin{aligned}&\left|\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2)\right|\\ &\quad= \Big|\mathbb{P}_{(I_{C_{21}}(h^*_{t\theta}(X_{t1})),\ldots,I_{C_{2k}}(h^*_{t\theta}(X_{tk})))|\mathbb{X}_{tk}}(\{(1,\ldots,1)\})\\ &\qquad - \mathbb{P}_{(I_{C_{21}}(h^*_{t\theta'}(X_{t1})),\ldots,I_{C_{2k}}(h^*_{t\theta'}(X_{tk})))|\mathbb{X}_{tk}}(\{(1,\ldots,1)\})\Big|.\end{aligned}$$

Therefore, the results of [2] (in the proof of their Lemma 3) imply that

$$\begin{aligned}&\left|\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2)\right|\\ &\leq 2^k\max_{y\in\{0,1\}^d}\max_{D\in\{1,\ldots,k\}^d}\Big|\mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\})\\ &\qquad - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\})\Big|.\end{aligned}$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2)\right|\right] \\
&\leq 2^k \mathbb{E}\Bigg[\max_{y\in\{0,1\}^d} \max_{D\in\{1,\ldots,k\}^d} \Big|\mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\}) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\})\Big|\Bigg] \\
&\leq 2^k \sum_{y\in\{0,1\}^d} \sum_{D\in\{1,\ldots,k\}^d} \mathbb{E}\Bigg[\Big|\mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\}) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\})\Big|\Bigg] \\
&\leq 2^{d+k} k^d \max_{y\in\{0,1\}^d} \max_{D\in\{1,\ldots,k\}^d} \mathbb{E}\Bigg[\Big|\mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\}) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j\in D}|\{X_{tj}\}_{j\in D}}(\{y\})\Big|\Bigg].
\end{aligned}
$$

Exchangeability implies this is at most

$$
\begin{aligned}
&2^{d+k} k^d \max_{y\in\{0,1\}^d} \sup_{t_1,\ldots,t_d\in\mathbb{R}} \mathbb{E}\Bigg[\Big|\mathbb{P}_{\{I_{(-\infty,t_j]}(h^*_{t\theta}(X_{tj}))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\}) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{\{I_{(-\infty,t_j]}(h^*_{t\theta'}(X_{tj}))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\})\Big|\Bigg] \\
&= 2^{d+k} k^d \max_{y\in\{0,1\}^d} \sup_{t_1,\ldots,t_d\in\mathbb{R}} \mathbb{E}\Bigg[\Big|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\}) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\})\Big|\Bigg].
\end{aligned}
$$

[2] argue that for all $y \in \{0,1\}^d$ and $t_1, \ldots, t_d \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\}) - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d|\mathbb{X}_{td}}(\{y\})\right|\right] \\
&\leq 4\sqrt{\|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d,\mathbb{X}_{td}} - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d,\mathbb{X}_{td}}\|}.
\end{aligned}
$$

Noting that

$$\|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^{d},\mathbb{X}_{td}} - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^{d},\mathbb{X}_{td}}\| \leq \|\mathbb{P}_{\mathcal{Z}_d^t(\theta)} - \mathbb{P}_{\mathcal{Z}_d^t(\theta')}\|$$

completes the proof. $\square$

We are now ready for the proof of Theorem 5.

*Proof of Theorem 5.* From Lemma 1, for any $\varepsilon > 0$, if $N(\varepsilon/4)$ is the $(\varepsilon/4)$-covering number of $\{\mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)} : \theta \in \Theta\}$, then for $T_\varepsilon = (64/\varepsilon^2)\ln(N(\varepsilon/4))$, for any $T \geq T_\varepsilon$, taking $\hat{\theta}_{T\theta_\star}$ so that $\mathbb{P}_{\mathcal{Z}_d^t(\hat{\theta}_{T\theta_\star})}$ is the minimum-distance $(\varepsilon/4)$-skeleton estimator of $\mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)}$ [5, 6] yields

$$\mathbb{E}\left[\|\mathbb{P}_{\mathcal{Z}_d^t(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)}\|\right] \leq \varepsilon.$$

Thus, letting $G_T = \inf\{\varepsilon > 0 : T \geq T_\varepsilon\}$, if we define $\hat{\theta}_{T\theta_\star}$ so that $\mathbb{P}_{\mathcal{Z}_d^t(\hat{\theta}_{T\theta_\star})}$ is the minimum-distance $(G_T/2)$-skeleton estimator of $\mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)}$, we have

$$\mathbb{E}\left[\|\mathbb{P}_{\mathcal{Z}_d^t(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)}\|\right] \leq 2G_T = o(1).$$

Letting $R'(T,\alpha)$ be any positive sequence with $2G_T \ll R'(T,\alpha) \ll 1$ and $R'(T,\alpha) \geq 2G_T/\alpha$, and letting $\delta(T,\alpha) = 2G_T/R'(T,\alpha) = o(1)$, Markov's inequality implies

$$\mathbb{P}\left(\|\mathbb{P}_{\mathcal{Z}_d^t(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_d^t(\theta_\star)}\| > R'(T,\alpha)\right) \leq \delta(T,\alpha) \leq \alpha. \tag{B.2}$$

Letting $R(T,\alpha) = \min_k \left(M_k\left(R'(T,\alpha)\right) + r_k\right)$, since $R'(T,\alpha) = o(1)$ and $r_k = o(1)$, we have $R(T,\alpha) = o(1)$. Furthermore, composing (B.2) with Lemmas 2, 3, and 4, we have

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

$\square$


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

[2] L. Yang, S. Hanneke, J. Carbonell, A theory of transfer learning with applications to active learning, Machine Learning 90 (2) (2013) 161–189.

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

[4] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Applications 16 (1971) 264–280.

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

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

[7] P. Billingsley, Probability and Measure, 3rd Edition, John Wiley & Sons, 1995.

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

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, Journal of the Association for Computing Machinery 36 (4) (1989) 929–965.

[10] S. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, 2000.

[11] Z. Bar-Yossef, Sampling lower bounds via information theory, in: Proceedings of the 35th Annual ACM Symposium on the Theory of Computing, 2003, pp. 335–344.

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

[13] A. Wald, Sequential tests of statistical hypotheses, The Annals of Mathematical Statistics 16 (2) (1945) 117–186.

[14] P. Cramton, Y. Shoham, R. Steinberg, Combinatorial Auctions, The MIT Press, 2006.

[15] M. Zinkevich, A. Blum, T. Sandholm, On polynomial-time preference elicitation with value queries, in: Proceedings of the 4th ACM Conference on Electronic Commerce, 2003, pp. 175–185.