SURROGATE LOSSES IN PASSIVE AND ACTIVE LEARNING

BY STEVE HANNEKE AND LIU YANG

Carnegie Mellon University

Active learning is a type of sequential design for supervised machine learning, in which the learning algorithm sequentially requests the labels of selected instances from a large pool of unlabeled data points. The objective is to produce a classifier of relatively low risk, as measured under the 0-1 loss, ideally using fewer label requests than the number of random labeled data points sufficient to achieve the same. This work investigates the potential uses of surrogate loss functions in the context of active learning. Specifically, it presents an active learning algorithm based on an arbitrary classification-calibrated surrogate loss function, along with an analysis of the number of label requests sufficient for the classifier returned by the algorithm to achieve a given risk under the 0-1 loss. Interestingly, these results cannot be obtained by simply optimizing the surrogate risk via active learning to an extent sufficient to provide a guarantee on the 0-1 loss, as is common practice in the analysis of surrogate losses for passive learning. Some of the results have additional implications for the use of surrogate losses in passive learning.

1. Introduction. In supervised machine learning, we are tasked with learning a classifier whose probability of making a mistake (i.e., error rate) is small. The study of when it is possible to learn an accurate classifier via a computationally efficient algorithm, and how to go about doing so, is a subtle and difficult topic, owing largely to nonconvexity of the loss function: namely, the 0-1 loss. While there is certainly an active literature on developing computationally efficient methods that succeed at this task, even under various noise conditions, it seems fair to say that at present, many of these advances have not yet reached the level of robustness, efficiency, and simplicity required for most applications. In the mean time, practitioners have turned to various heuristics in the design of practical learning methods, in attempts to circumvent these tough computational problems. One of the most common such heuristics is the use of a convex *surrogate* loss function in place of the 0-1 loss in various optimizations performed by the learning method. The convexity of the surrogate loss allows these optimizations to be performed efficiently, so that the methods can be applied within a reasonable execution time, even with only modest computational resources. Although classifiers arrived at in this

AMS 2000 subject classifications: Primary 62L05, 68Q32, 62H30, 68T05; secondary 68T10, 68Q10, 68Q25, 68W40, 62G99

Keywords and phrases: active learning, sequential design, selective sampling, statistical learning theory, surrogate loss functions, classification

way are not always guaranteed to be good classifiers when performance is measured under the 0-1 loss, in practice this heuristic has often proven quite effective. In light of this fact, most modern learning methods either explicitly make use of a surrogate loss in the formulation of optimization problems (e.g., SVM), or implicitly optimize a surrogate loss via iterative descent (e.g., AdaBoost). Indeed, the choice of a surrogate loss is often as fundamental a part of the process of approaching a learning problem as the choice of hypothesis class or learning bias. Thus it seems essential that we come to some understanding of how best to make use of surrogate losses in the design of learning methods, so that in the favorable scenario that this heuristic actually does work, we have methods taking full advantage of it.

In this work, we are primarily interested in how best to use surrogate losses in the context of *active learning*, which is a type of sequential design in which the learning algorithm is presented with a large pool of unlabeled data points (i.e., only the covariates are observable), and can sequentially request to observe the labels (response variables) of individual instances from the pool. The objective in active learning is to produce a classifier of low error rate while accessing a smaller number of labels than would be required for a method based on random labeled data points (i.e., *passive learning*) to achieve the same. We take as our starting point that we have already committed to use a given surrogate loss, and we restrict our attention to just those scenarios in which this heuristic actually *does* work. We are then interested in how best to make use of the surrogate loss toward the goal of producing a classifier with relatively small error rate. To be clear, we focus on the case where the minimizer of the surrogate risk also minimizes the error rate, and is contained in our function class.

We construct an active learning strategy based on optimizing the empirical surrogate risk over increasingly focused subsets of the instance space, and derive bounds on the number of label requests the method requires to achieve a given error rate. Interestingly, we find that the basic approach of optimizing the surrogate risk via active learning to a sufficient extent to guarantee small error rate generally does not lead to as strong of results. In fact, the method our results apply to typically *does not* optimize the surrogate risk (even in the limit). The insight leading to this algorithm is that, if we are truly only interested in achieving low 0-1 loss, then once we have identified the *sign* of the optimal function at a given point, we need not optimize the value of the function at that point any further, and can therefore focus the label requests elsewhere. As a byproduct of this analysis, we find this insight has implications for the use of certain surrogate losses in passive learning as well, though to a lesser extent.

Most of the mathematical tools used in this analysis are inspired by recently-developed techniques for the study of active learning [18, 19, 25], in conjunction with the results of Bartlett, Jordan, and McAuliffe [6] bounding the excess error rate in terms of the excess surrogate risk, and the works of Koltchinskii [23] and Bartlett, Bousquet, and Mendelson [7] on localized Rademacher complexity bounds.

1.1. *Related Work.* There are many previous works on the topic of surrogate losses in the context of passive learning. Perhaps the most relevant to our results below are the work of Bartlett, Jordan, and McAuliffe [6] and the related work of Zhang [38]. These develop a general theory for converting results on excess risk under the surrogate loss into results on excess risk under the 0-1 loss. Below, we describe the conclusions of that work in detail, and we build on many of the basic definitions and insights pioneered in these works.

Another related line of research, initiated by Audibert and Tsybakov [2], studies "plug-in rules," which make use of regression estimates obtained by optimizing a surrogate loss, and are then rounded to $\{-1, +1\}$ values to obtain classifiers. They prove results under smoothness assumptions on the actual regression function, which (remarkably) are often *better* than the known results for methods that directly optimize the 0-1 loss. Under similar conditions, Minsker [28] studies an analogous active learning method, which again makes use of a surrogate loss, and obtains improvements in label complexity compared to the passive learning method of Audibert and Tsybakov [2]; again, the results for this method based on a surrogate loss are actually better than those derived from existing active learning methods designed to directly optimize the 0-1 loss. The works of Audibert and Tsybakov [2] and Minsker [28] raise interesting questions about whether the general analyses of methods that optimize the 0-1 loss remain tight under complexity assumptions on the regression function, and potentially also about the design of optimal methods for classification when assumptions are phrased in terms of the regression function.

In the present work, we focus our attention on scenarios where the main purpose of using the surrogate loss is to ease the computational problems associated with minimizing an empirical risk, so that our statistical results are typically strongest when the surrogate loss is the 0-1 loss itself. Thus, in the specific scenarios studied by Minsker [28], our results are generally not optimal; rather, the main strength of our analysis lies in its generality. In this sense, our results are more closely related to those of Bartlett, Jordan, and McAuliffe [6] and Zhang [38] than to those of Audibert and Tsybakov [2] and Minsker [28]. That said, we note that several important elements of the design and analysis of the active learning method below are already present to some extent in the work of Minsker [28].

There are several interesting works on active learning methods that optimize a general loss function. Beygelzimer, Dasgupta, and Langford [8] and Koltchinskii [25] have both proposed active learning methods, and analyzed the number of label requests the methods make before achieving a given excess risk for that loss function. The former method is based on importance weighted sampling, while the latter makes clear an interesting connection to local Rademacher complexities. One natural idea for approaching the problem of active learning with a surrogate loss is to run one of these methods with the surrogate loss. The results of Bartlett, Jordan, and McAuliffe [6] allow us to determine a sufficiently small value γ such that any function with excess surrogate risk at most γ has excess error rate at most ε. Thus, by evaluating the established bounds on the number of label requests sufficient for these active learning methods to achieve excess surrogate risk γ, we immediately have a result on the number of label requests sufficient for them to achieve excess error rate ε. This is a common strategy to constructing and analyzing passive learning algorithms that make use of a surrogate loss. However, as we discuss below, this strategy does not generally lead to the best behavior in active learning, and often will not be much better than simply using a related passive learning method. Instead, we propose a new method that typically does not optimize the surrogate risk, but makes use of it in a different way so as to achieve stronger results when performance is measured under the 0-1 loss.

2. Definitions. Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space, where $\mathcal{X}$ is called the *instance space*; for convenience, we suppose this is a standard Borel space. Let $\mathcal{Y} = \{-1, +1\}$, and equip the space $\mathcal{X} \times \mathcal{Y}$ with its product σ-algebra: $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \otimes 2^{\mathcal{Y}}$. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, let $\mathcal{F}^*$ denote the set of all measurable functions $g : \mathcal{X} \to \bar{\mathbb{R}}$, and let $\mathcal{F} \subseteq \mathcal{F}^*$, where $\mathcal{F}$ is called the *function class*. Throughout, we fix a distribution $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, and we denote by $\mathcal{P}$ the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$. In the analysis below, we make the usual simplifying assumption that the events and functions in the definitions and proofs are indeed measurable. In most cases, this holds under simple conditions on $\mathcal{F}$ and $\mathcal{P}_{XY}$ [see e.g., 34]; when this is not the case, we may turn to outer probabilities. However, we will not discuss these technical issues further.

For any $h \in \mathcal{F}^*$, and any distribution P over $\mathcal{X} \times \mathcal{Y}$, denote the *error rate* by $\mathrm{er}(h; P) = P((x, y) : \mathrm{sign}(h(x)) \neq y)$; when $P = \mathcal{P}_{XY}$, we abbreviate this as $\mathrm{er}(h) = \mathrm{er}(h; \mathcal{P}_{XY})$. Also, let $\eta(X; P)$ be a version of $\mathbb{P}(Y = 1|X)$, for $(X, Y) \sim P$; when $P = \mathcal{P}_{XY}$, abbreviate this as $\eta(X) = \eta(X; \mathcal{P}_{XY})$. In particular, note that $\mathrm{er}(h; P)$ is minimized at any h with $\mathrm{sign}(h(x)) = \mathrm{sign}(\eta(x; P) - 1/2)$ for all $x \in \mathcal{X}$. In this work, we will also be interested in certain conditional distributions and modifications of functions, specified as follows. For any measurable $\mathcal{U} \subseteq \mathcal{X}$ with $\mathcal{P}(\mathcal{U}) > 0$, define the probability measure $\mathcal{P}_{\mathcal{U}}(\cdot) = \mathcal{P}_{XY}(\cdot|\mathcal{U} \times \mathcal{Y}) = \mathcal{P}_{XY}(\cdot \cap \mathcal{U} \times \mathcal{Y})/\mathcal{P}(\mathcal{U})$: that is, $\mathcal{P}_{\mathcal{U}}$ is the conditional distribution of $(X, Y) \sim \mathcal{P}_{XY}$ given that $X \in \mathcal{U}$. Also, for any $h, g \in \mathcal{F}^*$, define the spliced function $h_{\mathcal{U},g}(x) = h(x)\mathbb{1}_{\mathcal{U}}(x) + g(x)\mathbb{1}_{\mathcal{X} \setminus \mathcal{U}}(x)$. For a set $\mathcal{H} \subseteq \mathcal{F}^*$, denote $\mathcal{H}_{\mathcal{U},g} = \{h_{\mathcal{U},g} : h \in \mathcal{H}\}$.

For any $\mathcal{H} \subseteq \mathcal{F}^*$, define the *region of sign-disagreement* $\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x))\}$, and the *region of value-disagreement* $\mathrm{DISF}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}$, and denote by $\overline{\mathrm{DIS}}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H}) \times \mathcal{Y}$ and $\overline{\mathrm{DISF}}(\mathcal{H}) = \mathrm{DISF}(\mathcal{H}) \times \mathcal{Y}$. Additionally, we denote by $[\mathcal{H}] = \{f \in \mathcal{F}^* : \forall x \in \mathcal{X}, \inf_{h \in \mathcal{H}} h(x) \leq f(x) \leq \sup_{h \in \mathcal{H}} h(x)\}$ the minimal bracket set containing $\mathcal{H}$.

Our interest here is learning from data, so let $\mathcal{Z} = \{(X_1, Y_1), (X_2, Y_2), \ldots\}$ denote a sequence of independent $\mathcal{P}_{XY}$-distributed random variables, referred to as the *labeled data* sequence, while $\{X_1, X_2, \ldots\}$ is referred to as the *unlabeled data* sequence. For $m \in \mathbb{N}$, we also denote $\mathcal{Z}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$. Throughout, we will let $\delta \in (0, 1/4)$ denote an arbitrary confidence parameter, which will be referenced in the methods and theorem statements.

The *active learning* protocol is defined as follows. An active learning algorithm is initially permitted access to the sequence $X_1, X_2, \ldots$ of unlabeled data. It may then select an index $i_1 \in \mathbb{N}$ and *request* to observe Y_{i_1}; after observing Y_{i_1}, it may select another index $i_2 \in \mathbb{N}$, request to observe Y_{i_2}, and so on. After a number of such label requests not exceeding some specified budget n, the algorithm halts and returns a function $\hat{h} \in \mathcal{F}^*$. Formally, this protocol specifies a type of mapping that maps the random variable $\mathcal{Z}$ to a function $\hat{h}$, where $\hat{h}$ is conditionally independent of $\mathcal{Z}$ given $X_1, X_2, \ldots$ and $(i_1, Y_{i_1}), (i_2, Y_{i_2}), \ldots, (i_n, Y_{i_n})$, where each i_k is conditionally independent of $\mathcal{Z}$ and $i_{k+1}, \ldots, i_n$ given $X_1, X_2, \ldots$ and $(i_1, Y_{i_1}), \ldots, (i_{k-1}, Y_{i_{k-1}})$.

2.1. *Surrogate Loss Functions for Classification.* Throughout, we let $\ell : \bar{\mathbb{R}} \to [0, \infty]$ denote an arbitrary *surrogate loss function*; we will primarily be interested in functions ℓ that satisfy certain conditions discussed below. To simplify some statements below, it will be convenient to suppose $z \in \mathbb{R} \Rightarrow \ell(z) < \infty$. For any $g \in \mathcal{F}^*$ and distribution P over $\mathcal{X} \times \mathcal{Y}$, let $\mathrm{R}_\ell(g; P) = \mathbb{E}[\ell(g(X)Y)]$, where $(X, Y) \sim P$; in the case $P = \mathcal{P}_{XY}$, abbreviate $\mathrm{R}_\ell(g) = \mathrm{R}_\ell(g; \mathcal{P}_{XY})$. Also define $\bar{\ell} = 1 \vee \sup_{x \in \mathcal{X}} \sup_{h \in \mathcal{F}} \max_{y \in \{-1,+1\}} \ell(yh(x))$; we will generally suppose $\bar{\ell} < \infty$. In practice, this is more often a constraint on $\mathcal{F}$ than on ℓ; that is, we could have ℓ unbounded, but due to some normalization of the functions $h \in \mathcal{F}$, ℓ is bounded on the corresponding set of values.

Throughout this work, we will be interested in loss functions ℓ whose point-wise minimizer necessarily also optimizes the 0-1 loss. This property was nicely characterized by Bartlett, Jordan, and McAuliffe [6] as follows. For $\eta_0 \in [0, 1]$, define $\ell^\star(\eta_0) = \inf_{z \in \bar{\mathbb{R}}}(\eta_0 \ell(z) + (1 - \eta_0)\ell(-z))$, and $\ell^\star_-(\eta_0) = \inf_{z \in \bar{\mathbb{R}} : z(2\eta_0 - 1) \leq 0}(\eta_0 \ell(z) + (1 - \eta_0)\ell(-z))$.

DEFINITION 1. *The loss ℓ is* classification-calibrated *if,* $\forall \eta_0 \in [0, 1] \setminus \{1/2\}$, $\ell^\star_-(\eta_0) > \ell^\star(\eta_0)$. $\diamond$

In our context, for $X \sim \mathcal{P}$, $\ell^\star(\eta(X))$ represents the minimum value of the conditional ℓ-risk at X, so that $\mathbb{E}[\ell^\star(\eta(X))] = \inf_{h \in \mathcal{F}^*} \mathrm{R}_\ell(h)$, while $\ell^\star_-(\eta(X))$ represents the minimum conditional ℓ-risk at X, subject to having a sub-optimal conditional error rate at X: i.e., $\mathrm{sign}(h(X)) \neq \mathrm{sign}(\eta(X) - 1/2)$. Thus, being classification-calibrated implies the minimizer of the conditional ℓ-risk at X necessarily has the same sign as the minimizer of the conditional error rate at X. Since we are only interested here in using ℓ as a reasonable surrogate for the 0-1 loss, throughout the work below we suppose ℓ is classification-calibrated.

Though not strictly necessary for our results below, it will be convenient for us to suppose that, for all $\eta_0 \in [0,1]$, this infimum value $\ell^\star(\eta_0)$ is actually *obtained* as $\eta_0 \ell(z^\star(\eta_0)) + (1-\eta_0)\ell(-z^\star(\eta_0))$ for some $z^\star(\eta_0) \in \bar{\mathbb{R}}$ (not necessarily unique). For instance, this is the case for any nonincreasing right-continuous ℓ, or continuous and convex ℓ, which include most of the cases we are interested in using as surrogate losses anyway. The proofs can be modified in a natural way to handle the general case, simply substituting any z with conditional risk sufficiently close to the minimum value. For any distribution P, denote $f^\star_P(x) = z^\star(\eta(x;P))$ for all $x \in \mathcal{X}$. In particular, note that $f^\star_P$ obtains $\mathrm{R}_\ell(f^\star_P; P) = \inf_{g \in \mathcal{F}^*} \mathrm{R}_\ell(g; P)$. When $P = \mathcal{P}_{XY}$, we abbreviate this as $f^\star = f^\star_{\mathcal{P}_{XY}}$. Furthermore, if ℓ is classification-calibrated, then $\mathrm{sign}(f^\star_P(x)) = \mathrm{sign}(\eta(x;P) - 1/2)$ for all $x \in \mathcal{X}$ with $\eta(x;P) \neq 1/2$, and hence $\mathrm{er}(f^\star_P; P) = \inf_{h \in \mathcal{F}^*} \mathrm{er}(h;P)$ as well.

For any distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $h, g \in \mathcal{F}^*$, define the *loss distance* $\mathrm{D}_\ell(h,g;P) = \sqrt{\mathbb{E}\left[(\ell(h(X)Y) - \ell(g(X)Y))^2\right]}$, where $(X,Y) \sim P$. Also define the *loss diameter* of a class $\mathcal{H} \subseteq \mathcal{F}^*$ as $\mathrm{D}_\ell(\mathcal{H}; P) = \sup_{h,g \in \mathcal{H}} D_\ell(h,g;P)$, and the ℓ-risk ε-minimal set of $\mathcal{H}$ as $\mathcal{H}(\varepsilon; \ell, P) = \{h \in \mathcal{H} : \mathrm{R}_\ell(h;P) - \inf_{g \in \mathcal{H}} \mathrm{R}_\ell(g;P) \leq \varepsilon\}$. When $P = \mathcal{P}_{XY}$, we abbreviate these as $\mathrm{D}_\ell(h,g) = \mathrm{D}_\ell(h,g;\mathcal{P}_{XY})$, $\mathrm{D}_\ell(\mathcal{H}) = \mathrm{D}_\ell(\mathcal{H};\mathcal{P}_{XY})$, and $\mathcal{H}(\varepsilon;\ell) = \mathcal{H}(\varepsilon;\ell,\mathcal{P}_{XY})$. Also, for any $h \in \mathcal{F}^*$, abbreviate $h_{\mathcal{U}} = h_{\mathcal{U},f^\star}$, and for any $\mathcal{H} \subseteq \mathcal{F}^*$, define $\mathcal{H}_{\mathcal{U}} = \{h_{\mathcal{U}} : h \in \mathcal{H}\}$.

We additionally define related quantities for the 0-1 loss, as follows. Define the *distance* $\Delta_P(h,g) = \mathcal{P}(x : \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x)))$ and *radius* $\mathrm{radius}(\mathcal{H};P) = \sup_{h \in \mathcal{H}} \Delta_P(h, f^\star_P)$. Also define the ε-minimal set of $\mathcal{H}$ as $\mathcal{H}(\varepsilon; {\scriptstyle 01}, P) = \{h \in \mathcal{H} : \mathrm{er}(h;P) - \inf_{g \in \mathcal{H}} \mathrm{er}(g;P) \leq \varepsilon\}$, and for $r > 0$, define the r-ball centered at h in $\mathcal{H}$ by $\mathrm{B}_{\mathcal{H},P}(h,r) = \{g \in \mathcal{H} : \Delta_P(h,g) \leq r\}$. When $P = \mathcal{P}_{XY}$, we abbreviate these as $\Delta(h,g) = \Delta_{\mathcal{P}_{XY}}(h,g)$, $\mathrm{radius}(\mathcal{H}) = \mathrm{radius}(\mathcal{H};\mathcal{P}_{XY})$, $\mathcal{H}(\varepsilon; {\scriptstyle 01}) = \mathcal{H}(\varepsilon; {\scriptstyle 01}, \mathcal{P}_{XY})$, and $\mathrm{B}_{\mathcal{H}}(h,r) = \mathrm{B}_{\mathcal{H},\mathcal{P}_{XY}}(h,r)$; when $\mathcal{H} = \mathcal{F}$, further abbreviate $\mathrm{B}(h,r) = \mathrm{B}_{\mathcal{F}}(h,r)$.

We will be interested in transforming results concerning the excess surrogate risk into results on the excess error rate. As such, we will make use of the following abstract transformation.

DEFINITION 2. *For any distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $\varepsilon \in [0,1]$, define*

$$\Gamma_\ell(\varepsilon; P) = \sup\{\gamma > 0 : \mathcal{F}^*(\gamma; \ell, P) \subseteq \mathcal{F}^*(\varepsilon; {\scriptstyle 01}, P)\} \cup \{0\}.$$

Also, for any $\gamma \in [0, \infty)$, define the inverse

$$\mathcal{E}_\ell(\gamma; P) = \inf\{\varepsilon > 0 : \gamma \le \Gamma_\ell(\varepsilon; P)\}.$$

When $P = \mathcal{P}_{XY}$, abbreviate $\Gamma_\ell(\varepsilon) = \Gamma_\ell(\varepsilon; \mathcal{P}_{XY})$ and $\mathcal{E}_\ell(\gamma) = \mathcal{E}_\ell(\gamma; \mathcal{P}_{XY})$. ◇

By definition, for classification-calibrated ℓ, Γ_ℓ has the property that

$$\text{(1)} \quad \forall h \in \mathcal{F}^*, \forall \varepsilon \in [0,1], \quad \mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star) < \Gamma_\ell(\varepsilon) \implies \mathrm{er}(h) - \mathrm{er}(f^\star) \le \varepsilon.$$

In fact, Γ_ℓ is defined to be maximal with this property, in that *any* Γ'_ℓ for which (1) is satisfied must have $\Gamma'_\ell(\varepsilon) \le \Gamma_\ell(\varepsilon)$ for all $\varepsilon \in [0,1]$.

In our context, we will typically be interested in calculating lower bounds on Γ_ℓ for any particular scenario of interest. Bartlett, Jordan, and McAuliffe [6] studied various lower bounds of this type. Specifically, for $\zeta \in [-1,1]$, define $\tilde{\psi}_\ell(\zeta) = \ell^\star_-\left(\frac{1+\zeta}{2}\right) - \ell^\star\left(\frac{1+\zeta}{2}\right)$, and let ψ_ℓ be the largest convex lower bound of $\tilde{\psi}_\ell$ on $[0,1]$, which is well-defined in this context [6]. Bartlett, Jordan, and McAuliffe [6] show ψ_ℓ is continuous and nondecreasing on $(0,1)$, and in fact that $x \mapsto \psi_\ell(x)/x$ is nondecreasing on $(0,\infty)$. They also show every $h \in \mathcal{F}^*$ has $\psi_\ell(\mathrm{er}(h) - \mathrm{er}(f^\star)) \le \mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star)$, so that $\psi_\ell \le \Gamma_\ell$, and they find this inequality can be tight for a particular choice of $\mathcal{P}_{XY}$. They further study more subtle relationships between excess ℓ-risk and excess error rate holding for any classification-calibrated ℓ. In particular, following the same argument as in the proof of their Theorem 3, one can show that if ℓ is classification-calibrated, every $h \in \mathcal{F}^*$ satisfies

$$\Delta(h, f^\star) \cdot \psi_\ell\left(\frac{\mathrm{er}(h) - \mathrm{er}(f^\star)}{2\Delta(h, f^\star)}\right) \le \mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star).$$

The implication of this in our context is the following. Fix any nondecreasing function $\Psi_\ell : [0,1] \to [0,\infty)$ such that $\forall \varepsilon \ge 0$,

$$\text{(2)} \qquad \Psi_\ell(\varepsilon) \le \mathrm{radius}(\mathcal{F}^*(\varepsilon; {\scriptstyle 01}))\psi_\ell\left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}^*(\varepsilon; {\scriptstyle 01}))}\right).$$

Any $h \in \mathcal{F}^*$ with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star) < \Psi_\ell(\varepsilon)$ also has $\Delta(h, f^\star)\psi_\ell\left(\frac{\mathrm{er}(h)-\mathrm{er}(f^\star)}{2\Delta(h,f^\star)}\right) < \Psi_\ell(\varepsilon)$; combined with the fact that $x \mapsto \psi_\ell(x)/x$ is nondecreasing on $(0,\infty)$, this implies $\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h) - \mathrm{er}(f^\star); {\scriptstyle 01}))\psi_\ell\left(\frac{\mathrm{er}(h)-\mathrm{er}(f^\star)}{2\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h)-\mathrm{er}(f^\star); {\scriptstyle 01}))}\right) < \Psi_\ell(\varepsilon)$;

this means $\Psi_\ell(\mathrm{er}(h) - \mathrm{er}(f^\star)) < \Psi_\ell(\varepsilon)$, and monotonicity of Ψ_ℓ implies $\mathrm{er}(h) - \mathrm{er}(f^\star) < \varepsilon$. Altogether, this implies $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$. In fact, though we do not present the details here, with only minor modifications to the proofs below, when $f^\star \in \mathcal{F}$, all of our results involving $\Gamma_\ell(\varepsilon)$ will also hold while replacing $\Gamma_\ell(\varepsilon)$ with any nondecreasing Ψ'_ℓ such that $\forall \varepsilon \geq 0$,

$$\Psi'_\ell(\varepsilon) \leq \mathrm{radius}(\mathcal{F}(\varepsilon; {}_{01}))\psi_\ell\left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}(\varepsilon; {}_{01}))}\right), \tag{3}$$

which can sometimes lead to tighter results.

Some of our stronger results below will be stated for a restricted family of losses, originally explored by Bartlett, Jordan, and McAuliffe [6]: namely, smooth losses whose convexity is quantified by a polynomial. Specifically, this restriction is characterized by the following condition.

CONDITION 3. *$\mathcal{F}$ is convex, with $\forall x \in \mathcal{X}, \sup_{f\in\mathcal{F}} |f(x)| \leq \bar{B}$ for some constant $\bar{B} \in (0, \infty)$, and there exists a pseudometric $d_\ell : [-\bar{B}, \bar{B}]^2 \to [0, \bar{d}_\ell]$ for some constant $\bar{d}_\ell \in (0, \infty)$, and constants $L, C_\ell \in (0, \infty)$ and $r_\ell \in (0, \infty]$ such that $\forall x, y \in [-\bar{B}, \bar{B}], |\ell(x) - \ell(y)| \leq L d_\ell(x, y)$ and the function $\bar{\delta}_\ell(\varepsilon) = \inf\left\{\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) : x, y \in [-\bar{B}, \bar{B}], d_\ell(x, y) \geq \varepsilon\right\} \cup \{\infty\}$ satisfies $\forall \varepsilon \in (0, 1)$, $\bar{\delta}_\ell(\varepsilon) \geq C_\ell \varepsilon^{r_\ell}$.* $\diamond$

In particular, note that if $\mathcal{F}$ is convex, the functions in $\mathcal{F}$ are uniformly bounded, and ℓ is continuous, Condition 3 is always satisfied (though possibly with $r_\ell = \infty$).

2.2. *A Few Examples of Loss Functions.* Here we briefly mention a few loss functions ℓ in common practical use, all of which are classification-calibrated. These examples are taken directly from the work of Bartlett, Jordan, and McAuliffe [6], which additionally discusses many other interesting examples of classification-calibrated loss functions and their corresponding ψ_ℓ functions.

Example 1. The *exponential loss* is specified as $\ell(x) = e^{-x}$. This loss function appears in many contexts in machine learning; for instance, the popular AdaBoost method can be viewed as an algorithm that greedily optimizes the exponential loss [13]. Bartlett, Jordan, and McAuliffe [6] show that under the exponential loss, $\psi_\ell(x) = 1 - \sqrt{1 - x^2}$, which is tightly approximated by $x^2/2$ for small x. They also show this loss satisfies the conditions on ℓ in Condition 3 with $d_\ell(x, y) = |x - y|$, $L = e^{\bar{B}}$, $C_\ell = e^{-\bar{B}}/8$, and $r_\ell = 2$.

Example 2. The *hinge loss*, specified as $\ell(x) = \max\{1 - x, 0\}$, is another common surrogate loss in machine learning practice today. For instance, it is used in the objective of the Support Vector Machine (along with a regularization term) [10]. Bartlett, Jordan, and McAuliffe [6] show that for the hinge loss, $\psi_\ell(x) = |x|$.

The hinge loss is Lipschitz continuous, with Lipschitz constant 1. However, for the remaining conditions on ℓ in Condition 3, any $x, y \leq 1$ have $\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) = \ell(\frac{1}{2}x + \frac{1}{2}y)$, so that $\bar{\delta}_\ell(\varepsilon) = 0$; hence, $r_\ell = \infty$ is required.

Example 3. The *quadratic loss* (or squared loss), specified as $\ell(x) = (1 - x)^2$, is often used in so-called *plug-in* classifiers [2], which approach the problem of learning a classifier by estimating the regression function $\mathbb{E}[Y|X = x] = 2\eta(x) - 1$, and then taking the sign of this estimator to get a binary classifier. The quadratic loss has the convenient property that for any distribution P over $\mathcal{X} \times \mathcal{Y}$, $f_P^\star(\cdot) = 2\eta(\cdot; P) - 1$, so that it is straightforward to describe the set of distributions P satisfying the assumption $f_P^\star \in \mathcal{F}$. Bartlett, Jordan, and McAuliffe [6] show that for the quadratic loss, $\psi_\ell(x) = x^2$. They also show the quadratic loss satisfies the conditions on ℓ in Condition 3, with $L = 2(\bar{B} + 1)$, $C_\ell = 1/4$, and $r_\ell = 2$. In fact, they study the general family of losses $\ell(x) = |1 - x|^p$, for $p \in (1, \infty)$, and show that $\psi_\ell(x)$ and r_ℓ exhibit a range of behaviors varying with p.

Example 4. The *truncated quadratic loss* is specified as $\ell(x) = (\max\{1-x, 0\})^2$. Bartlett, Jordan, and McAuliffe [6] show that in this case, $\psi_\ell(x) = x^2$. They also show that, under the pseudometric $d_\ell(a, b) = |\min\{a, 1\} - \min\{b, 1\}|$, the truncated quadratic loss satisfies the conditions on ℓ in Condition 3, with $L = 2(\bar{B}+1)$, $C_\ell = 1/4$, and $r_\ell = 2$.

2.3. *Empirical ℓ-Risk Minimization.* For any $m \in \mathbb{N}$, $g : \mathcal{X} \to \bar{\mathbb{R}}$, and $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, define the *empirical ℓ-risk* as $\mathrm{R}_\ell(g; S) = m^{-1} \sum_{i=1}^m \ell(g(x_i)y_i)$. At times it will be convenient to keep track of the indices for a subsequence of $\mathcal{Z}$, and for this reason we also overload the notation, so that for any $Q = \{(i_1, y_1), \ldots, (i_m, y_m)\} \in (\mathbb{N} \times \mathcal{Y})^m$, we define $S[Q] = \{(X_{i_1}, y_1), \ldots, (X_{i_m}, y_m)\}$ and $\mathrm{R}_\ell(g; Q) = \mathrm{R}_\ell(g; S[Q])$. For completeness, we also generally define $\mathrm{R}_\ell(g; \emptyset) = 0$. The method of empirical ℓ-risk minimization, here denoted by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$, is characterized by the property that it returns $\hat{h} = \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{R}_\ell(h; \mathcal{Z}_m)$. This is a well-studied and classical passive learning method, presently in popular use in applications, and as such it will serve as our baseline for passive learning methods.

2.4. *Localized Sample Complexities.* The derivation of localized excess risk bounds can essentially be motivated as follows. Suppose we are interested in bounding the excess ℓ-risk of $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$. Further suppose we have a coarse guarantee $U_\ell(\mathcal{H}, m)$ on the excess ℓ-risk of the $\hat{h}$ returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$: that is, $\mathrm{R}_\ell(\hat{h}) - \mathrm{R}_\ell(f^\star) \leq U_\ell(\mathcal{H}, m)$. In some sense, this guarantee identifies a set $\mathcal{H}' \subseteq \mathcal{H}$ of functions that a priori have the *potential* to be returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ (namely, $\mathcal{H}' = \mathcal{H}(U_\ell(\mathcal{H}, m); \ell)$), while those in $\mathcal{H} \setminus \mathcal{H}'$ do not. With this information in hand, we can think of $\mathcal{H}'$ as a kind of *effective* function class, and we can then think of $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ as equivalent to $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m)$. We may then repeat this same reasoning for $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m)$, calculating $U_\ell(\mathcal{H}', m)$ to determine a set $\mathcal{H}'' = \mathcal{H}'(U_\ell(\mathcal{H}', m); \ell) \subseteq \mathcal{H}'$ of potential return values for *this* empirical minimizer, so that $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m) = \mathrm{ERM}_\ell(\mathcal{H}'', \mathcal{Z}_m)$, and so on. This repeats until we identify a fixed-point set $\mathcal{H}^{(\infty)}$ of functions such that $\mathcal{H}^{(\infty)}(U_\ell(\mathcal{H}^{(\infty)}, m); \ell) = \mathcal{H}^{(\infty)}$, so that no further reduction is possible. Following this chain of reasoning back to the beginning, we find that $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m) = \mathrm{ERM}_\ell(\mathcal{H}^{(\infty)}, \mathcal{Z}_m)$, so that the function $\hat{h}$ returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ has excess ℓ-risk at most $U_\ell(\mathcal{H}^{(\infty)}, m)$, which may be significantly smaller than $U_\ell(\mathcal{H}, m)$, depending on how refined the original $U_\ell(\mathcal{H}, m)$ bound was.

To formalize this fixed-point argument for $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$, Koltchinskii [23] makes use of the following quantities to define the coarse bound $U_\ell(\mathcal{H}, m)$ [see also 7, 15]. For any $\mathcal{H} \subseteq [\mathcal{F}]$, $m \in \mathbb{N}$, $s \in [1, \infty)$, and any distribution P on $\mathcal{X} \times \mathcal{Y}$, letting $Q \sim P^m$, define

$$\phi_\ell(\mathcal{H}; m, P) = \mathbb{E}\left[\sup_{h,g\in\mathcal{H}} (\mathrm{R}_\ell(h;P) - \mathrm{R}_\ell(g;P)) - (\mathrm{R}_\ell(h;Q) - \mathrm{R}_\ell(g;Q))\right],$$

$$\bar{U}_\ell(\mathcal{H}; P, m, s) = \bar{K}_1 \phi_\ell(\mathcal{H}; m, P) + \bar{K}_2 \mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{s}{m}} + \frac{\bar{K}_3 \bar{\ell} s}{m},$$

$$\tilde{U}_\ell(\mathcal{H}; P, m, s) = \tilde{K}\left(\phi_\ell(\mathcal{H}; m, P) + \mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{s}{m}} + \frac{\bar{\ell} s}{m}\right),$$

where $\bar{K}_1$, $\bar{K}_2$, $\bar{K}_3$, and $\tilde{K}$ are appropriately chosen constants.

We will be interested in having access to these quantities in the context of our algorithms; however, since $\mathcal{P}_{XY}$ is not directly accessible to the algorithm, we will need to approximate these by data-dependent estimators. Toward this end, we define the following quantities, again taken from the work of Koltchinskii [23]. For $\varepsilon > 0$, let $\mathbb{Z}_\varepsilon = \{j \in \mathbb{Z} : 2^j \geq \varepsilon\}$. For any $\mathcal{H} \subseteq [\mathcal{F}]$, $q \in \mathbb{N}$, and $S = \{(x_1, y_1), \ldots, (x_q, y_q)\} \in (\mathcal{X} \times \{-1, +1\})^q$, let $\mathcal{H}(\varepsilon; \ell, S) = \{h \in \mathcal{H} : \mathrm{R}_\ell(h; S) - \inf_{g\in\mathcal{H}} \mathrm{R}_\ell(g; S) \leq \varepsilon\}$; then for any sequence $\Xi = \{\xi_k\}_{k=1}^q \in \{-1, +1\}^q$, and any $s \in [1, \infty)$, define

$$\hat{\phi}_\ell(\mathcal{H}; S, \Xi) = \sup_{h,g\in\mathcal{H}} \frac{1}{q}\sum_{k=1}^q \xi_k \cdot (\ell(h(x_k)y_k) - \ell(g(x_k)y_k)),$$

$$\hat{\mathrm{D}}_\ell(\mathcal{H}; S)^2 = \sup_{h,g\in\mathcal{H}} \frac{1}{q}\sum_{k=1}^q (\ell(h(x_k)y_k) - \ell(g(x_k)y_k))^2,$$

$$\hat{U}_\ell(\mathcal{H}; S, \Xi, s) = 12\hat{\phi}_\ell(\mathcal{H}; S, \Xi) + 34\hat{\mathrm{D}}_\ell(\mathcal{H}; S)\sqrt{\frac{s}{q}} + \frac{752\bar{\ell} s}{q}.$$

For completeness, define $\hat{\phi}_\ell(\mathcal{H};\emptyset,\emptyset) = \hat{\mathrm{D}}_\ell(\mathcal{H};\emptyset) = 0$, and $\hat{U}_\ell(\mathcal{H};\emptyset,\emptyset,s) = 752\bar{\ell}s$.

The above quantities (with appropriate choices of $\bar{K}_1$, $\bar{K}_2$, $\bar{K}_3$, and $\tilde{K}$) can be formally related to each other and to the excess ℓ-risk of functions in $\mathcal{H}$ via the following general result; this variant is due to Koltchinskii [23].

LEMMA 4. *For any $\mathcal{H} \subseteq [\mathcal{F}]$, $s \in [1,\infty)$, distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $m \in \mathbb{N}$, if $Q \sim P^m$ and $\Xi = \{\xi_1,\ldots,\xi_m\} \sim \mathrm{Uniform}(\{-1,+1\})^m$ are independent, and $h^* \in \mathcal{H}$ has $\mathrm{R}_\ell(h^*;P) = \inf_{h\in\mathcal{H}} \mathrm{R}_\ell(h;P)$, then with probability at least $1-6e^{-s}$, the following claims hold.*

$$\forall h \in \mathcal{H}, \mathrm{R}_\ell(h;P) - \mathrm{R}_\ell(h^*;P) \le \mathrm{R}_\ell(h;Q) - \mathrm{R}_\ell(h^*;Q) + \bar{U}_\ell(\mathcal{H};P,m,s),$$
$$\forall h \in \mathcal{H}, \mathrm{R}_\ell(h;Q) - \inf_{g\in\mathcal{H}} \mathrm{R}_\ell(g;Q) \le \mathrm{R}_\ell(h;P) - \mathrm{R}_\ell(h^*;P) + \bar{U}_\ell(\mathcal{H};P,m,s),$$
$$\bar{U}_\ell(\mathcal{H};P,m,s) < \hat{U}_\ell(\mathcal{H};Q,\Xi,s) < \tilde{U}_\ell(\mathcal{H};P,m,s).$$

$\diamond$

We typically expect the $\bar{U}$, $\hat{U}$, and $\tilde{U}$ quantities to be roughly within constant factors of each other. Following Koltchinskii [23] and Giné and Koltchinskii [15], we can use this result to derive localized bounds on the number of samples sufficient for $\mathrm{ERM}_\ell(\mathcal{H},\mathcal{Z}_m)$ to achieve a given excess ℓ-risk. Specifically, for $\mathcal{H} \subseteq [\mathcal{F}]$, distribution P over $\mathcal{X} \times \mathcal{Y}$, values $\gamma,\gamma_1,\gamma_2 \ge 0$, $s \in [1,\infty)$, and any function $\mathfrak{s} : (0,\infty)^2 \to [1,\infty)$, define the following quantities.

$$\bar{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) = \min\left\{m \in \mathbb{N} : \bar{U}_\ell(\mathcal{H}(\gamma_2;\ell,P);P,m,s) < \gamma_1\right\},$$
$$\bar{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}) = \sup_{\gamma'\ge\gamma} \bar{\mathrm{M}}_\ell(\gamma'/2,\gamma';\mathcal{H},P,\mathfrak{s}(\gamma,\gamma')),$$
$$\tilde{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) = \min\left\{m \in \mathbb{N} : \tilde{U}_\ell(\mathcal{H}(\gamma_2;\ell,P);P,m,s) \le \gamma_1\right\},$$
$$\tilde{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}) = \sup_{\gamma'\ge\gamma} \tilde{\mathrm{M}}_\ell(\gamma'/2,\gamma';\mathcal{H},P,\mathfrak{s}(\gamma,\gamma')).$$

These quantities are well-defined for $\gamma_1,\gamma_2,\gamma > 0$ when $\lim_{m\to\infty} \phi_\ell(\mathcal{H};m,P) = 0$. In other cases, for completeness, we define them to be ∞.

In particular, the quantity $\bar{\mathrm{M}}_\ell(\gamma;\mathcal{F},\mathcal{P}_{XY},\mathfrak{s})$ is used in Theorem 6 below to quantify the performance of $\mathrm{ERM}_\ell(\mathcal{F},\mathcal{Z}_m)$. The primary practical challenge in calculating $\bar{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s})$ is handling the $\phi_\ell(\mathcal{H}(\gamma';\ell,P);m,P)$ quantity. In the literature, the typical (only?) way such calculations are approached is by first deriving a bound on $\phi_\ell(\mathcal{H}';m,P)$ for every $\mathcal{H}' \subseteq \mathcal{H}$ in terms of some natural measure of complexity for the full class $\mathcal{H}$ (e.g., entropy numbers) and some very basic measure of complexity for $\mathcal{H}'$: most often $\mathrm{D}_\ell(\mathcal{H}';P)$ and sometimes a seminorm of an envelope function for $\mathcal{H}'$. After this, one then proceeds to bound these basic measures of complexity for the specific subsets $\mathcal{H}(\gamma';\ell,P)$, as a function of γ'. Composing these two results is then sufficient to bound $\phi_\ell(\mathcal{H}(\gamma';\ell,P);m,P)$. For instance, bounds based on an entropy integral tend to follow this strategy. This approach effectively decomposes the problem of calculating the complexity of $\mathcal{H}(\gamma';\ell,P)$ into the problem of calculating the complexity of $\mathcal{H}$ and the problem of calculating some much more basic properties of $\mathcal{H}(\gamma';\ell,P)$. See [6, 15, 23, 35], or Section 5 below, for several explicit examples of this technique.

Another technique often (though not always) used in conjunction with the above strategy when deriving explicit rates of convergence is to relax $\mathrm{D}_\ell(\mathcal{H}(\gamma';\ell,P);P)$ to $\mathrm{D}_\ell(\mathcal{F}^*(\gamma';\ell,P);P)$ or $\mathrm{D}_\ell([\mathcal{H}](\gamma';\ell,P);P)$. This relaxation can sometimes be a source of slack; however, in many interesting cases, such as for certain losses ℓ [e.g., 6], or even certain noise conditions [e.g., 27, 33], this relaxed quantity can still lead to nearly tight bounds.

For our purposes, it will be convenient to make these common techniques explicit in the results. In later sections, this will make the benefits of our proposed methods more explicit, while still allowing us to state results in a form abstract enough to capture the variety of specific complexity measures most often used in conjunction with the above approach. Toward this end, we have the following definition.

DEFINITION 5. *For every distribution P over $\mathcal{X}\times\mathcal{Y}$, let $\mathring{\phi}_\ell(\sigma,\mathcal{H};m,P)$ be a quantity defined for every $\sigma\in[0,\infty]$, $\mathcal{H}\subseteq[\mathcal{F}]$, and $m\in\mathbb{N}$, such that the following conditions are satisfied when $f_P^\star\in\mathcal{H}$.*

$$\text{If } 0\le\sigma\le\sigma', \mathcal{H}\subseteq\mathcal{H}'\subseteq[\mathcal{F}], \mathcal{U}\subseteq\mathcal{X}, \text{ and } m'\le m,$$
$$\text{then } \mathring{\phi}_\ell(\sigma,\mathcal{H}_{\mathcal{U},f_P^\star};m,P)\le\mathring{\phi}_\ell(\sigma',\mathcal{H}';m',P). \tag{4}$$
$$\forall\sigma\ge\mathrm{D}_\ell(\mathcal{H};P), \phi_\ell(\mathcal{H};m,P)\le\mathring{\phi}_\ell(\sigma,\mathcal{H};m,P). \tag{5}$$

$\diamond$

For instance, most bounds based on entropy integrals can be made to satisfy this. See Section 5.3 for explicit examples of quantities $\mathring{\phi}_\ell$ from the literature that satisfy this definition. Given a function $\mathring{\phi}_\ell$ of this type, we define the following quantity for $m\in\mathbb{N}$, $s\in[1,\infty)$, $\zeta\in[0,\infty]$, $\mathcal{H}\subseteq[\mathcal{F}]$, and a distribution P over $\mathcal{X}\times\mathcal{Y}$.

$$\mathring{U}_\ell(\mathcal{H},\zeta;P,m,s)$$
$$=\tilde{K}\left(\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\zeta;\ell,P);P),\mathcal{H};m,P)+\mathrm{D}_\ell([\mathcal{H}](\zeta;\ell,P);P)\sqrt{\frac{s}{m}}+\frac{\bar{\ell}s}{m}\right).$$

Note that when $f_P^\star\in\mathcal{H}$, since $\mathrm{D}_\ell([\mathcal{H}](\gamma;\ell,P);P)\ge\mathrm{D}_\ell(\mathcal{H}(\gamma;\ell,P);P)$, Definition 5 implies $\phi_\ell(\mathcal{H}(\gamma;\ell,P);m,P)\le\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma;\ell,P);P),\mathcal{H}(\gamma;\ell,P);P,m)$, and furthermore $\mathcal{H}(\gamma;\ell,P) \subseteq \mathcal{H}$ so that $\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma;\ell,P);P),\mathcal{H}(\gamma;\ell,P);P,m) \leq \mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma;\ell,P);P),\mathcal{H};P,m)$. Thus,

$$\tilde{U}_\ell(\mathcal{H}(\gamma;\ell,P);P,m,s) \leq \mathring{U}_\ell(\mathcal{H}(\gamma;\ell,P),\gamma;P,m,s) \leq \mathring{U}_\ell(\mathcal{H},\gamma;P,m,s). \tag{6}$$

Furthermore, when $f_P^\star \in \mathcal{H}$, for any measurable $\mathcal{U} \subseteq \mathcal{U}' \subseteq \mathcal{X}$, any $\gamma' \geq \gamma \geq 0$, and any $\mathcal{H}' \subseteq [\mathcal{F}]$ with $\mathcal{H} \subseteq \mathcal{H}'$,

$$\mathring{U}_\ell(\mathcal{H}_{\mathcal{U},f_P^\star},\gamma;P,m,s) \leq \mathring{U}_\ell(\mathcal{H}'_{\mathcal{U}',f_P^\star},\gamma';P,m,s). \tag{7}$$

Note that the fact that we use $\mathrm{D}_\ell([\mathcal{H}](\gamma;\ell,P);P)$ instead of $\mathrm{D}_\ell(\mathcal{H}(\gamma;\ell,P);P)$ in the definition of $\mathring{U}_\ell$ is crucial for these inequalities to hold; specifically, it is not necessarily true that $\mathrm{D}_\ell(\mathcal{H}_{\mathcal{U},f_P^\star}(\gamma;\ell,P);P) \leq \mathrm{D}_\ell(\mathcal{H}_{\mathcal{U}',f_P^\star}(\gamma;\ell,P);P)$, but it is always the case that $[\mathcal{H}_{\mathcal{U},f_P^\star}](\gamma;\ell,P) \subseteq [\mathcal{H}_{\mathcal{U}',f_P^\star}](\gamma;\ell,P)$ when $f_P^\star \in [\mathcal{H}]$, so that $\mathrm{D}_\ell([\mathcal{H}_{\mathcal{U},f_P^\star}](\gamma;\ell,P);P) \leq \mathrm{D}_\ell([\mathcal{H}_{\mathcal{U}',f_P^\star}](\gamma;\ell,P);P)$.

Finally, for $\mathcal{H} \subseteq [\mathcal{F}]$, distribution P over $\mathcal{X} \times \mathcal{Y}$, values $\gamma,\gamma_1,\gamma_2 \geq 0$, $s \in [1,\infty)$, and any function $\mathfrak{s} : (0,\infty)^2 \to [1,\infty)$, define

$$\mathring{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) = \min\left\{m \in \mathbb{N} : \mathring{U}_\ell(\mathcal{H},\gamma_2;P,m,s) \leq \gamma_1\right\},$$
$$\mathring{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}) = \sup_{\gamma' \geq \gamma} \mathring{\mathrm{M}}_\ell(\gamma'/2,\gamma';\mathcal{H},P,\mathfrak{s}(\gamma,\gamma')).$$

For completeness, define $\mathring{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) = \infty$ when $\mathring{U}_\ell(\mathcal{H},\gamma_2;P,m,s) > \gamma_1$ for every $m \in \mathbb{N}$.

It will often be convenient to isolate the terms in $\mathring{U}_\ell$ when inverting for a sufficient m, thus arriving at an upper bound on $\mathring{\mathrm{M}}_\ell$. Specifically, define

$$\dot{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) = \min\left\{m \in \mathbb{N} : \mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);P)\sqrt{\frac{s}{m}} + \frac{\bar{\ell}s}{m} \leq \gamma_1\right\},$$
$$\ddot{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P) = \min\left\{m \in \mathbb{N} : \mathring{\phi}_\ell\left(\mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);P),\mathcal{H};P,m\right) \leq \gamma_1\right\}.$$

This way, for $\tilde{c} = 1/(2\tilde{K})$, we have

$$\mathring{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) \leq \max\left\{\ddot{\mathrm{M}}_\ell(\tilde{c}\gamma_1,\gamma_2;\mathcal{H},P),\dot{\mathrm{M}}_\ell(\tilde{c}\gamma_1,\gamma_2;\mathcal{H},P,s)\right\}. \tag{8}$$

Also note that we clearly have

$$\dot{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},P,s) \leq s \cdot \max\left\{\frac{4\mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);\ell,P)^2}{\gamma_1^2},\frac{2\bar{\ell}}{\gamma_1}\right\}, \tag{9}$$

so that, in the task of bounding $\mathring{\mathrm{M}}_\ell$, we can simply focus on bounding $\ddot{\mathrm{M}}_\ell$.

We will express our main abstract results below in terms of the incremental values $\mathring{\mathrm{M}}_\ell(\gamma_1,\gamma_2;\mathcal{H},\mathcal{P}_{XY},s)$; the quantity $\mathring{\mathrm{M}}_\ell(\gamma;\mathcal{H},\mathcal{P}_{XY},\mathfrak{s})$ will also be useful in deriving analogous results for ERM_ℓ. When $f_P^\star \in \mathcal{H}$, (6) implies

$$\bar{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}) \leq \tilde{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}) \leq \mathring{\mathrm{M}}_\ell(\gamma;\mathcal{H},P,\mathfrak{s}). \tag{10}$$

3. Methods Based on Optimizing the Surrogate Risk. Perhaps the simplest way to make use of a surrogate loss function is to try to optimize $\mathrm{R}_\ell(h)$ over $h \in \mathcal{F}$, until identifying $h \in \mathcal{F}$ with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star) < \Gamma_\ell(\varepsilon)$, at which point we are guaranteed $\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \varepsilon$. In this section, we briefly discuss some known results for this basic idea, along with a comment on the potential drawbacks of this approach for active learning.

3.1. *Passive Learning: Empirical Risk Minimization.* In the context of passive learning, the method of *empirical ℓ-risk minimization* is one of the most-studied methods for optimizing $\mathrm{R}_\ell(h)$ over $h \in \mathcal{F}$. Based on Lemma 4 and the above definitions, one can derive a bound on the number of labeled data points m sufficient for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ to achieve a given excess error rate. Specifically, the following theorem is due to Koltchinskii [23] (slightly modified here, following Giné and Koltchinskii [15], to allow for general $\mathfrak{s}$ functions). It will serve as our baseline for comparison in the applications below.

THEOREM 6. *Fix any function* $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$. *If* $f^\star \in \mathcal{F}$, *then for any* $m \geq \bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s})$, *with probability at least* $1 - \sum_{j \in \mathbb{Z}_{\Gamma_\ell(\varepsilon)}} 6e^{-\mathfrak{s}(\Gamma_\ell(\varepsilon), 2^j)}$, $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ *produces a function* $\hat{h}$ *such that* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. ⋄

3.2. *Negative Results for Active Learning.* As mentioned, there are several active learning methods designed to optimize a general loss function [8, 25]. However, it turns out that for many interesting loss functions, the number of labels required for active learning to achieve a given excess surrogate risk value is not significantly smaller than that sufficient for passive learning by ERM_ℓ.

Specifically, consider a problem with $\mathcal{X} = \{x_0, x_1\}$, let $z \in (0, 1/2)$ be a constant, and for $\varepsilon \in (0, z)$, let $\mathcal{P}(\{x_1\}) = \varepsilon/(2z)$, $\mathcal{P}(\{x_0\}) = 1 - \mathcal{P}(\{x_1\})$, and suppose $\mathcal{F}$ and ℓ are such that for $\eta(x_1) = 1/2 + z$ and any $\eta(x_0) \in [4/6, 5/6]$, we have $f^\star \in \mathcal{F}$. For this problem, any function h with $\mathrm{sign}(h(x_1)) \neq +1$ has $\mathrm{er}(h) - \mathrm{er}(f^\star) \geq \varepsilon$, so that $\Gamma_\ell(\varepsilon) \leq (\varepsilon/(2z))(\ell^\star_-(\eta(x_1)) - \ell^\star(\eta(x_1)))$; when ℓ is classification-calibrated and $\bar{\ell} < \infty$, this is $c\varepsilon$, for some ℓ-dependent $c \in (0, \infty)$. Any function h with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star) \leq c\varepsilon$ for this problem must have $\mathrm{R}_\ell(h; \mathcal{P}_{\{x_0\}}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{\{x_0\}}) \leq c\varepsilon/\mathcal{P}(\{x_0\}) = O(\varepsilon)$. Existing results of Hanneke and Yang [21] (with a slight modification to rescale for $\eta(x_0) \in [4/6, 5/6]$) imply that, for many classification-calibrated losses ℓ, the minimax optimal number of labels sufficient for an active learning algorithm to achieve this is $\Theta(1/\varepsilon)$. Hanneke and Yang [21] specifically show this for losses ℓ that are strictly positive, decreasing, strictly convex, and twice differentiable with continuous second derivative; however, that result can easily be extended to a wide variety of other classification-calibrated losses, such as the quadratic loss, which satisfy these conditions in a neighborhood of 0. It is also known [6] (see also below) that for many such losses (specifically, those satisfying Condition 3 with $r_\ell = 2$), $\Theta(1/\varepsilon)$ random labeled samples are sufficient for ERM_ℓ to achieve this same guarantee, so that results that only bound the surrogate risk of the function produced by an active learning method in this scenario can be at most a constant factor smaller than those provable for passive learning methods.

In the next section, we provide an active learning algorithm and a general analysis of its performance which, in the special case described above, guarantees excess error rate less than ε with high probability, using a number of label requests $O(\log(1/\varepsilon)\log\log(1/\varepsilon))$. The implication is that, to identify the improvements achievable by active learning with a surrogate loss, it is not sufficient to merely analyze the surrogate risk of the function produced by a given active learning algorithm. Indeed, since we are not particularly interested in the surrogate risk itself, we may even consider active learning algorithms that do not actually optimize $\mathrm{R}_\ell(h)$ over $h \in \mathcal{F}$ (even in the limit).

4. Alternative Use of the Surrogate Loss. Given that we are interested in ℓ only insofar as it helps us to optimize the error rate with computational efficiency, we should ask whether there is a method that sometimes makes more effective use of ℓ in terms of optimizing the error rate, while maintaining essentially the same computational advantages. The following method is essentially a relaxation of the methods of Koltchinskii [25] and Hanneke [20]. Similar results should also hold for analogous relaxations of the related methods of Balcan, Beygelzimer, and Langford [3], Dasgupta, Hsu, and Monteleoni [11], Balcan, Beygelzimer, and Langford [4], and Beygelzimer, Dasgupta, and Langford [8].

Algorithm 1:
Input: surrogate loss ℓ, unlabeled sample budget u, labeled sample budget n
Output: classifier $\hat{h}$

0. $V \leftarrow \mathcal{F}$, $Q \leftarrow \emptyset$, $m \leftarrow 0$, $t \leftarrow 0$, $k \leftarrow 1$, $m_1 \leftarrow 0$, $\hat{\gamma}_1 \leftarrow \bar{\ell}$
1. While $m < u$ and $t < n$
2. $\quad m \leftarrow m+1$
3. $\quad$ If $X_m \in \mathrm{DIS}(V)$
4. $\quad\quad$ Request label Y_m and let $Q \leftarrow Q \cup \{(m, Y_m)\}$, $t \leftarrow t+1$
5. $\quad$ If $\log_2(m - m_k) \in \mathbb{N}$ and $\hat{T}_\ell(V; Q, m, k)\frac{|Q| \vee 1}{m - m_k} \leq \hat{\gamma}_k/2$
6. $\quad\quad \hat{\gamma}_{k+1} \leftarrow \hat{T}_\ell(V; Q, m, k)\frac{|Q| \vee 1}{m - m_k}$, $m_{k+1} \leftarrow m$
7. $\quad\quad V \leftarrow \left\{h \in V : \mathrm{R}_\ell(h; Q) - \inf_{g \in V} \mathrm{R}_\ell(g; Q) \leq \hat{T}_\ell(V; Q, m, k)\right\}$
8. $\quad\quad Q \leftarrow \emptyset$, $k \leftarrow k+1$
9. Return $\hat{h} = \mathrm{argmin}_{h \in V} \mathrm{R}_\ell(h; Q)$

The intuition behind this algorithm is that, since we are only interested in achieving low error rate, once we have identified $\mathrm{sign}(f^\star(x))$ for a given $x \in \mathcal{X}$, there is no need to further optimize the value $\mathbb{E}[\ell(\hat{h}(X)Y)|X = x]$. Thus, as long as we maintain $f^\star \in V$, the data points $X_m \notin \mathrm{DIS}(V)$ are typically less informative than those $X_m \in \mathrm{DIS}(V)$. We therefore focus the label requests on those $X_m \in \mathrm{DIS}(V)$, since there remains some uncertainty about $\mathrm{sign}(f^\star(X_m))$ for these points. The algorithm updates V once enough samples have accumulated to estimate the excess risks under the current sampling distribution up to some desired precision. This update (Step 7) essentially removes from V those functions h whose excess empirical risks (under the current sampling distribution) are relatively large; by setting this threshold $\hat{T}_\ell$ appropriately, we can guarantee the excess empirical risk of $f^\star$ is smaller than $\hat{T}_\ell$. Thus, the algorithm maintains $f^\star \in V$ as an invariant, while focusing the sampling region $\mathrm{DIS}(V)$.

In practice, the set V can be maintained implicitly, simply by keeping track of the constraints (Step 7) that define it; then the condition in Step 3 can be checked by solving two constraint satisfaction problems (one for each sign); likewise, the value $\inf_{g \in V} \mathrm{R}_\ell(g; Q)$ in these constraints, as well as the final $\hat{h}$, can be found by solving constrained optimization problems. The quantity $\hat{T}_\ell$ in Algorithm 1 can be defined in one of several possible ways. In our context, we consider the following definition. Let $\{\xi'_k\}_{k \in \mathbb{N}}$ denote independent Rademacher random variables (i.e., uniform in $\{-1, +1\}$), also independent from $\mathcal{Z}$; these should be considered internal random bits used by the algorithm, which is therefore a randomized algorithm. For any $q \in \mathbb{N} \cup \{0\}$ and $Q = \{(i_1, y_1), \ldots, (i_q, y_q)\} \in (\mathbb{N} \times \{-1, +1\})^q$, let $S[Q] = \{(X_{i_1}, y_1), \ldots, (X_{i_q}, y_q)\}$, $\Xi[Q] = \{\xi'_{i_k}\}_{k=1}^q$. For $s \in [1, \infty)$, define

$$\hat{U}_\ell(\mathcal{H}; Q, s) = \hat{U}_\ell(\mathcal{H}; S[Q], \Xi[Q], s).$$

Then we can define the quantity $\hat{T}_\ell$ in the method above as

$$\hat{T}_\ell(\mathcal{H}; Q, m, k) = \hat{U}_\ell(\mathcal{H}; Q, \hat{\mathfrak{s}}(\hat{\gamma}_k, m - m_k)), \tag{11}$$

for some $\hat{\mathfrak{s}} : (0, \infty) \times \mathbb{N} \to [1, \infty)$. This definition has the appealing property that it allows us to interpret the update in Step 7 in two complementary ways: as comparing the empirical risks of functions in V under the conditional distribution given the region of disagreement $\mathcal{P}_{\mathrm{DIS}(V)}$, and as comparing the empirical risks of the functions in $V_{\mathrm{DIS}(V)}$ under the original distribution $\mathcal{P}_{XY}$.

For convenience, we will also suppose the function $\hat{\mathfrak{s}}$ in (11) satisfies, $\forall \gamma > 0$ and $m \in \mathbb{N}$,

$$\hat{\mathfrak{s}}(\gamma, m) = \hat{\mathfrak{s}}(2^{\lceil \log_2(\gamma) \rceil}, m), \tag{12}$$

so that we can effectively round γ to a power of 2.

We have the following theorem, which represents our main abstract result. The proof is included in Appendix A.

THEOREM 7. *For each* $j \geq -\lceil \log_2(\bar{\ell}) \rceil$, *let* $\mathfrak{s}_j(\cdot) = \hat{\mathfrak{s}}(2^{-j}, \cdot)$, *for* $\hat{\mathfrak{s}}$ *satisfying* (12), *let* $\mathcal{F}_j = \mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {\scriptstyle 01})_{\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); 01))}$, $\mathcal{U}_j = \mathrm{DIS}(\mathcal{F}_j)$, *and let* $u_j \in \mathbb{N}$ *satisfy* $\log_2(u_j) \in \mathbb{N}$ *and*

$$u_j \geq \mathring{\mathrm{M}}_\ell(2^{-j-2}, 2^{1-j}; \mathcal{F}_j, \mathcal{P}_{XY}, \mathfrak{s}_j(u_j)). \tag{13}$$

Suppose $f^\star \in \mathcal{F}$. *For any* $\varepsilon \in (0, 1)$, *and* $s \in [1, \infty)$, *if*

$$u \geq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} u_j \qquad \text{and} \qquad n \geq s + 2e \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \mathcal{P}(\mathcal{U}_j) u_j,$$

then, with arguments ℓ, u, *and* n, *Algorithm 1 uses at most* u *unlabeled samples and makes at most* n *label requests, and with probability at least*

$$1 - 2^{-s} - \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{i=1}^{\log_2(u_j)} 6 e^{-\mathfrak{s}_j(2^i)},$$

returns a function $\hat{h}$ *with* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. ◇

The number of label requests indicated by Theorem 7 can often (though not always) be significantly smaller than the number of random labeled data points sufficient for ERM_ℓ to achieve the same, as indicated by Theorem 6. This is typically the case when $\mathcal{P}(\mathcal{U}_j) \to 0$ as $j \to \infty$. When this is the case, the number of labels requested by the algorithm is sublinear in the number of unlabeled samples it processes; below, we will derive more explicit results for certain types of function classes $\mathcal{F}$, by characterizing the rate at which $\mathcal{P}(\mathcal{U}_j)$ vanishes in terms of a complexity measure known as the disagreement coefficient.

For the purpose of calculating the values $\mathring{\mathrm{M}}_\ell$ in Theorem 7, it is sometimes convenient to use the alternative interpretation of Algorithm 1, in terms of sampling Q from the conditional distribution $\mathcal{P}_{\mathrm{DIS}(V)}$. Specifically, the following lemma allows us to replace calculations in terms of $\mathcal{F}_j$ and $\mathcal{P}_{XY}$ with calculations in terms of $\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {\scriptstyle 01})$ and $\mathcal{P}_{\mathrm{DIS}(\mathcal{F}_j)}$. Its proof is included in Appendix A

LEMMA 8. *Let* $\mathring{\phi}_\ell$ *be any function satisfying Definition 5. Let* P *be any distribution over* $\mathcal{X} \times \mathcal{Y}$. *For any measurable* $\mathcal{U} \subseteq \mathcal{X} \times \mathcal{Y}$ *with* $P(\mathcal{U}) > 0$, *define* $P_{\mathcal{U}}(\cdot) = P(\cdot|\mathcal{U})$. *Also, for any* $\sigma \geq 0$, $\mathcal{H} \subseteq [\mathcal{F}]$, *and* $m \in \mathbb{N}$, *if* $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$, *define*

(14) $$\mathring{\phi}'_{\ell}(\sigma, \mathcal{H}; m, P) =$$
$$32\left(\inf_{\substack{\mathcal{U}=\mathcal{U}'\times\mathcal{Y}:\\ \mathcal{U}'\supseteq \mathrm{DISF}(\mathcal{H})}} P(\mathcal{U})\mathring{\phi}_{\ell}\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil(1/2)P(\mathcal{U})m\rceil, P_{\mathcal{U}}\right) + \frac{\bar{\ell}}{m} + \sigma\sqrt{\frac{1}{m}}\right),$$

and otherwise define $\mathring{\phi}'_{\ell}(\sigma, \mathcal{H}; m, P) = 0$. *Then the function* $\mathring{\phi}'_{\ell}$ *also satisfies Definition 5.* ◇

Plugging this $\mathring{\phi}'_{\ell}$ function into Theorem 7 immediately yields the following corollary, the proof of which is included in Appendix A.

COROLLARY 9. *For each* $j \geq -\lceil\log_2(\bar{\ell})\rceil$, *let* $\mathcal{F}_j$, $\mathcal{U}_j$, *and* $\mathfrak{s}_j$ *be as in Theorem 7, and if* $\mathcal{P}(\mathcal{U}_j) > 0$, *let* $u_j \in \mathbb{N}$ *satisfy* $\log_2(u_j) \in \mathbb{N}$ *and*

(15) $$u_j \geq 2\mathcal{P}(\mathcal{U}_j)^{-1}\mathring{\mathrm{M}}_{\ell}\left(\frac{2^{-j-8}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \mathfrak{s}_j(u_j)\right).$$

If $\mathcal{P}(\mathcal{U}_j) = 0$, *let* $u_j \in \mathbb{N}$ *satisfy* $\log_2(u_j) \in \mathbb{N}$ *and* $u_j \geq \tilde{K}\bar{\ell}\mathfrak{s}_j(u_j)2^{j+2}$. *Suppose* $f^{\star} \in \mathcal{F}$. *For any* $\varepsilon \in (0, 1)$ *and* $s \in [1, \infty)$, *if*

$$u \geq \sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Gamma_{\ell}(\varepsilon))\rfloor} u_j \qquad \textit{and} \qquad n \geq s + 2e\sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Gamma_{\ell}(\varepsilon))\rfloor} \mathcal{P}(\mathcal{U}_j)u_j,$$

then, with arguments ℓ, u, *and* n, *Algorithm 1 uses at most* u *unlabeled samples and makes at most* n *label requests, and with probability at least*

$$1 - 2^{-s} - \sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Gamma_{\ell}(\varepsilon))\rfloor}\sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)},$$

returns a function $\hat{h}$ *with* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^{\star}) \leq \varepsilon$. ◇

Algorithm 1 can be modified in a variety of interesting ways, leading to related methods that can be analyzed analogously. One simple modification is to use a more involved bound to define the quantity $\hat{T}_{\ell}$. For instance, for Q as above, and a function $\hat{\mathfrak{s}}_k : (0, \infty) \times \mathbb{N} \to [1, \infty)$, one could define

$$\hat{T}_{\ell}(\mathcal{H}; Q, m, k) = (3/2)q^{-1}\inf\Big\{\lambda > 0 : \forall j \in \mathbb{Z}_{\lambda},$$
$$\hat{U}_{\ell}\left(\mathcal{H}\left(3q^{-1}2^{j-1}; \ell, S[Q]\right); Q, \hat{\mathfrak{s}}_k\left(3q^{-1}2^{j-1}, m - m_k\right)\right) \leq 2^{j-4}q^{-1}\Big\},$$

for which one can also prove a result similar to Lemma 4 [see 15, 23]. This definition shares the convenient dual-interpretations property mentioned above about $\hat{U}_\ell(\mathcal{H}; Q, \hat{\mathfrak{s}}(\hat{\gamma}_k, m-m_k))$; furthermore, the results above for Algorithm 1 also hold under this definition (for appropriate $\hat{\mathfrak{s}}_k$ functions), with only minor modifications to constants and event probabilities.

The update trigger in Step 5 can be modified in several ways, leading to interesting related methods. One simple change would be replacing it with $\log_2(m) \in \mathbb{N}$, as in the methods of Hanneke [20], which simplifies the algorithm to some extent. In most applications of interest, this still yields a result similar to Theorem 7, since we might expect the value $\mathring{\mathrm{M}}_\ell(2^{-j-2}, 2^{1-j}; \mathcal{F}_j, \mathcal{P}_{XY}, \mathfrak{s}_j(u_j))$ to be at least twice as large as $\mathring{\mathrm{M}}_\ell(2^{-j-1}, 2^{2-j}; \mathcal{F}_{j-1}, \mathcal{P}_{XY}, \mathfrak{s}_{j-1}(u_{j-1}))$ anyway. Another interesting possibility is to replace the last condition in Step 5 with a check for $\hat{T}_\ell(V; Q, m, k)\frac{|Q|\vee 1}{m-m_k} < \Gamma_\ell(2^{-k})$. Of course, the value $\Gamma_\ell(2^{-k})$ is typically not directly available to us, but we could substitute a distribution-independent lower bound on $\Gamma_\ell(2^{-k})$, for instance based on the ψ_ℓ function of Bartlett, Jordan, and McAuliffe [6]; in the active learning context, we could potentially use unlabeled samples to estimate a $\mathcal{P}$-dependent lower bound on $\Gamma_\ell(2^{-k})$, or even $\mathrm{diam}(V)\psi_\ell(2^{-k}/2\mathrm{diam}(V))$, based on (3), where $\mathrm{diam}(V) = \sup_{h,g\in V} \Delta(h, g)$.

5. Applications.

In this section, we apply the abstract results from above to a few commonly-studied scenarios: namely, VC subgraph classes and entropy conditions, with some additional mention of VC major classes and VC hull classes. In the interest of making the results more concise and explicit, we express them in terms of well-known conditions relating distances to excess risks. We also express them in terms of a lower bound on $\Gamma_\ell(\varepsilon)$ of the type in (2), with convenient properties that allow for closed-form expression of the results. To simplify the presentation, we often omit numerical constant factors in the inequalities below, and for this we use the common notation $f(x) \lesssim g(x)$ to mean that $f(x) \leq cg(x)$ for some implicit universal constant $c \in (0, \infty)$.

5.1. *Diameter Conditions.*

To begin, we first state some general characterizations relating distances to excess risks; these characterizations will make it easier to express our results more concretely below, and make for a more straightforward comparison between results for the above methods. The following condition, introduced by Mammen and Tsybakov [27] and Tsybakov [33], is a well-known noise condition, about which there is now an extensive literature [e.g., 6, 19, 20, 23].

CONDITION 10. *For some $a \in [1, \infty)$ and $\alpha \in [0, 1]$, for every $g \in \mathcal{F}^*$,*

$$\Delta(g, f^\star) \leq a\left(\mathrm{er}(g) - \mathrm{er}(f^\star)\right)^\alpha.$$

$\diamond$

Condition 10 can be equivalently expressed in terms of certain noise conditions [6, 27, 33]. Specifically, satisfying Condition 10 with some $\alpha < 1$ is equivalent to the existence of some $a' \in [1, \infty)$ such that, for all $\varepsilon > 0$,

$$\mathcal{P}\left(x : |\eta(x) - 1/2| \leq \varepsilon\right) \leq a' \varepsilon^{\alpha/(1-\alpha)},$$

which is often referred to as a *low noise* condition. Additionally, satisfying Condition 10 with $\alpha = 1$ is equivalent to having some $a' \in [1, \infty)$ such that

$$\mathcal{P}\left(x : |\eta(x) - 1/2| \leq 1/a'\right) = 0,$$

often referred to as a *bounded noise* condition.

For simplicity, we formulate our results in terms of a and α from Condition 10. However, for the abstract results in this section, the results remain valid under the weaker condition that replaces $\mathcal{F}^*$ by $\mathcal{F}$, and adds the condition that $f^\star \in \mathcal{F}$. In fact, the specific results in this section also remain valid using this weaker condition while additionally using (3) in place of (2), as remarked above.

An analogous condition can be defined for the surrogate loss function, as follows. Similar notions have been explored by Bartlett, Jordan, and McAuliffe [6] and Koltchinskii [23].

CONDITION 11. *For some $b \in [1, \infty)$ and $\beta \in [0, 1]$, for every $g \in [\mathcal{F}]$,*

$$\mathrm{D}_\ell\left(g, f^\star; P\right)^2 \leq b\left(\mathrm{R}_\ell(g; P) - \mathrm{R}_\ell(f^\star; P)\right)^\beta.$$

◇

Note that these conditions are *always* satisfied for *some* values of a, b, α, β, since $\alpha = \beta = 0$ trivially satisfies the conditions. However, in more benign scenarios, values of α and β strictly greater than 0 can be satisfied. Furthermore, for some loss functions ℓ, Condition 11 can even be satisfied *universally*, in the sense that a value of $\beta > 0$ is satisfied for *all* distributions. In particular, Bartlett, Jordan, and McAuliffe [6] show that this is the case under Condition 3, as stated in the following lemma [see 6, for the proof].

LEMMA 12. *Suppose Condition 3 is satisfied. Let $\beta = \min\{1, \frac{2}{r_\ell}\}$ and $b = (2C'_\ell)^{-\beta} L^2$, where $C'_\ell = C_\ell$ for $r_\ell \geq 2$, and $C'_\ell = C_\ell \bar{d}_\ell^{r_\ell - 2}$ otherwise. Then* every *distribution P over $\mathcal{X} \times \mathcal{Y}$ with $f^\star_P \in [\mathcal{F}]$ satisfies Condition 11 with these values of b and β.* ◇

Under Condition 10, it is particularly straightforward to obtain bounds on $\Gamma_\ell(\varepsilon)$ based on a function $\Psi_\ell(\varepsilon)$ satisfying (2). For instance, since $x \mapsto x\psi_\ell(1/x)$ is nonincreasing on $(0, \infty)$ [6], the function

$$\Psi_\ell(\varepsilon) = a\varepsilon^\alpha \psi_\ell\left(\varepsilon^{1-\alpha}/(2a)\right) \tag{16}$$

satisfies $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ [6]. Furthermore, for classification-calibrated ℓ, Ψ_ℓ in (16) is strictly increasing, nonnegative, and continuous on $[0,1]$ [6], and has $\Psi_\ell(0) = 0$; thus, the inverse $\Psi_\ell^{-1}(\gamma)$, defined for all $\gamma > 0$ by

$$\Psi_\ell^{-1}(\gamma) = \inf\{\varepsilon > 0 : \gamma \leq \Psi_\ell(\varepsilon)\} \cup \{1\}, \tag{17}$$

is strictly increasing, nonnegative, and continuous on $(0, \Psi_\ell(1))$. Furthermore, one can easily show $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing on $(0,\infty)$. Also note that $\forall \gamma > 0, \mathcal{E}_\ell(\gamma) \leq \Psi_\ell^{-1}(\gamma)$.

5.2. *The Disagreement Coefficient.* In order to more concisely state our results, it will be convenient to bound $\mathcal{P}(\mathrm{DIS}(\mathcal{H}))$ by a linear function of $\mathrm{radius}(\mathcal{H})$, for $\mathrm{radius}(\mathcal{H})$ in a given range. This type of relaxation has been used extensively in the active learning literature [5, 8, 11, 14, 17–20, 25, 26, 32, 37], and the coefficient in the linear function is typically referred to as the *disagreement coefficient*. Specifically, the following definition is due to Hanneke [17, 19]; related quantities have been explored by Alexander [1] and Giné and Koltchinskii [15].

DEFINITION 13. *For any* $r_0 > 0$, *define the* disagreement coefficient *of a function* $h : \mathcal{X} \to \mathbb{R}$ *with respect to* $\mathcal{F}$ *under* $\mathcal{P}$ *as*

$$\theta_h(r_0) = \sup_{r > r_0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} \vee 1.$$

If $f^\star \in \mathcal{F}$, *define the disagreement coefficient of the class* $\mathcal{F}$ *as* $\theta(r_0) = \theta_{f^\star}(r_0)$. $\diamond$

The value of $\theta(\varepsilon)$ has been studied and bounded for various function classes $\mathcal{F}$ under various conditions on $\mathcal{P}$. In many cases of interest, $\theta(\varepsilon)$ is known to be bounded by a finite constant [5, 14, 17, 19, 26], while in other cases, $\theta(\varepsilon)$ may have an interesting dependence on ε [5, 32, 37]. The reader is referred to the works of Hanneke [19, 20] for detailed discussions on the disagreement coefficient.

5.3. *Specification of* $\mathring{\phi}_\ell$. Next, we recall a few well-known bounds on the ϕ_ℓ function, which leads to a more concrete instance of a function $\mathring{\phi}_\ell$ satisfying Definition 5. Below, we let $\mathcal{G}^*$ denote the set of measurable functions $g : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$. Also, for $\mathcal{G} \subseteq \mathcal{G}^*$, let $\mathrm{F}(\mathcal{G}) = \sup_{g\in\mathcal{G}} |g|$ denote the minimal *envelope* function for $\mathcal{G}$, and for $g \in \mathcal{G}^*$ let $\|g\|_P^2 = \int g^2 \mathrm{d}P$ denote the squared $L_2(P)$ seminorm of g; we will generally assume $\mathrm{F}(\mathcal{G})$ is measurable in the discussion below.

Uniform Entropy: The first bound is based on the work of van der Vaart and Wellner [34]; related bounds have been studied by Giné and Koltchinskii [15], Giné, Koltchinskii, and Wellner [16], van der Vaart and Wellner [35], and others. For a distribution P over $\mathcal{X} \times \mathcal{Y}$, a set $\mathcal{G} \subseteq \mathcal{G}^*$, and $\varepsilon \geq 0$, let $\mathcal{N}(\varepsilon, \mathcal{G}, L_2(P))$ denote the size of a minimal ε-cover of $\mathcal{G}$ (that is, the minimum number of balls of radius at most ε sufficient to cover $\mathcal{G}$), where distances are measured in terms of the $L_2(P)$ pseudo-metric: $(f, g) \mapsto \|f - g\|_P$. For $\sigma \geq 0$ and $\mathrm{F} \in \mathcal{G}^*$, define the function

$$J(\sigma, \mathcal{G}, \mathrm{F}) = \sup_Q \int_0^\sigma \sqrt{1 + \ln \mathcal{N}(\varepsilon \|\mathrm{F}\|_Q, \mathcal{G}, L_2(Q))} \mathrm{d}\varepsilon,$$

where Q ranges over all finitely discrete probability measures.

Fix any distribution P over $\mathcal{X} \times \mathcal{Y}$ and any $\mathcal{H} \subseteq [\mathcal{F}]$ with $f_P^\star \in \mathcal{H}$, and let

$$\begin{aligned} &\mathcal{G}_{\mathcal{H}} = \{(x, y) \mapsto \ell(h(x)y) : h \in \mathcal{H}\}, \\ \text{(18)} \qquad &\text{and } \mathcal{G}_{\mathcal{H},P} = \{(x, y) \mapsto \ell(h(x)y) - \ell(f_P^\star(x)y) : h \in \mathcal{H}\}. \end{aligned}$$

Then, since $J(\sigma, \mathcal{G}_{\mathcal{H}}, \mathrm{F}) = J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$, it follows from Theorem 2.1 of van der Vaart and Wellner [34] (and a triangle inequality) that for some universal constant $c \in [1, \infty)$, for any $m \in \mathbb{N}$, $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, and $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$,

$$\text{(19)} \qquad \phi_\ell(\mathcal{H}; P, m) \leq cJ\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_{\mathcal{H}}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_{\mathcal{H}}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\sigma^2 m}\right).$$

Based on (19), it is straightforward to define a function $\mathring{\phi}_\ell$ that satisfies Definition 5. Specifically, define

$$\text{(20)} \qquad \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P) = \inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})} \inf_{\lambda \geq \sigma} cJ\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_{\mathcal{H}}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_{\mathcal{H}}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\lambda^2 m}\right),$$

for c as in (19). By (19), $\mathring{\phi}_\ell^{(1)}$ satisfies (5). Also note that $m \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is nonincreasing, while $\sigma \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is nondecreasing. Furthermore, $\mathcal{H} \mapsto \mathcal{N}(\varepsilon, \mathcal{G}_{\mathcal{H}}, L_2(Q))$ is nondecreasing for all Q, so that $\mathcal{H} \mapsto J(\sigma, \mathcal{G}_{\mathcal{H}}, \mathrm{F})$ is nondecreasing as well; since $\mathcal{H} \mapsto \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$ is also nondecreasing, we see that $\mathcal{H} \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is nondecreasing. Similarly, for $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{N}(\varepsilon, \mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star}}, L_2(Q)) \leq \mathcal{N}(\varepsilon, \mathcal{G}_{\mathcal{H}}, L_2(Q))$ for all Q, so that $J(\sigma, \mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star}}, \mathrm{F}) \leq J(\sigma, \mathcal{G}_{\mathcal{H}}, \mathrm{F})$; because $\mathrm{F}(\mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star},P}) \leq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, we have $\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}_{\mathcal{U},f_P^\star}; m, P) \leq \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ as well. Thus, to satisfy Definition 5, it suffices to take $\mathring{\phi}_\ell = \mathring{\phi}_\ell^{(1)}$.

Bracketing Entropy: Our second bound is a classic result in empirical process theory. For functions $g_1 \leq g_2$, a *bracket* $[g_1, g_2]$ is the set of functions $g \in \mathcal{G}^*$ with $g_1 \leq g \leq g_2$; $[g_1, g_2]$ is called an ε-bracket under $L_2(P)$ if $\|g_1 - g_2\|_P < \varepsilon$. Then $\mathcal{N}_{[]}(\varepsilon, \mathcal{G}, L_2(P))$ denotes the smallest number of ε-brackets (under $L_2(P)$) sufficient to cover $\mathcal{G}$. For $\sigma \geq 0$, define the function

$$J_{[]}(\sigma, \mathcal{G}, P) = \int_0^\sigma \sqrt{1 + \ln \mathcal{N}_{[]}(\varepsilon, \mathcal{G}, L_2(P))} \mathrm{d}\varepsilon.$$

Fix any $\mathcal{H} \subseteq [\mathcal{F}]$, and let $\mathcal{G}_\mathcal{H}$ and $\mathcal{G}_{\mathcal{H},P}$ be as above. Then since $J_{[]}(\sigma, \mathcal{G}_\mathcal{H}, P) = J_{[]}(\sigma, \mathcal{G}_{\mathcal{H},P}, P)$, Lemma 3.4.2 of van der Vaart and Wellner [35] and a triangle inequality imply that for some universal constant $c \in [1, \infty)$, for any $m \in \mathbb{N}$ and $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$,

$$\phi_\ell(\mathcal{H}; P, m) \leq c J_{[]}\left(\sigma, \mathcal{G}_\mathcal{H}, P\right) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}\left(\sigma, \mathcal{G}_\mathcal{H}, P\right) \bar{\ell}}{\sigma^2 m} \right). \tag{21}$$

As-is, the right side of (21) nearly satisfies Definition 5 already. Only a slight modification is required to fulfill the requirement of monotonicity in σ. Specifically, define

$$\mathring{\phi}_\ell^{(2)}(\sigma, \mathcal{H}; P, m) = \inf_{\lambda \geq \sigma} c J_{[]}\left(\lambda, \mathcal{G}_\mathcal{H}, P\right) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}\left(\lambda, \mathcal{G}_\mathcal{H}, P\right) \bar{\ell}}{\lambda^2 m} \right), \tag{22}$$

for c as in (21). Then taking $\mathring{\phi}_\ell = \mathring{\phi}_\ell^{(2)}$ suffices to satisfy Definition 5.

Since Definition 5 is satisfied for both $\mathring{\phi}_\ell^{(1)}$ and $\mathring{\phi}_\ell^{(2)}$, it is also satisfied for

$$\mathring{\phi}_\ell = \min\left\{ \mathring{\phi}_\ell^{(1)}, \mathring{\phi}_\ell^{(2)} \right\}. \tag{23}$$

For the remainder of this section, we suppose $\mathring{\phi}_\ell$ is defined as in (23) (for all distributions P over $\mathcal{X} \times \mathcal{Y}$), and study the implications arising from the combination of this definition with the abstract theorems above.

5.4. *VC Subgraph Classes.* For a collection $\mathcal{A}$ of sets, a set $\{z_1, \ldots, z_k\}$ of points is said to be *shattered* by $\mathcal{A}$ if $|\{A \cap \{z_1, \ldots, z_k\} : A \in \mathcal{A}\}| = 2^k$. The VC dimension $\mathrm{vc}(\mathcal{A})$ of $\mathcal{A}$ is then defined as the largest integer k for which there exist k points $\{z_1, \ldots, z_k\}$ shattered by $\mathcal{A}$ [36]; if no such largest k exists, we define $\mathrm{vc}(\mathcal{A}) = \infty$. For a set $\mathcal{G}$ of real-valued functions, denote by $\mathrm{vc}(\mathcal{G})$ the VC dimension of the collection $\{\{(x, y) : y < g(x)\} : g \in \mathcal{G}\}$ of subgraphs of functions in $\mathcal{G}$ (called the pseudo-dimension [22, 31]); to simplify the statement of results below, we adopt the convention that when the VC dimension of this collection is 0, we let $\mathrm{vc}(\mathcal{G}) = 1$. A set $\mathcal{G}$ is said to be a VC subgraph class if $\mathrm{vc}(\mathcal{G}) < \infty$ [35].

Because we are interested in results concerning values of $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star)$, for functions h in certain subsets $\mathcal{H} \subseteq [\mathcal{F}]$, we will formulate results below in terms of $\mathrm{vc}(\mathcal{G}_\mathcal{H})$, for $\mathcal{G}_\mathcal{H}$ defined as above. Depending on certain properties of ℓ, these results can often be restated directly in terms of $\mathrm{vc}(\mathcal{H})$; for instance, this is true when ℓ is monotone, since $\mathrm{vc}(\mathcal{G}_\mathcal{H}) \leq \mathrm{vc}(\mathcal{H})$ in that case [12, 22, 29].

The following is a well-known result for VC subgraph classes [see e.g., 35], derived from the works of Pollard [30] and Haussler [22].

LEMMA 14. *For any $\mathcal{G} \subseteq \mathcal{G}^*$, for any measurable $\mathrm{F} \geq \mathrm{F}(\mathcal{G})$, for any distribution Q such that $\|\mathrm{F}\|_Q > 0$, for any $\varepsilon \in (0, 1)$,*

$$\mathcal{N}(\varepsilon\|\mathrm{F}\|_Q, \mathcal{G}, L_2(Q)) \leq A(\mathcal{G}) \left(\frac{1}{\varepsilon}\right)^{2\mathrm{vc}(\mathcal{G})}.$$

where $A(\mathcal{G}) \lesssim (\mathrm{vc}(\mathcal{G}) + 1)(16e)^{\mathrm{vc}(\mathcal{G})}$. $\diamond$

In particular, Lemma 14 implies that any $\mathcal{G} \subseteq \mathcal{G}^*$ has, $\forall \sigma \in (0, 1]$,

$$\begin{aligned} J(\sigma, \mathcal{G}, \mathrm{F}) &\leq \int_0^\sigma \sqrt{\ln(eA(\mathcal{G})) + 2\mathrm{vc}(\mathcal{G}) \ln(1/\varepsilon)} \mathrm{d}\varepsilon \\ &\leq 2\sigma\sqrt{\ln(eA(\mathcal{G}))} + \sqrt{8\mathrm{vc}(\mathcal{G})} \int_0^\sigma \sqrt{\ln(1/\varepsilon)}\mathrm{d}\varepsilon \\ &= 2\sigma\sqrt{\ln(eA(\mathcal{G}))} + \sigma\sqrt{8\mathrm{vc}(\mathcal{G}) \ln(1/\sigma)} + \sqrt{2\pi\mathrm{vc}(\mathcal{G})}\mathrm{erfc}\left(\sqrt{\ln(1/\sigma)}\right). \end{aligned} \tag{24}$$

Since $\mathrm{erfc}(x) \leq \exp\{-x^2\}$ for all $x \geq 0$, (24) implies $\forall \sigma \in (0, 1]$,

$$J(\sigma, \mathcal{G}, \mathrm{F}) \lesssim \sigma\sqrt{\mathrm{vc}(\mathcal{G})\mathrm{Log}(1/\sigma)}. \tag{25}$$

Applying these observations to bound $J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$ for $\mathcal{H} \subseteq [\mathcal{F}]$ and $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, noting $J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F}) = J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$ and $\mathrm{vc}(\mathcal{G}_{\mathcal{H},P}) = \mathrm{vc}(\mathcal{G}_\mathcal{H})$, and plugging the resulting bound into (20) yields the following well-known bound on $\phi_\ell^{(1)}$ due to Giné and Koltchinskii [15]. For any $m \in \mathbb{N}$ and $\sigma > 0$,

$$\begin{aligned} &\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P) \\ &\quad\lesssim \inf_{\lambda \geq \sigma} \lambda\sqrt{\frac{\mathrm{vc}(\mathcal{G}_\mathcal{H})\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}} + \frac{\mathrm{vc}(\mathcal{G}_\mathcal{H})\bar{\ell}\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}. \end{aligned} \tag{26}$$

Specifically, to arrive at (26), we relaxed the $\inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})}$ in (20) by taking $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$ such that $\|\mathrm{F}\|_P = \max\{\sigma, \|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P\}$, thus maintaining $\lambda/\|\mathrm{F}\|_P \in$

$(0,1]$ for the minimizing λ value, so that (25) remains valid; we also made use of the fact that $\mathrm{Log} \geq 1$, which gives us $\mathrm{Log}(\|\mathrm{F}\|_P/\lambda) = \mathrm{Log}(\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P/\lambda)$ for this case.

In particular, (26) implies

$$\ddot{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P) \lesssim \inf_{\sigma \geq \mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);P)} \left(\frac{\sigma^2}{\gamma_1^2} + \frac{\bar{\ell}}{\gamma_1}\right) \mathrm{vc}(\mathcal{G}_{\mathcal{H}}) \mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\sigma}\right). \tag{27}$$

Following Giné and Koltchinskii [15], for $r > 0$, define $\mathrm{B}_{\mathcal{H},P}(f_P^\star, r; \ell) = \{g \in \mathcal{H} : \mathrm{D}_\ell(g, f_P^\star; P)^2 \leq r\}$, and for $r_0 \geq 0$, define

$$\tau_\ell(r_0; \mathcal{H}, P) = \sup_{r > r_0} \frac{\left\|\mathrm{F}\left(\mathcal{G}_{\mathrm{B}_{\mathcal{H},P}(f_P^\star, r;\ell),P}\right)\right\|_P^2}{r} \vee 1.$$

When $P = \mathcal{P}_{XY}$, abbreviate this as $\tau_\ell(r_0; \mathcal{H}) = \tau_\ell(r_0; \mathcal{H}, \mathcal{P}_{XY})$, and when $\mathcal{H} = \mathcal{F}$, further abbreviate $\tau_\ell(r_0) = \tau_\ell(r_0; \mathcal{F}, \mathcal{P}_{XY})$. For $\lambda > 0$, when $f_P^\star \in \mathcal{H}$ and P satisfies Condition 11, (27) implies that,

$$\sup_{\gamma \geq \lambda} \ddot{\mathrm{M}}_\ell(\gamma/(4\tilde{K}), \gamma; \mathcal{H}(\gamma; \ell, P), P) \lesssim \left(\frac{b}{\lambda^{2-\beta}} + \frac{\bar{\ell}}{\lambda}\right) \mathrm{vc}(\mathcal{G}_{\mathcal{H}}) \mathrm{Log}\left(\tau_\ell\left(b\lambda^\beta; \mathcal{H}, P\right)\right). \tag{28}$$

Combining this observation with (6), (8), (9), (10), and Theorem 6, we arrive at a result for the sample complexity of empirical ℓ-risk minimization with a general VC subgraph class under Conditions 10 and 11. Specifically, for $\mathfrak{s} : (0,\infty)^2 \to [1,\infty)$, when $f^\star \in \mathcal{F}$, (6) implies that

$$\begin{aligned} \bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s}) &\leq \tilde{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s}) \\ &= \sup_{\gamma \geq \Gamma_\ell(\varepsilon)} \tilde{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma)) \\ &\leq \sup_{\gamma \geq \Gamma_\ell(\varepsilon)} \mathring{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma)). \end{aligned} \tag{29}$$

Supposing $\mathcal{P}_{XY}$ satisfies Conditions 10 and 11, applying (8), (9), and (28) to (29), and taking $\mathfrak{s}(\lambda, \gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at the following theorem, which is implicit in the work of Giné and Koltchinskii [15].

THEOREM 15. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10 and Condition 11, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (16), then for any $\varepsilon \in (0,1)$, letting $\tau_\ell = \tau_\ell\left(b\Psi_\ell(\varepsilon)^\beta\right)$, for any $m \in \mathbb{N}$ with*

$$m \geq c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\tau_\ell\right) + \mathrm{Log}\left(1/\delta\right)\right), \tag{30}$$

with probability at least $1-\delta$, $\mathrm{ERM}_\ell(\mathcal{F},\mathcal{Z}_m)$ produces $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. $\diamond$

As noted by Giné and Koltchinskii [15], in the special case when ℓ is itself the 0-1 loss, the bound in Theorem 15 simplifies quite nicely, since in that case $\|\mathrm{F}(\mathcal{G}_{\mathrm{B}_{\mathcal{F},\mathcal{P}_{XY}}(f^\star,r;\ell),\mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} = \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star,r\right)\right)\right)$, so that $\tau_\ell(r_0) = \theta(r_0)$; in this case, we also have $\mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \leq \mathrm{vc}(\mathcal{F})$ and $\Psi_\ell(\varepsilon) = \varepsilon/2$, and we can take $\beta = \alpha$ and $b = a$, so that it suffices to have

$$m \geq ca\varepsilon^{\alpha-2}\left(\mathrm{vc}(\mathcal{F})\mathrm{Log}\left(\theta\right) + \mathrm{Log}\left(1/\delta\right)\right), \tag{31}$$

where $\theta = \theta\left(a\varepsilon^\alpha\right)$ and $c \in [1,\infty)$ is a universal constant. It is known that this is sometimes the minimax optimal number of samples sufficient for passive learning [9, 19, 32].

Next, we turn to the performance of Algorithm 1 under the conditions of Theorem 15. Specifically, suppose $\mathcal{P}_{XY}$ satisfies Conditions 10 and 11, and for $\gamma_0 \geq 0$, define

$$\chi_\ell(\gamma_0) = \sup_{\gamma > \gamma_0} \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star, a\mathcal{E}_\ell\left(\gamma\right)^\alpha\right)\right)\right)}{b\gamma^\beta} \vee 1.$$

Note that $\|\mathrm{F}(\mathcal{G}_{\mathcal{F}_j,\mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} \leq \bar{\ell}^2\mathcal{P}\left(\mathrm{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{1-j}\right);_{01}\right)\right)\right)$. Thus, by (27), for $-\lceil\log_2(\bar{\ell})\rceil \leq j \leq \lfloor\log_2(2/\Psi_\ell(\varepsilon))\rfloor$,

$$\ddot{\mathrm{M}}_\ell(2^{-j-3}\tilde{K}^{-1}, 2^{1-j}; \mathcal{F}_j, \mathcal{P}_{XY}) \lesssim \left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\chi_\ell\left(\Psi_\ell(\varepsilon)\right)\bar{\ell}\right). \tag{32}$$

With a little additional work to define an appropriate $\mathfrak{s}_j$ function and derive closed-form bounds on the summations in Theorem 7, we arrive at the following theorem regarding the performance of Algorithm 1 for VC subgraph classes. For completeness, the remaining technical details of the proof are included in Appendix A

THEOREM 16. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10 and Condition 11, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (16), for any $\varepsilon \in (0,1)$, letting $\theta = \theta\left(a\varepsilon^\alpha\right)$, $\chi_\ell = \chi_\ell(\Psi_\ell(\varepsilon))$, $A_1 =$*

$\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}(\chi_\ell \bar{\ell}) + \mathrm{Log}(1/\delta)$, $B_1 = \min\left\{\frac{1}{1-2^{(\alpha+\beta-2)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, *and* $C_1 = \min\left\{\frac{1}{1-2^{(\alpha-1)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, *if*

$$u \geq c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right) A_1 \tag{33}$$

and

$$n \geq c\theta a \varepsilon^{\alpha}\left(\frac{b(A_1 + \mathrm{Log}(B_1))B_1}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1 + \mathrm{Log}(C_1))C_1}{\Psi_\ell(\varepsilon)}\right), \tag{34}$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function satisfying (12)*, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1-\delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.* ◇

To be clear, in specifying B_1 and C_1, we have adopted the convention that $1/0 = \infty$ and $\min\{\infty, x\} = x$ for any $x \in \mathbb{R}$, so that B_1 and C_1 are well-defined even when $\alpha = \beta = 1$, or $\alpha = 1$, respectively. Note that, when $\alpha + \beta < 2$, $B_1 = O(1)$, so that the asymptotic dependence on ε in (34) is $O\left(\theta\varepsilon^{\alpha}\Psi_\ell(\varepsilon)^{\beta-2}\mathrm{Log}(\chi_\ell)\right)$, while in the case of $\alpha = \beta = 1$, it is $O\left(\theta\mathrm{Log}(1/\varepsilon)(\mathrm{Log}(\theta) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)))\right)$. It is likely that the logarithmic and constant factors can be improved in many cases (particularly the $\mathrm{Log}(\chi_\ell\bar{\ell})$, B_1, and C_1 factors).

Comparing the result in Theorem 16 to Theorem 15, we see that the condition on u in (33) is almost identical to the condition on m in (30), aside from a change in the logarithmic factor, so that the total number of data points needed is roughly the same. However, the number of *labels* indicated by (34) may often be significantly smaller than the condition in (30), reducing it by a factor of roughly $\theta a\varepsilon^{\alpha}$. This reduction is particularly strong when θ is bounded by a finite constant. Moreover, this is the same *type* of improvement that is known to occur when ℓ is itself the 0-1 loss [19], so that in particular these results agree with the existing analysis in this special case, and are therefore sometimes nearly minimax [19, 32]. Regarding the slight difference between (33) and (30) from replacing τ_ℓ by $\chi_\ell\bar{\ell}$, the effect is somewhat mixed, and which of these is smaller may depend on the particular class $\mathcal{F}$ and loss ℓ; we can generally bound χ_ℓ as a function of $\theta(a\varepsilon^{\alpha})$, ψ_ℓ, a, α, b, and β. In the special case of ℓ equal the 0-1 loss, both τ_ℓ and $\chi_\ell\bar{\ell}$ are equal to $\theta(a(\varepsilon/2)^{\alpha})$.

We note that the values $\hat{\mathfrak{s}}(\gamma, m)$ used in the proof of Theorem 16 have a direct dependence on the parameters b, β, a, and α from Condition 11 and Condition 10. Such a dependence may be undesirable for many applications, where information about these values is not available. However, one can easily follow this same proof, taking $\hat{\mathfrak{s}}(2^{-j}, m) = \mathrm{Log}\left(\frac{12\log_2(4\bar{\ell}2^j)^2\log_2(2m)^2}{\delta}\right)$ instead, which only leads to an increase by a $\log\log$ factor: specifically, replacing the factor of A_1 in (33), and the factors $(A_1 + \mathrm{Log}(B_1))$ and $(A_1 + \mathrm{Log}(C_1))$ in (34), with a factor of $(A_1 + \mathrm{Log}(\mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))))$. It is not clear whether it is always possible to achieve the slightly tighter result of Theorem 16 without having direct access to the values b, β, a, and α in the algorithm.

In the special case when ℓ satisfies Condition 3, we can derive a sometimes-stronger result via Corollary 9. Specifically, we can combine (27), (8), (9), and Lemma 12, to get that if $f^\star \in \mathcal{F}$ and Condition 3 is satisfied, then for $j \geq -\lceil \log_2(\bar{\ell}) \rceil$ in Corollary 9,

$$\text{(35)}\quad \mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-8}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) \lesssim \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + 2^j\bar{\ell}\mathcal{P}(\mathcal{U}_j)\right)\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\bar{\ell}2^{j\beta}\mathcal{P}(\mathcal{U}_j)^\beta/b\right) + s\right),$$

where b and β are as in Lemma 12. Plugging this into Corollary 9, with $\hat{\mathfrak{s}}$ defined analogous to that used in the proof of Theorem 16, and bounding the summations in the conditions for u and n in Corollary 9, we arrive at the following theorem. The details of the proof proceed along similar lines as the proof of Theorem 16, and a sketch of the remaining technical details is included in Appendix A.

THEOREM 17. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10, ℓ is classification-calibrated and satisfies Condition 3, $f^\star \in \mathcal{F}$, Ψ_ℓ is as in (16), and b and β are as in Lemma 12, then for any $\varepsilon \in (0,1)$, letting $\theta = \theta(a\varepsilon^\alpha)$, $A_2 = \mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\left(\bar{\ell}/b\right)\left(a\theta\varepsilon^\alpha/\Psi_\ell(\varepsilon)\right)^\beta\right) + \mathrm{Log}\left(1/\delta\right)$, $B_2 = \min\left\{\frac{1}{1-2^{(\alpha-1)(2-\beta)}}, \mathrm{Log}\left(\bar{\ell}/\Psi_\ell(\varepsilon)\right)\right\}$, and $C_2 = \min\left\{\frac{1}{1-2^{(\alpha-1)}}, \mathrm{Log}\left(\bar{\ell}/\Psi_\ell(\varepsilon)\right)\right\}$, if*

$$\text{(36)}\qquad u \geq c\left(\frac{b\left(a\theta\varepsilon^\alpha\right)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)A_2,$$

and

(37)
$$n \geq c\left(b(A_2 + \mathrm{Log}(B_2))B_2\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}(A_2 + \mathrm{Log}(C_2))C_2\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right),$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function satisfying (12), Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1-\delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.
$\diamond$

Examining the asymptotic dependence on ε in the above result, the sufficient number of unlabeled samples is $O\left(\frac{(\theta\varepsilon^{\alpha})^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}}\mathrm{Log}\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right)$, and the number of label requests is $O\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{2-\beta}\mathrm{Log}\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right)$ in the case that $\alpha < 1$, or $O\left(\theta^{2-\beta}\mathrm{Log}(1/\varepsilon)\mathrm{Log}\left(\theta^{\beta}\mathrm{Log}(1/\varepsilon)\right)\right)$ in the case that $\alpha = 1$. This is noteworthy in the case $\alpha > 0$ and $r_\ell > 2$, for at least two reasons. First, the number of label requests indicated by this result can often be smaller than that indicated by Theorem 16, by a factor of roughly $\tilde{O}\left((\theta\varepsilon^{\alpha})^{1-\beta}\right)$; this is particularly interesting when θ is bounded by a finite constant. The second interesting feature of this result is that even the sufficient number of *unlabeled* samples, as indicated by (36), can often be smaller than the number of *labeled* samples sufficient for ERM_ℓ, as indicated by Theorem 15, again by a factor of roughly $\tilde{O}\left((\theta\varepsilon^{\alpha})^{1-\beta}\right)$. This indicates that, in the case of a surrogate loss ℓ satisfying Condition 3 with $r_\ell > 2$, when Theorem 15 is tight, even if we have complete access to a fully labeled data set, we may still prefer to use Algorithm 1 rather than ERM_ℓ; this is somewhat surprising, since (as (37) indicates) we expect Algorithm 1 to ignore the vast majority of the labels in this case. That said, it is not clear whether there exist natural classification-calibrated losses ℓ satisfying Condition 3 with $r_\ell > 2$ for which the indicated sufficient size of m in Theorem 15 is ever competitive with the known results for methods that directly optimize the empirical 0-1 risk (i.e., Theorem 15 with ℓ the 0-1 loss); thus, the improvements in u and n reflected by Theorem 17 may simply indicate that Algorithm 1 is, to some extent, compensating for a choice of loss ℓ that would otherwise lead to suboptimal label complexities.

We note that, as in Theorem 16, the values $\hat{\mathfrak{s}}$ used to obtain this result have a direct dependence on certain values, which are typically not directly accessible in practice: in this case, a, α, and θ. However, as was the case for Theorem 16, we can obtain only slightly worse results by instead taking $\hat{\mathfrak{s}}(2^{-j}, m) = \mathrm{Log}\left(\frac{12\log_2(4\bar{\ell}2^{j})^2\log_2(2m)^2}{\delta}\right)$, which again only leads to an increase by a $\log\log$ factor: replacing the factor of A_2 in (36), and the factors $(A_2 + \mathrm{Log}(B_2))$ and $(A_2 + \mathrm{Log}(C_2))$ in (37), with a factor of $(A_2 + \mathrm{Log}(\mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))))$. As before, it is not clear whether the slightly tighter result of Theorem 17 is always available, without requiring direct dependence on these quantities.

5.5. *Entropy Conditions.* Next we turn to problems satisfying certain entropy conditions. In particular, the following represent two commonly-studied conditions, which allow for concise statement of results below.

CONDITION 18. *For some* $q \geq 1$, $\rho \in (0,1)$, *and* $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F},\mathcal{P}_{XY}})$, *either*

$\forall \varepsilon > 0$,

$$\ln \mathcal{N}_{[]}(\varepsilon \|\mathrm{F}\|_{\mathcal{P}_{XY}}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})) \leq q\varepsilon^{-2\rho}, \tag{38}$$

or for all finitely discrete P, $\forall \varepsilon > 0$,

$$\ln \mathcal{N}(\varepsilon \|\mathrm{F}\|_{P}, \mathcal{G}_{\mathcal{F}}, L_2(P)) \leq q\varepsilon^{-2\rho}. \tag{39}$$

$\diamond$

In particular, note that when $\mathcal{F}$ satisfies Condition 18, for $0 \leq \sigma \leq 2\|\mathrm{F}\|_{\mathcal{P}_{XY}}$,

$$\mathring{\phi}_{\ell}(\sigma, \mathcal{F}; \mathcal{P}_{XY}, m) \lesssim \max\left\{ \frac{\sqrt{q}\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{\rho}\sigma^{1-\rho}}{(1-\rho)m^{1/2}}, \frac{\bar{\ell}^{\frac{1-\rho}{1+\rho}} q^{\frac{1}{1+\rho}} \|\mathrm{F}\|_{\mathcal{P}_{XY}}^{\frac{2\rho}{1+\rho}}}{(1-\rho)^{\frac{2}{1+\rho}} m^{\frac{1}{1+\rho}}} \right\}. \tag{40}$$

Since $\mathrm{D}_{\ell}([\mathcal{F}]) \leq 2\|\mathrm{F}\|_{\mathcal{P}_{XY}}$, this implies that for any numerical constant $c \in (0,1]$, for every $\gamma \in (0,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 11, then

$$\ddot{\mathrm{M}}_{\ell}(c\gamma, \gamma; \mathcal{F}, \mathcal{P}_{XY}) \lesssim \frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2} \max\left\{ b^{1-\rho}\gamma^{\beta(1-\rho)-2}, \bar{\ell}^{1-\rho}\gamma^{-(1+\rho)} \right\}. \tag{41}$$

Combined with (8), (9), (10), and Theorem 6, taking $\mathfrak{s}(\lambda,\gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at the following classic result [e.g., 6, 35].

THEOREM 19. *For a universal constant* $c \in [1,\infty)$, *if* $\mathcal{P}_{XY}$ *satisfies Condition 10 and Condition 11,* $\mathcal{F}$ *and* $\mathcal{P}_{XY}$ *satisfy Condition 18,* ℓ *is classification-calibrated,* $f^{\star} \in \mathcal{F}$, *and* Ψ_{ℓ} *is as in (16), then for any* $\varepsilon \in (0,1)$ *and* m *with*

$$m \geq c\frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2}\left(\frac{b^{1-\rho}}{\Psi_{\ell}(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}}{\Psi_{\ell}(\varepsilon)^{1+\rho}}\right) + c\left(\frac{b}{\Psi_{\ell}(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_{\ell}(\varepsilon)}\right)\mathrm{Log}\left(\frac{1}{\delta}\right),$$

with probability at least $1-\delta$, $\mathrm{ERM}_{\ell}(\mathcal{F}, \mathcal{Z}_m)$ *produces* $\hat{h}$ *with* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^{\star}) \leq \varepsilon$.

$\diamond$

Next, turning to the analysis of Algorithm 1 under these same conditions, combining (41) with (8), (9), and Theorem 7, we have the following result. The details of the proof follow analogously to the proof of Theorem 16, and are therefore omitted for brevity.

THEOREM 20. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10 and Condition 11, $\mathcal{F}$ and $\mathcal{P}_{XY}$ satisfy Condition 18, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (16), then for any $\varepsilon \in (0,1)$, letting B_1 and C_1 be as in Theorem 16, $B_3 = \min\left\{\frac{1}{1-2^{(\alpha+\beta(1-\rho)-2)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, $C_3 = \min\left\{\frac{1}{1-2^{(\alpha-(1+\rho))}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, and $\theta = \theta(a\varepsilon^\alpha)$, if*

$$u \geq c\frac{q\|\mathrm{F}\|^{2\rho}_{\mathcal{P}_{XY}}}{(1-\rho)^2}\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}\left(\frac{1}{\delta}\right) \tag{42}$$

and

$$n \geq c\theta a\varepsilon^\alpha\frac{q\|\mathrm{F}\|^{2\rho}_{\mathcal{P}_{XY}}}{(1-\rho)^2}\left(\frac{b^{1-\rho}B_3}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}C_3}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\theta a\varepsilon^\alpha\left(\frac{bB_1\mathrm{Log}(B_1/\delta)}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}C_1\mathrm{Log}(C_1/\delta)}{\Psi_\ell(\varepsilon)}\right), \tag{43}$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function satisfying (12), Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1-\delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. ◇

The sufficient size of u in Theorem 20 is essentially identical (up to the constant factors) to the number of labels sufficient for ERM_ℓ to achieve the same, as indicated by Theorem 19. In particular, the dependence on ε in these results is $O\left(\Psi_\ell(\varepsilon)^{\beta(1-\rho)-2}\right)$. On the other hand, when $\theta(\varepsilon^\alpha) = o(\varepsilon^{-\alpha})$, the sufficient size of n in Theorem 20 *does* reflect an improvement in the number of labels indicated by Theorem 19, by a factor with dependence on ε of $O(\theta\varepsilon^\alpha)$.

As before, in the special case when ℓ satisfies Condition 3, we can derive sometimes stronger results via Corollary 9. In this case, we will distinguish between the cases of (39) and (38), as we find a slightly stronger result for the former.

First, suppose (39) is satisfied for all finitely discrete P and all $\varepsilon > 0$, with $\mathrm{F} \leq \bar{\ell}$. Then following the derivation of (41) above, combined with (9), (8), and

Lemma 12, for values of $j \geq -\lceil \log_2(\bar{\ell}) \rceil$ in Corollary 9,

$$\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-8}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) \lesssim \frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\left(b^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{1+\rho}\right) + \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + \bar{\ell}2^j\mathcal{P}(\mathcal{U}_j)\right)s,$$

where q and ρ are from Lemma 12. This immediately leads to the following result by reasoning analogous to the proof of Theorem 17.

THEOREM 21. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10, ℓ is classification-calibrated and satisfies Condition 3, $f^\star \in \mathcal{F}$, Ψ_ℓ is as in* (16), *b and β are as in Lemma 12, and* (39) *is satisfied for all finitely discrete P and all $\varepsilon > 0$, with $\mathrm{F} \leq \bar{\ell}$, then for any $\varepsilon \in (0,1)$, letting B_2 and C_2 be as in Theorem 17, $B_4 = \min\left\{\frac{1}{1-2^{(\alpha-1)(2-\beta(1-\rho))}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, $C_4 = \min\left\{\frac{1}{1-2^{(\alpha-1)(1+\rho)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, and $\theta = \theta(a\varepsilon^\alpha)$, if*

$$u \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\left(\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta(1-\rho)} + \left(\frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{\rho}\right) + c\left(\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}(1/\delta)$$

and

$$n \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\left(B_4 b^{1-\rho}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta(1-\rho)} + C_4\bar{\ell}^{1-\rho}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+\rho}\right) + c\left(B_2\mathrm{Log}(B_2/\delta)b\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + C_2\mathrm{Log}(C_2/\delta)\bar{\ell}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right),$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function satisfying (12), *Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1-\delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.* $\diamond$

Compared to Theorem 20, in terms of the asymptotic dependence on ε, the sufficient sizes for both u and n here may be smaller by a factor of $O\left((\theta\varepsilon^\alpha)^{1-\beta(1-\rho)}\right)$, which sometimes represents a significant refinement, particularly when θ is much smaller than $\varepsilon^{-\alpha}$. In particular, as was the case in Theorem 17, when $\theta(\varepsilon) = o(1/\varepsilon)$, the size of u indicated by Theorem 21 is smaller than the known results for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ from Theorem 19.

The case where (38) is satisfied can be treated similarly, though the result we obtain here is slightly weaker. Specifically, for simplicity suppose (38) is satisfied with $\mathrm{F} = \bar{\ell}$ constant. In this case, we have $\bar{\ell} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}})$ as well, while $\mathcal{N}_{[]}(\varepsilon\bar{\ell}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{\mathcal{U}_j})) = \mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{XY}))$, which is no larger than $\mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY}))$, so that $\mathcal{F}_j$ and $\mathcal{P}_{\mathcal{U}_j}$ also satisfy (38) with $\mathrm{F} = \bar{\ell}$; specifically,

$$\ln \mathcal{N}_{[]}\left(\varepsilon\bar{\ell}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{\mathcal{U}_j})\right) \leq q\mathcal{P}(\mathcal{U}_j)^{-\rho}\varepsilon^{-2\rho}.$$

Thus, based on (41), (8), (9), and Lemma 12, we have that if $f^\star \in \mathcal{F}$ and Condition 3 is satisfied, then for $j \geq -\lceil \log_2(\bar{\ell}) \rceil$ in Corollary 9,

$$\begin{aligned}
&\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-8}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) \\
&\lesssim \left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\mathcal{P}(\mathcal{U}_j)^{-\rho}\left(b^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{1+\rho}\right) \\
&\qquad + \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + \bar{\ell}2^j\mathcal{P}(\mathcal{U}_j)\right)s,
\end{aligned}$$

where b and β are as in Lemma 12. Combining this with Corollary 9 and reasoning analogously to the proof of Theorem 17, we have the following result.

THEOREM 22. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 10, ℓ is classification-calibrated and satisfies Condition 3, $f^\star \in \mathcal{F}$, Ψ_ℓ is as in (16), b and β are as in Lemma 12, and (38) is satisfied with $\mathrm{F} = \bar{\ell}$ constant, then for any $\varepsilon \in (0,1)$, letting B_2 and C_2 be as in Theorem 17, $B_5 = \min\left\{\frac{1}{1-2^{(\alpha-1)(2-\beta(1-\rho))-\alpha\rho}}, \mathrm{Log}\left(\frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\right\}$, $C_5 = \min\left\{\frac{1}{1-2^{\alpha-1-\rho}}, \mathrm{Log}\left(\frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\right\}$, and $\theta = \theta(a\varepsilon^\alpha)$, if*

$$\begin{aligned}
u \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)&\left(\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{(1-\beta)(1-\rho)} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) \\
&+ c\left(\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}(1/\delta)
\end{aligned}$$

and

$$n \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\left(\left(\frac{B_5 b^{1-\rho}}{\Psi_\ell(\varepsilon)^\rho}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+(1-\beta)(1-\rho)} + \frac{C_5\bar{\ell}^{1-\rho}a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\left(bB_2\mathrm{Log}(B_2/\delta)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}C_2\mathrm{Log}(C_2/\delta)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right),$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function satisfying (12)*, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1-\delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.* $\diamond$

In this case, compared to Theorem 20, in terms of the asymptotic dependence on ε, the sufficient sizes for both u and n here may be smaller by a factor of $O\left((\theta\varepsilon^\alpha)^{(1-\beta)(1-\rho)}\right)$, which may sometimes be significant, though not quite as dramatic a refinement as we found under (39) in Theorem 21. As with Theorem 21, when $\theta(\varepsilon) = o(1/\varepsilon)$, the size of u indicated by Theorem 22 is smaller than the known results for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ from Theorem 19.

5.6. *Remarks on VC Major and VC Hull Classes.* Another widely-studied family of function classes includes *VC Major* classes. Specifically, we say $\mathcal{G}$ is a VC Major class with index d if $d = \mathrm{vc}(\{\{z : g(z) \geq t\} : g \in \mathcal{G}, t \in \mathbb{R}\}) < \infty$. We can derive results for VC Major classes, analogously to the above, as follows. For brevity, we leave many of the details as an exercise for the reader. For any VC Major class $\mathcal{G} \subseteq \mathcal{G}^*$ with index d, by reasoning similar to that of Giné and Koltchinskii [15], one can show that if $\mathrm{F} = \bar{\ell}\mathbb{1}_{\mathcal{U}} \geq \mathrm{F}(\mathcal{G})$ for some measurable $\mathcal{U} \subseteq \mathcal{X} \times \mathcal{Y}$, then for any distribution P and $\varepsilon > 0$,

$$\ln \mathcal{N}\left(\varepsilon\|\mathrm{F}\|_P, \mathcal{G}, L_2(P)\right) \lesssim \frac{d}{\varepsilon}\log\left(\frac{\bar{\ell}}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right).$$

This implies that for $\mathcal{F}$ a VC Major class, and ℓ classification-calibrated and either nonincreasing or Lipschitz, if $f^\star \in \mathcal{F}$ and $\mathcal{P}_{XY}$ satisfies Condition 10 and Condition 11, then the conditions of Theorem 7 can be satisfied with the probability bound being at least $1-\delta$, for some $u = \tilde{O}\left(\frac{\theta^{1/2}\varepsilon^{\alpha/2}}{\Psi_\ell(\varepsilon)^{2-\beta/2}} + \Psi_\ell(\varepsilon)^{\beta-2}\right)$ and $n = \tilde{O}\left(\frac{\theta^{3/2}\varepsilon^{3\alpha/2}}{\Psi_\ell(\varepsilon)^{2-\beta/2}} + \theta\varepsilon^\alpha\Psi_\ell(\varepsilon)^{\beta-2}\right)$, where $\theta = \theta(a\varepsilon^\alpha)$, and $\tilde{O}(\cdot)$ hides logarithmic and constant factors. Under Condition 3, with β as in Lemma 12, the conditions of Corollary 9 can be satisfied with the probability bound being at least $1-\delta$, for some $u = \tilde{O}\left(\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta/2}\right)$ and $n = \tilde{O}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta/2}\right)$.

For example, for $\mathcal{X} = [0,1]$ and $\mathcal{F}$ the class of all nondecreasing functions mapping $\mathcal{X}$ to $[-1,1]$, $\mathcal{F}$ is a VC Major class with index 1, and $\theta(0) \leq 2$ for all distributions $\mathcal{P}$. Thus, for instance, if η is nondecreasing and ℓ is the quadratic loss, then $f^\star \in \mathcal{F}$, and Algorithm 1 achieves excess error rate ε with high probability for some $u = \tilde{O}\left(\varepsilon^{2\alpha-3}\right)$ and $n = \tilde{O}\left(\varepsilon^{3(\alpha-1)}\right)$.

VC Major classes are contained in special types of *VC Hull* classes, which are more generally defined as follows. Let $\mathcal{C}$ be a VC Subgraph class of functions on $\mathcal{X}$, with bounded envelope, and for $B \in (0,\infty)$, let $\mathcal{F} = B\mathrm{conv}(\mathcal{C}) = \left\{x \mapsto B\sum_j \lambda_j h_j(x) : \sum_j |\lambda_j| \leq 1, h_j \in \mathcal{C}\right\}$ denote the scaled symmetric convex hull of $\mathcal{C}$; then $\mathcal{F}$ is called a VC Hull class. For instance, these spaces are often used in conjunction with the popular AdaBoost learning algorithm. One can derive results for VC Hull classes following analogously to the above. Specifically, for a VC Hull class $\mathcal{F} = B\mathrm{conv}(\mathcal{C})$ with $d = \mathrm{vc}(\mathcal{C})$, if ℓ is classification-calibrated and Lipschitz, $f^\star \in \mathcal{F}$, and $\mathcal{P}_{XY}$ satisfies Condition 10 and Condition 11, then the conditions of Theorem 7 can be satisfied with the probability bound being at least $1-\delta$, for some $u = \tilde{O}\left((\theta\varepsilon^\alpha)^{\frac{d}{d+2}}\,\Psi_\ell(\varepsilon)^{\frac{2\beta}{d+2}-2}\right)$ and $n = \tilde{O}\left((\theta\varepsilon^\alpha)^{\frac{2d+2}{d+2}}\,\Psi_\ell(\varepsilon)^{\frac{2\beta}{d+2}-2}\right)$. Under Condition 3, with β as in Lemma 12, the conditions of Corollary 9 can be satisfied with the probability bound being at least $1-\delta$, for some $u = \tilde{O}\left(\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\frac{2\beta}{d+2}}\right)$ and $n = \tilde{O}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\frac{2\beta}{d+2}}\right)$. However, it is not clear whether these results for VC Hull classes have any practical implications, since we do not know of any examples of VC Hull classes where these results reflect an improvement over a more direct analysis of ERM_ℓ for these scenarios.

APPENDIX A: PROOFS

PROOF OF THEOREM 7. The proof has two main components: first, showing that, with high probability, $f^\star \in V$ is maintained as an invariant, and second, showing that, with high probability, the set V will be sufficiently reduced to provide the guarantee on $\hat{h}$ after at most the stated number of label requests, given the value of u is as large as stated. Both of these components are served by the following application of Lemma 4.

Let K denote the set of values of $k \in \mathbb{N}$ obtained in Algorithm 1. Let S denote the set of pairs (k',m') such that $k' \in K$ and Algorithm 1 reaches the value $m = m'$ in Step 2 while $k = k'$. For each $k \in K$, let $V^{(k)}$ denote the value of V upon obtaining that value of k in Algorithm 1 (either in Step 0 or Step 8), and let $D_k = \mathrm{DIS}(V^{(k)})$. For each $(k,m) \in S$, let Q_m denote the value of Q in Step 5 on the round that Algorithm 1 obtains that value of m.

Consider any $(k,m) \in S$. Let $\mathcal{L}_m = \{(m_k+1, Y_{m_k+1}), \ldots, (m, Y_m)\}$. Note that $\forall h, g \in V^{(k)}$,

$$(|Q_m| \vee 1) \left(\mathrm{R}_\ell(h; Q_m) - \mathrm{R}_\ell(g; Q_m)\right) = (m - m_k) \left(\mathrm{R}_\ell(h_{D_k}; \mathcal{L}_m) - \mathrm{R}_\ell(g_{D_k}; \mathcal{L}_m)\right), \tag{44}$$

and furthermore that

$$(|Q_m| \vee 1)\hat{U}_\ell(V^{(k)}; Q_m, \hat{\mathfrak{s}}(\hat{\gamma}_k, m - m_k)) = (m - m_k)\hat{U}_\ell(V^{(k)}_{D_k}; \mathcal{L}_m, \hat{\mathfrak{s}}(\hat{\gamma}_k, m - m_k)). \tag{45}$$

Applying Lemma 4 under the conditional distribution given k, $V^{(k)}$, m_k, and $\hat{\gamma}_k$, we have that for any $m > m_k$, on an event of (conditional) probability at least $1 - 6e^{-\hat{\mathfrak{s}}(\hat{\gamma}_k, m-m_k)}$, if $f^\star \in V^{(k)}$ and $(k, m) \in S$, then letting $\hat{u}_{k,m} = \hat{U}_\ell\left(V^{(k)}_{D_k}; \mathcal{L}_m, \hat{\mathfrak{s}}\left(\hat{\gamma}_k, m - m_k\right)\right)$, every $h_{D_k} \in V^{(k)}_{D_k}$ has

$$\mathrm{R}_\ell(h_{D_k}) - \mathrm{R}_\ell(f^\star) < \mathrm{R}_\ell(h_{D_k}; \mathcal{L}_m) - \mathrm{R}_\ell(f^\star; \mathcal{L}_m) + \hat{u}_{k,m}, \tag{46}$$

$$\mathrm{R}_\ell(h_{D_k}; \mathcal{L}_m) - \min_{g_{D_k} \in V^{(k)}_{D_k}} \mathrm{R}_\ell(g_{D_k}; \mathcal{L}_m) < \mathrm{R}_\ell(h_{D_k}) - \mathrm{R}_\ell(f^\star) + \hat{u}_{k,m}, \tag{47}$$

and furthermore

$$\hat{u}_{k,m} < \tilde{U}_\ell\left(V^{(k)}_{D_k}; \mathcal{P}_{XY}, m - m_k, \hat{\mathfrak{s}}\left(\hat{\gamma}_k, m - m_k\right)\right). \tag{48}$$

Let $j_k = \lfloor \log_2(1/\hat{\gamma}_k) \rfloor$ for values of $k \in K$. Then (12) implies $\hat{\mathfrak{s}}(\hat{\gamma}_k, m - m_k) = \mathfrak{s}_{j_k}(m - m_k)$. By a union bound and the law of total probability, on an event of probability at least

$$1 - \mathbb{E}\left[\sum_{k \in K : \hat{\gamma}_k \geq \Gamma_\ell(\varepsilon)/2} \sum_{i=1}^{\log_2(u_{j_k})} 6e^{-\mathfrak{s}_{j_k}(2^i)}\right],$$

for every $(k, m) \in S$ with $\hat{\gamma}_k \geq \Gamma_\ell(\varepsilon)/2$, $m \leq m_k + u_{j_k}$, $\log_2(m - m_k) \in \mathbb{N}$, and $f^\star \in V^{(k)}$, the inequalities (46), (47), and (48) hold. Call this event E. Note that $\hat{\gamma}_k \geq \Gamma_\ell(\varepsilon)/2$ implies $j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$. Furthermore, since each $k \in K$ with $k > 1$ has $\hat{\gamma}_k \leq \hat{\gamma}_{k-1}/2$, and $\hat{\gamma}_1 = \bar{\ell}$, we have $j_{k+1} \geq j_k + 1$ and $j_k \geq k - \lceil \log_2(2\bar{\ell}) \rceil$. This implies $\sum_{k \in K : \hat{\gamma}_k \geq \Gamma_\ell(\varepsilon)/2} \sum_{i=1}^{\log_2(u_{j_k})} 6e^{-\mathfrak{s}_{j_k}(2^i)} \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)}$, so that event E has probability at least

$$1 - \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)}.$$

For the remainder of this proof, we will suppose the event E occurs.

Define $j_0 = -\infty$ and $m_0 = u_{j_0} = 0$. We proceed by induction, establishing the following claims for all $k \in K \cup \{0\}$ having $j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$.

Claim 1: $\max\{m \in \mathbb{N} : (k, m) \in S\} \cup \{m_k\} \leq m_k + u_{j_k}$. If equality is obtained, then we also have $k + 1 \in K$.

Claim 2: If $k+1 \in K$, then $\forall h \in V^{(k+1)}$, $\mathrm{R}_\ell(h_{D_{k+1}}) - \mathrm{R}_\ell(f^\star) < 2\hat{\gamma}_{k+1}$.

Claim 3: If $k+1 \in K$, then $f^\star \in V^{(k+1)}$.

We can think of $k = 0$ as a base case for this inductive proof, since then the first claim is trivially satisfied, while the second claim is satisfied due to $\mathrm{R}_\ell(h_{D_1}) \leq \bar{\ell} < 2\hat{\gamma}_1$, and the third claim is satisfied by assumption (since $V^{(1)} = \mathcal{F}$). Now suppose these three claims hold for k equal $k' - 1$, for some $k' \in \mathbb{N}$ with $k' \in K$ and $j_{k'} \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$.

If it happens that $(k', m_{k'} + u_{j_{k'}}) \in S$, then by definition of $u_{j_{k'}}$ and monotonicity of $m \mapsto \mathring{U}_\ell(\cdot, \cdot; \cdot, m, \cdot)$, we have

$$\mathring{U}_\ell\left(\mathcal{F}_{j_{k'}}, 2^{1-j_{k'}}; \mathcal{P}_{XY}, u_{j_{k'}}, \mathfrak{s}_{j_{k'}}\left(u_{j_{k'}}\right)\right) \leq 2^{-j_{k'}-2}.$$

Plugging in the definition of $j_{k'}$, by (12) and (7), this implies

$$\mathring{U}_\ell\left(\mathcal{F}_{j_{k'}}, 2\hat{\gamma}_{k'}; \mathcal{P}_{XY}, u_{j_{k'}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \leq \hat{\gamma}_{k'}/2. \tag{49}$$

Furthermore, since $f^\star \in \mathcal{F}$, Claim 2 and the definition of $\mathcal{E}_\ell(\cdot)$ imply $V_{D_{k'}}^{(k')} \subseteq [\mathcal{F}]\left(\mathcal{E}_\ell\left(2\hat{\gamma}_{k'}\right); {\scriptstyle 01}\right)$. Since Claim 3 and the definition of $D_{k'}$ imply $\mathrm{sign}(h_{D_{k'}}) = \mathrm{sign}(h)$ for all $h \in V^{(k')}$, we have $\mathrm{er}(h) = \mathrm{er}(h_{D_{k'}})$ for all $h \in V^{(k')}$, so that $V^{(k')} \subseteq [\mathcal{F}]\left(\mathcal{E}_\ell\left(2\hat{\gamma}_{k'}\right); {\scriptstyle 01}\right)$; we also have $V^{k'} \subseteq \mathcal{F}$, so that together these imply

$$V^{(k')} \subseteq \mathcal{F}\left(\mathcal{E}_\ell\left(2\hat{\gamma}_{k'}\right); {\scriptstyle 01}\right) \subseteq \mathcal{F}\left(\mathcal{E}_\ell\left(2^{1-j_{k'}}\right); {\scriptstyle 01}\right). \tag{50}$$

This also implies $D_{k'} \subseteq \mathrm{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{1-j_{k'}}\right); {\scriptstyle 01}\right)\right)$. Combined with (49) and (7), these imply

$$\mathring{U}_\ell\left(V_{D_{k'}}^{(k')}, 2\hat{\gamma}_{k'}; \mathcal{P}_{XY}, u_{j_{k'}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \leq \hat{\gamma}_{k'}/2.$$

Together with (6), this implies

$$\tilde{U}_\ell\left(V_{D_{k'}}^{(k')}\left(2\hat{\gamma}_{k'}; \ell\right); \mathcal{P}_{XY}, u_{j_{k'}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \leq \hat{\gamma}_{k'}/2.$$

Claim 2 implies $V_{D_{k'}}^{(k')} = V_{D_{k'}}^{(k')}\left(2\hat{\gamma}_{k'}; \ell\right)$, which means

$$\tilde{U}_\ell\left(V_{D_{k'}}^{(k')}; \mathcal{P}_{XY}, u_{j_{k'}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \leq \hat{\gamma}_{k'}/2.$$

Since $\log_2(u_{j_{k'}}) \in \mathbb{N}$, $j_{k'} \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$, and Claim 3 implies $f^\star \in V^{(k')}$, combining the above with (48) implies that on the event E,

$$\hat{U}_\ell\left(V^{(k')}_{D_{k'}}; \mathcal{L}_{m_{k'}+u_{j_{k'}}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \leq \hat{\gamma}_{k'}/2.$$

By (45), this also means

$$\hat{U}_\ell\left(V^{(k')}; Q_{m_{k'}+u_{j_{k'}}}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, u_{j_{k'}}\right)\right) \frac{|Q_{m_{k'}+u_{j_{k'}}}| \vee 1}{u_{j_{k'}}} \leq \hat{\gamma}_{k'}/2.$$

The left hand side of this inequality is precisely the value

$$\hat{T}_\ell\left(V^{(k')}; Q_{m_{k'}+u_{j_{k'}}}, m_{k'} + u_{j_{k'}}, k'\right) \frac{|Q_{m_{k'}+u_{j_{k'}}}| \vee 1}{u_{j_{k'}}},$$

so that the condition in Step 5 of Algorithm 1 will be satisfied if and when $k = k'$ and $m = m_{k'} + u_{j_{k'}}$. In summary, we have shown that if $(k', m_{k'} + u_{j_{k'}}) \in S$, then $\max\{m \in \mathbb{N} : (k', m) \in S\} \cup \{m_{k'}\} = m_{k'} + u_{j_{k'}}$ and $k' + 1 \in K$. Furthermore, since $\{m \in \mathbb{N} : (k', m) \in S\} \cup \{m_{k'}\}$ is a sequence of *consecutive* integers including $m_{k'}$, if $(k', m_{k'} + u_{j_{k'}}) \notin S$, then $\max\{m \in \mathbb{N} : (k', m) \in S\} \cup \{m_{k'}\} < m_{k'} + u_{j_{k'}}$. In either case, we have established Claim 1 for k equal to k'.

Next we consider Claim 2 and Claim 3. If $k' + 1 \notin K$, then Claim 2 and Claim 3 are trivially satisfied for k equal to k'. Otherwise, suppose $k' + 1 \in K$. Let $m' = \max\{m \in \mathbb{N} : (k', m) \in S\} \cup \{m_{k'}\}$. By Claim 1, we have $m' \leq m_{k'} + u_{j_{k'}}$. Furthermore, $k' + 1 \in K$ implies that the condition in Step 5 in Algorithm 1 is satisfied for k equal k' and m equal m', so that $\log_2(m' - m_{k'}) \in \mathbb{N}$ and

$$V^{(k'+1)} = \left\{ h \in V^{(k')} : \right.$$
$$\left. \mathrm{R}_\ell(h; Q_{m'}) - \min_{g \in V^{(k')}} \mathrm{R}_\ell(g; Q_{m'}) \leq \hat{U}_\ell\left(V^{(k')}; Q_{m'}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, m' - m_{k'}\right)\right) \right\}.$$

By (44) and the definition of $\hat{\gamma}_{k'+1}$, this is equivalently expressed as

(51)
$$V^{(k'+1)} = \left\{ h \in V^{(k')} : \mathrm{R}_\ell(h_{D_{k'}}; \mathcal{L}_{m'}) - \min_{g \in V^{(k')}} \mathrm{R}_\ell(g_{D_{k'}}; \mathcal{L}_{m'}) \leq \hat{\gamma}_{k'+1} \right\}.$$

By Claim 3, $f^\star \in V^{(k')}$; thus, (46) implies that on the event E, every $h \in V^{(k'+1)}$ has

$$\mathrm{R}_\ell(h_{D_{k'}}) - \mathrm{R}_\ell(f^\star) < \hat{\gamma}_{k'+1} + \hat{U}_\ell\left(V^{(k')}_{D_{k'}}; \mathcal{L}_{m'}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, m' - m_{k'}\right)\right).$$

By (45), this is equivalently expressed as

$$\mathrm{R}_\ell(h_{D_{k'}}) - \mathrm{R}_\ell(f^\star) < 2\hat{\gamma}_{k'+1}.$$

Since $\mathrm{R}_\ell(h_{D_{k'+1}}) \leq \mathrm{R}_\ell(h_{D_{k'}})$, we have established Claim 2 for k equal to k'.

Furthermore, Claim 3 implies $f^\star \in V^{(k')}$, so that by (47), on the event E, we have

$$\mathrm{R}_\ell(f^\star; \mathcal{L}_{m'}) - \min_{g \in V^{(k')}} \mathrm{R}_\ell(g_{D_{k'}}; \mathcal{L}_{m'}) < \hat{U}_\ell\left(V^{(k')}_{D_{k'}}; \mathcal{L}_{m'}, \hat{\mathfrak{s}}\left(\hat{\gamma}_{k'}, m' - m_{k'}\right)\right).$$

By (45) and the definition of $\hat{\gamma}_{k'+1}$, the right hand side of this inequality is equal to $\hat{\gamma}_{k'+1}$. In particular, combined with (51), this implies $f^\star \in V^{(k'+1)}$, which establishes Claim 3 for k equal to k'.

Finally, note that j_k is nondecreasing, so that the values of $k \in K \cup \{0\}$ with $j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$ form a sequence of consecutive integers starting with 0. Thus, by the principle of induction, these three claims hold (on event E) for all $k \in K$ for which $j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$.

Similar to above, by Claim 2, for all $k \in K$ with $j_{k-1} \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$,

$$V^{(k)}_{D_k} \subseteq \mathcal{F}^*\left(\mathcal{E}_\ell\left(2\hat{\gamma}_k\right); {\scriptstyle 01}\right). \tag{52}$$

Since every $h \in V^{(k)}$ has $\mathrm{sign}(h(x)) = \mathrm{sign}(f^\star(x)) = \mathrm{sign}(h_{D_k}(x))$ for all $x \notin D_k$, we have that $\forall h \in V^{(k)}$, $\mathrm{er}(h) = \mathrm{er}(h_{D_k})$. Thus, since $V^{(k)} \subseteq \mathcal{F}$ and $f^\star \in \mathcal{F}$, (52) implies

$$V^{(k)} \subseteq \mathcal{F}\left(\mathcal{E}_\ell\left(2\hat{\gamma}_k\right); {\scriptstyle 01}\right). \tag{53}$$

In particular, letting $k^* = \max\{k \in K : j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor\}$, if $k^* + 1 \in K$, then $j_{k^*+1} > \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$, so that $\hat{\gamma}_{k^*+1} < \Gamma_\ell(\varepsilon)/2$, which means $\mathcal{E}_\ell(2\hat{\gamma}_{k^*+1}) \leq \varepsilon$. Together with (53), this implies $V^{(k^*+1)} \subseteq \mathcal{F}^*(\varepsilon; {\scriptstyle 01})$. Since the update in Step 7 always keeps at least one element in V, the function $\hat{h}$ in Step 9 exists, and has $\hat{h} \in V^{(\max K)} = \bigcap_{k \in K} V^{(k)} \subseteq V^{(k^*+1)} \subseteq \mathcal{F}^*(\varepsilon; {\scriptstyle 01})$, so that $\mathrm{er}\left(\hat{h}\right) - \mathrm{er}\left(f^\star\right) \leq \varepsilon$, as claimed.

All that remains is to bound the sizes of u and n sufficient to guarantee $k^* + 1 \in K$. By Claim 1, $k^* + 1 \in K$ would be guaranteed as long as

$$u \geq \sum_{k=1}^{k^*} u_{j_k} \quad \text{and} \quad n \geq \sum_{m=m_{k^*}+1}^{m_{k^*}+u_{j_{k^*}}} \mathbb{1}_{D_{k^*}}(X_m) + \sum_{k=1}^{k^*-1} \sum_{m=m_k+1}^{m_{k+1}} \mathbb{1}_{D_k}(X_m). \tag{54}$$

Since every $k \leq k^*$ has $-\lceil \log_2(\bar{\ell}) \rceil \leq j_k \leq \lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor$, and (as noted above) $j_k \geq j_{k-1} + 1$, we have that

$$\sum_{k=1}^{k^*} u_{j_k} \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} u_j. \tag{55}$$

Furthermore, for all $k \leq k^*$, (53) and monotonicity imply that

$$D_k \subseteq \mathrm{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{1-j_k}\right); {}_{01}\right)\right) = \mathrm{DIS}\left(\mathcal{F}_{j_k}\right) = \mathcal{U}_{j_k},$$

so that

$$\sum_{m=m_{k^*}+1}^{m_{k^*}+u_{j_{k^*}}} \mathbb{1}_{D_{k^*}}(X_m) + \sum_{k=1}^{k^*-1} \sum_{m=m_k+1}^{m_{k+1}} \mathbb{1}_{D_k}(X_m) \leq \sum_{m=m_{k^*}+1}^{m_{k^*}+u_{j_{k^*}}} \mathbb{1}_{\mathcal{U}_{j_{k^*}}}(X_m) + \sum_{k=1}^{k^*-1} \sum_{m=m_k+1}^{m_{k+1}} \mathbb{1}_{\mathcal{U}_{j_k}}(X_m).$$

Since $\mathcal{U}_j$ is nonincreasing in j, we have that $\mathbb{1}_{\mathcal{U}_j}(X_m)$ is nonincreasing in j for all m. Combining this with Claim 1 and the above properties of j_k, we have

$$\sum_{m=m_{k^*}+1}^{m_{k^*}+u_{j_{k^*}}} \mathbb{1}_{\mathcal{U}_{j_{k^*}}}(X_m) + \sum_{k=1}^{k^*-1} \sum_{m=m_k+1}^{m_{k+1}} \mathbb{1}_{\mathcal{U}_{j_k}}(X_m) \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{m=1+\sum_{i=-\lceil \log_2(\bar{\ell}) \rceil}^{j-1} u_i}^{\sum_{i=-\lceil \log_2(\bar{\ell}) \rceil}^{j} u_j} \mathbb{1}_{\mathcal{U}_j}(X_m).$$

In summary, we have

$$\sum_{m=m_{k^*}+1}^{m_{k^*}+u_{j_{k^*}}} \mathbb{1}_{D_{k^*}}(X_m) + \sum_{k=1}^{k^*-1} \sum_{m=m_k+1}^{m_{k+1}} \mathbb{1}_{D_k}(X_m) \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{m=1+\sum_{i=-\lceil \log_2(\bar{\ell}) \rceil}^{j-1} u_i}^{\sum_{i=-\lceil \log_2(\bar{\ell}) \rceil}^{j} u_j} \mathbb{1}_{\mathcal{U}_j}(X_m). \tag{56}$$

Note that the indicators $\mathbb{1}_{\mathcal{U}_j}(X_m)$ in the summation on the right hand side of (56) are independent, so that a Chernoff bound implies that on an event E' of probability at least $1 - 2^{-s}$,

(57)
$$\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \sum_{m=1+\sum_{i=-\lceil \log_2(\bar{\ell})\rceil}^{j-1} u_i}^{\sum_{i=-\lceil \log_2(\bar{\ell})\rceil}^{j} u_j} \mathbb{1}_{\mathcal{U}_j}(X_m) \leq s + 2e \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \mathcal{P}(\mathcal{U}_j)u_j.$$

Combining (54), (55), (56), and (57) implies that, for u and n as in the statement of Theorem 7, on the event $E \cap E'$, we have $k^* + 1 \in K$. A union bound implies that the event $E \cap E'$ has probability at least

$$1 - 2^{-s} - \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)},$$

as required. □

PROOF OF LEMMA 8. If $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) = 0$, then $\phi_\ell(\mathcal{H}; m, P) = 0$, so that in this case, $\mathring{\phi}'_\ell$ trivially satisfies (5). Otherwise, suppose $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$. By the classic symmetrization inequality [e.g., 35, Lemma 2.3.1],

$$\phi_\ell(\mathcal{H}, m, P) \leq 2\mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q, \Xi_{[m]})\right|\right],$$

where $Q \sim P^m$ and $\Xi_{[m]} = \{\xi_1, \ldots, \xi_m\} \sim \mathrm{Uniform}(\{-1, +1\}^m)$ are independent. Fix any measurable $\mathcal{U} \supseteq \overline{\mathrm{DISF}}(\mathcal{H})$. Then

(58) $$\mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q, \Xi_{[m]})\right|\right] = \mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q \cap \mathcal{U}, \Xi_{[|Q\cap\mathcal{U}|]})\right| \frac{|Q \cap \mathcal{U}|}{m}\right],$$

where $\Xi_{[q]} = \{\xi_1, \ldots, \xi_q\}$ for any $q \in \{0, \ldots, m\}$. By the classic desymmetrization inequality [see e.g., 24], applied under the conditional distribution given $|Q \cap \mathcal{U}|$, the right hand side of (58) is at most

(59)
$$\mathbb{E}\left[2\phi_\ell(\mathcal{H}, |Q \cap \mathcal{U}|, P_{\mathcal{U}})\frac{|Q \cap \mathcal{U}|}{m}\right] + \sup_{h,g\in\mathcal{H}} |\mathrm{R}_\ell(h; P_{\mathcal{U}}) - \mathrm{R}_\ell(g; P_{\mathcal{U}})| \frac{\mathbb{E}\left[\sqrt{|Q \cap \mathcal{U}|}\right]}{m}.$$

By Jensen's inequality, the second term in (59) is at most

$$\sup_{h,g\in\mathcal{H}} |\mathrm{R}_\ell(h; P_{\mathcal{U}}) - \mathrm{R}_\ell(g; P_{\mathcal{U}})| \sqrt{\frac{P(\mathcal{U})}{m}} \leq \mathrm{D}_\ell(\mathcal{H}; P_{\mathcal{U}})\sqrt{\frac{P(\mathcal{U})}{m}} = \mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{1}{m}}.$$

Decomposing based on $|Q \cap \mathcal{U}|$, the first term in (59) is at most

(60) $$\mathbb{E}\left[2\phi_\ell(\mathcal{H}, |Q \cap \mathcal{U}|, P_{\mathcal{U}})\frac{|Q \cap \mathcal{U}|}{m}\mathbb{1}\left[|Q \cap \mathcal{U}| \geq (1/2)P(\mathcal{U})m\right]\right] + 2\bar{\ell}P(\mathcal{U})\mathbb{P}\left(|Q \cap \mathcal{U}| < (1/2)P(\mathcal{U})m\right).$$

Since $|Q \cap \mathcal{U}| \geq (1/2)P(\mathcal{U})m \Rightarrow |Q \cap \mathcal{U}| \geq \lceil (1/2)P(\mathcal{U})m \rceil$, and $\phi_\ell(\mathcal{H}, q, P_\mathcal{U})$ is nonincreasing in q, the first term in (60) is at most

$$2\phi_\ell(\mathcal{H}, \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U})\mathbb{E}\left[\frac{|Q \cap \mathcal{U}|}{m}\right] = 2\phi_\ell(\mathcal{H}, \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U})P(\mathcal{U}),$$

while a Chernoff bound implies the second term in (60) is at most

$$2\bar{\ell}P(\mathcal{U})\exp\{-P(\mathcal{U})m/8\} \leq \frac{16\bar{\ell}}{m}.$$

Plugging back into (59), we have
(61)
$$\phi_\ell(\mathcal{H}, m, P) \leq 4\phi_\ell(\mathcal{H}, \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U})P(\mathcal{U}) + \frac{32\bar{\ell}}{m} + 2\mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{1}{m}}.$$

Next, note that, for any $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$, $\frac{\sigma}{\sqrt{P(\mathcal{U})}} \geq \mathrm{D}_\ell(\mathcal{H}; P_\mathcal{U})$. Also, if $\mathcal{U} = \mathcal{U}' \times \mathcal{Y}$ for some $\mathcal{U}' \supseteq \mathrm{DISF}(\mathcal{H})$, then $f^\star_{P_\mathcal{U}} = f^\star_P$, so that if $f^\star_P \in \mathcal{H}$, (5) implies

$$\phi_\ell(\mathcal{H}, \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U}) \leq \mathring{\phi}_\ell\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U}\right). \tag{62}$$

Combining (61) with (62), we see that $\mathring{\phi}'_\ell$ satisfies the condition (5) of Definition 5.

Furthermore, by the fact that $\mathring{\phi}_\ell$ satisfies (4) of Definition 5, combined with the monotonicity imposed by the infimum in the definition of $\mathring{\phi}'_\ell$, it is easy to check that $\mathring{\phi}'_\ell$ also satisfies (4) of Definition 5. In particular, note that any $\mathcal{H}'' \subseteq \mathcal{H}' \subseteq [\mathcal{F}]$ and $\mathcal{U}'' \subseteq \mathcal{X}$ have $\mathrm{DISF}(\mathcal{H}''_{\mathcal{U}''}) \subseteq \mathrm{DISF}(\mathcal{H}')$, so that the range of $\mathcal{U}$ in the infimum is never smaller for $\mathcal{H} = \mathcal{H}''_{\mathcal{U}''}$ relative to that for $\mathcal{H} = \mathcal{H}'$. □

PROOF OF COROLLARY 9. Let $\mathring{\phi}'_\ell$ be as in Lemma 8, and define for any $m \in \mathbb{N}$, $s \in [1, \infty)$, $\zeta \in [0, \infty]$, and $\mathcal{H} \subseteq [\mathcal{F}]$,

$$\begin{aligned}&\mathring{U}'_\ell(\mathcal{H}, \zeta; \mathcal{P}_{XY}, m, s)\\ &\quad = \tilde{K}\left(\mathring{\phi}'_\ell(\mathrm{D}_\ell([\mathcal{H}](\zeta; \ell)), \mathcal{H}; m, \mathcal{P}_{XY}) + \mathrm{D}_\ell([\mathcal{H}](\zeta; \ell))\sqrt{\frac{s}{m}} + \frac{\bar{\ell}s}{m}\right).\end{aligned}$$

That is, $\mathring{U}'_\ell$ is the function $\mathring{U}_\ell$ that would result from using $\mathring{\phi}'_\ell$ in place of $\mathring{\phi}_\ell$. Let $\mathcal{U} = \mathrm{DISF}(\mathcal{H})$, and suppose $\mathcal{P}(\mathcal{U}) > 0$. Then since $\mathrm{DISF}([\mathcal{H}]) = \mathrm{DISF}(\mathcal{H})$ implies

$$\begin{aligned}\mathrm{D}_\ell([\mathcal{H}](\zeta; \ell)) &= \mathrm{D}_\ell([\mathcal{H}](\zeta; \ell); \mathcal{P}_\mathcal{U})\sqrt{\mathcal{P}(\mathcal{U})}\\ &= \mathrm{D}_\ell([\mathcal{H}](\zeta/\mathcal{P}(\mathcal{U}); \ell, \mathcal{P}_\mathcal{U}); \mathcal{P}_\mathcal{U})\sqrt{\mathcal{P}(\mathcal{U})},\end{aligned}$$

a little algebra reveals that for $m \geq 2\mathcal{P}(\mathcal{U})^{-1}$,

$$\mathring{U}'_\ell(\mathcal{H},\zeta;\mathcal{P}_{XY},m,s) \leq 33\mathcal{P}(\mathcal{U})\mathring{U}_\ell(\mathcal{H},\zeta/\mathcal{P}(\mathcal{U});\mathcal{P}_{\mathcal{U}},\lceil(1/2)\mathcal{P}(\mathcal{U})m\rceil,s). \tag{63}$$

In particular, for $j \geq -\lceil\log_2(\bar{\ell})\rceil$, taking $\mathcal{H} = \mathcal{F}_j$, we have (from the definition of $\mathcal{F}_j$) $\mathcal{U} = \mathrm{DISF}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H}) = \mathcal{U}_j$, so that when $\mathcal{P}(\mathcal{U}_j) > 0$, any

$$m \geq 2\mathcal{P}(\mathcal{U}_j)^{-1}\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-2}}{33\mathcal{P}(\mathcal{U}_j)},\frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)};\mathcal{F}_j,\mathcal{P}_{\mathcal{U}_j},\mathfrak{s}_j(m)\right)$$

suffices to make the right side of (63) (with $s = \mathfrak{s}_j(m)$ and $\zeta = 2^{1-j}$) at most 2^{-j-2}; in particular, this means taking u_j equal to any such m (with $\log_2(m) \in \mathbb{N}$) suffices to satisfy (13) (with the $\mathring{\mathrm{M}}_\ell$ in (13) defined with respect to the $\mathring{\phi}'_\ell$ function); monotonicity of $\zeta \mapsto \mathring{\mathrm{M}}_\ell\left(\zeta,\frac{2^{1-j}}{\mathcal{P}(\mathcal{U}_j)};\mathcal{F}_j,\mathcal{P}_{\mathcal{U}_j},\mathfrak{s}_j(m)\right)$ implies (15) is a sufficient condition for this. In the special case where $\mathcal{P}(\mathcal{U}_j) = 0$, $\mathring{U}'_\ell(\mathcal{F}_j,2^{1-j};\mathcal{P}_{XY},m,s) = \tilde{K}\frac{\bar{\ell}s}{m}$, so that taking $u_j \geq \tilde{K}\bar{\ell}\mathfrak{s}_j(u_j)2^{j+2}$ suffices to satisfy (13) (again, with the $\mathring{\mathrm{M}}_\ell$ in (13) defined in terms of $\mathring{\phi}'_\ell$). Plugging these values into Theorem 7 completes the proof. □

PROOF OF THEOREM 16. For $-\lceil\log_2(\bar{\ell})\rceil \leq j \leq \lfloor\log_2(2/\Psi_\ell(\varepsilon))\rfloor$, let $s_j = \mathrm{Log}\left(\frac{48(\lfloor\log_2(8/\Psi_\ell(\varepsilon))\rfloor-j)^2}{\delta}\right)$, and define $u_j = 2^{\lceil\log_2(u'_j)\rceil}$, where

$$u'_j = c'\left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(\mathrm{vc}\left(\mathcal{G}_{\mathcal{F}}\right)\mathrm{Log}\left(\chi_\ell\bar{\ell}\right) + s_j\right), \tag{64}$$

for an appropriate universal constant $c' \in [1,\infty)$. Note that, by (32), (8), and (9), we can choose the constant c' so that these u_j satisfy (13) when we define

$$m \mapsto \mathfrak{s}_j(m) = \mathrm{Log}\left(\frac{12\log_2\left(4u_j/m\right)^2\left(\lfloor\log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right)^2}{\delta}\right).$$

Additionally, let $s = \log_2(2/\delta)$.

Next, note that

$$\begin{aligned}
\sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Gamma_\ell(\varepsilon))\rfloor} u_j &\leq \sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Psi_\ell(\varepsilon))\rfloor} u_j \\
&\leq 2c'\left(\frac{8b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{4\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\mathrm{vc}\left(\mathcal{G}_{\mathcal{F}}\right)\mathrm{Log}\left(\chi_\ell\bar{\ell}\right) + \mathrm{Log}\left(\frac{48}{\delta}\right)\right) \\
&\quad + 4c'\left(\frac{2b}{\Psi_\ell(\varepsilon)^{1-\beta}} + \bar{\ell}\right)\sum_{j=-\lceil\log_2(\bar{\ell})\rceil}^{\lfloor\log_2(2/\Psi_\ell(\varepsilon))\rfloor} 2^j\mathrm{Log}\left(\lfloor\log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right).
\end{aligned} \tag{65}$$

We can bound this last summation by noting that

$$\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} 2^j \mathrm{Log}\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right) \tag{66}$$
$$\leq \frac{2}{\Psi_\ell(\varepsilon)} \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} 2^{j-\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \mathrm{Log}\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right)$$
$$\leq \frac{2}{\Psi_\ell(\varepsilon)} \sum_{i=0}^{\infty} 2^{-i}\mathrm{Log}\left(2+i\right) \leq \frac{2}{\Psi_\ell(\varepsilon)} \sum_{i=0}^{\infty} 2^{-i}(i+1) = \frac{8}{\Psi_\ell(\varepsilon)}.$$

Plugging this into (65), we have that $\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} u_j$ is at most

$$8c'\left(\frac{2b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\mathrm{vc}\left(\mathcal{G}_{\mathcal{F}}\right)\mathrm{Log}\left(\chi_\ell \bar{\ell}\right) + \mathrm{Log}\left(\frac{48e^4}{\delta}\right)\right).$$

Thus, by choosing $c \geq 160c'$, any u satisfying (33) has $u \geq \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} u_j$, as required by Theorem 7.

For $\mathcal{U}_j$ as in Theorem 7, note that by Condition 10 and the definition of θ,

$$\mathcal{P}\left(\mathcal{U}_j\right) = \mathcal{P}\left(\mathrm{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{1-j}\right);{\scriptstyle 01}\right)\right)\right) \leq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star, a\mathcal{E}_\ell\left(2^{1-j}\right)^\alpha\right)\right)\right)$$
$$\leq \theta \max\left\{a\mathcal{E}_\ell\left(2^{1-j}\right)^\alpha, a\varepsilon^\alpha\right\} \leq \theta\max\left\{a\Psi_\ell^{-1}\left(2^{1-j}\right)^\alpha, a\varepsilon^\alpha\right\}.$$

Because Ψ_ℓ is strictly increasing on $(0,1)$, for $j \leq \lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor$, $\Psi_\ell^{-1}\left(2^{1-j}\right) \geq \varepsilon$, so that this last expression is equal to $\theta a \Psi_\ell^{-1}\left(2^{1-j}\right)^\alpha$. This implies

$$\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \mathcal{P}\left(\mathcal{U}_j\right)u_j \leq \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \mathcal{P}\left(\mathcal{U}_j\right)u_j \lesssim$$

$$\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} a\theta\Psi_\ell^{-1}\left(2^{1-j}\right)^\alpha\left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(A_1 + \mathrm{Log}(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j)\right). \tag{67}$$

We can change the order of summation in the above expression by letting $i = \lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor - j$ and summing from 0 to $N = \lceil \log_2(\bar{\ell})\rceil + \lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor$. In particular, since $2^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \leq 2/\Psi_\ell(\varepsilon)$, (67) is at most

$$\sum_{i=0}^{N} a\theta\Psi_\ell^{-1}\left(2^{1-\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor}2^i\right)^\alpha\left(\frac{4b2^{i(\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{2\bar{\ell}2^{-i}}{\Psi_\ell(\varepsilon)}\right)\left(A_1 + \mathrm{Log}(i+2)\right). \tag{68}$$

Since $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing on $(0,\infty)$, $\Psi_\ell^{-1}\left(2^{1-\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor}2^i\right) \leq 2^{i+1}\Psi_\ell^{-1}\left(2^{-\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor}\right)$, and since Ψ_ℓ^{-1} is increasing, this latter expression is at most $2^{i+1}\Psi_\ell^{-1}(\Psi_\ell(\varepsilon)) = 2^{i+1}\varepsilon$. Thus, (68) is at most

$$8a\theta\varepsilon^\alpha \sum_{i=0}^{N}\left(\frac{b2^{i(\alpha+\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}2^{i(\alpha-1)}}{\Psi_\ell(\varepsilon)}\right)(A_1 + \mathrm{Log}(i+2)). \tag{69}$$

In general, $\mathrm{Log}(i+2) \leq \mathrm{Log}(N+2)$, so that $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)}(A_1+\mathrm{Log}(i+2)) \leq (A_1+\mathrm{Log}(N+2))(N+1)$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_1+\mathrm{Log}(i+2)) \leq (A_1+\mathrm{Log}(N+2))(N+1)$. When $\alpha+\beta<2$, we also have $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)} \leq \sum_{i=0}^{\infty} 2^{i(\alpha+\beta-2)} = \frac{1}{1-2^{(\alpha+\beta-2)}}$ and $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \sum_{i=0}^{\infty} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \frac{2}{1-2^{(\alpha+\beta-2)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha+\beta-2)}}\right)$. Similarly, if $\alpha<1$, $\sum_{i=0}^{N} 2^{i(\alpha-1)} \leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)} = \frac{1}{1-2^{(\alpha-1)}}$ and likewise $\sum_{i=0}^{N} 2^{i(\alpha-1)}\mathrm{Log}(i+2) \leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)}\mathrm{Log}(i+2) \leq \frac{2}{1-2^{(\alpha-1)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)}}\right)$. By combining these observations (along with a convention that $\frac{1}{1-2^{(\alpha-1)}} = \infty$ when $\alpha = 1$, and $\frac{1}{1-2^{(\alpha+\beta-2)}} = \infty$ when $\alpha=\beta=1$), we find that (69) is

$$\lesssim a\theta\varepsilon^\alpha\left(\frac{b(A_1+\mathrm{Log}(B_1))B_1}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1+\mathrm{Log}(C_1))C_1}{\Psi_\ell(\varepsilon)}\right).$$

Thus, for an appropriately large numerical constant c, any n satisfying (34) has

$$n \geq s + 2e \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \mathcal{P}(\mathcal{U}_j)u_j,$$

as required by Theorem 7.

Finally, we need to show the success probability from Theorem 7 is at least $1-\delta$, for $\mathfrak{s}_j$ and s as above. Toward this end, note that

$$\begin{aligned}
&\sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon))\rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)} \\
&\leq \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \sum_{i=1}^{\log_2(u_j)} \frac{\delta}{2\left(\log_2(4u_j)-i\right)^2\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right)^2} \\
&= \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \sum_{t=1}^{\log_2(u_j)} \frac{\delta}{2(t+1)^2\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right)^2} \\
&< \sum_{j=-\lceil \log_2(\bar{\ell})\rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon))\rfloor} \frac{\delta}{2\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon))\rfloor - j\right)^2} < \sum_{t=1}^{\infty}\frac{\delta}{2(t+1)^2} < \delta/2.
\end{aligned}$$

Noting that $2^{-s} = \delta/2$, we find that indeed

$$1 - 2^{-s} - \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)} \geq 1 - \delta.$$

Thus, by taking $\hat{\mathfrak{s}}$ to be the function satisfying (12) such that $\hat{\mathfrak{s}}(2^{-j}, \cdot) = \mathfrak{s}_j(\cdot)$ for all $j \in \mathbb{Z}$, Theorem 7 now implies the stated result. □

PROOF SKETCH OF THEOREM 17. The proof follows analogously to that of Theorem 16, with the exception that now, for each integer j with $-\lceil \log_2(\bar{\ell}) \rceil \leq j \leq \lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor$, we replace the definition of u'_j from (64) with the following definition. Letting $c_j = \mathrm{vc}(\mathcal{G}_\mathcal{F}) \mathrm{Log}\left(\left(\bar{\ell}/b\right) \left(a\theta 2^j \Psi_\ell^{-1}(2^{1-j})^\alpha \right)^\beta \right)$, define

$$u'_j = c' \left(b 2^{j(2-\beta)} \left(a\theta \Psi_\ell^{-1}(2^{1-j})^\alpha \right)^{1-\beta} + \bar{\ell} 2^j \right) (c_j + s_j),$$

where $c' \in [1, \infty)$ is an appropriate universal constant, and s_j is as in the proof of Theorem 16. With this substitution in place, the values u_j and s, and functions $\mathfrak{s}_j$ and $\hat{\mathfrak{s}}$, are then defined as in the proof of Theorem 16. By (35), (9), (8), and Lemma 12, we can choose the constant c' so that these u_j satisfy (15). By an identical argument to that used in Theorem 16, we have

$$1 - 2^{-s} - \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \sum_{i=1}^{\log_2(u_j)} 6e^{-\mathfrak{s}_j(2^i)} \geq 1 - \delta.$$

It remains only to show that any values of u and n satisfying (36) and (37), respectively, necessarily also satisfy the respective conditions for u and n in Corollary 9.

Toward this end, note that since $x \mapsto x\Psi_\ell^{-1}(1/x)$ is nondecreasing on $(0, \infty)$, we have that

$$\begin{aligned}
&\sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} u_j \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} u_j \\
&\lesssim \left(b \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{1-\beta} + \bar{\ell} \right) \left(\frac{A_2}{\Psi_\ell(\varepsilon)} + \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} 2^j \mathrm{Log}\left(\lfloor \log_2(8/\Psi_\ell(\varepsilon)) \rfloor - j \right) \right) \\
&\lesssim \left(\frac{b\left(a\theta\varepsilon^\alpha\right)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) A_2,
\end{aligned}$$

where this last inequality is due to (66). Thus, for an appropriate choice of c, any u satisfying (36) has $u \geq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} u_j$, as required by Corollary 9.

Finally, note that for $\mathcal{U}_j$ as in Theorem 7, and $i_j = \lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor - j$,

$$\sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \mathcal{P}(\mathcal{U}_j) u_j \leq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} a\theta \Psi_\ell^{-1}(2^{1-j})^\alpha u_j$$
$$\lesssim \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} b \left(a\theta 2^j \Psi_\ell^{-1}(2^{1-j})^\alpha \right)^{2-\beta} (A_2 + \mathrm{Log}\,(i_j + 2))$$
$$+ \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} \bar{\ell} a\theta 2^j \Psi_\ell^{-1}(2^{1-j})^\alpha (A_2 + \mathrm{Log}\,(i_j + 2)) .$$

By changing the order of summation, now summing over values of i_j from 0 to $\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor + \lceil \log_2(\bar{\ell}) \rceil$, letting $N = \lceil \log_2(2\bar{\ell}/\Psi_\ell(\varepsilon)) \rceil$, and noting $2^{\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} \leq 2/\Psi_\ell(\varepsilon)$, and $\Psi_\ell^{-1}(2^{-\lfloor \log_2(2/\Psi_\ell(\varepsilon)) \rfloor} 2^{1+i}) \leq 2^{1+i}\varepsilon$ for $i \geq 0$, this last expression is

$$\lesssim \sum_{i=0}^{N} b \left(\frac{a\theta 2^{i(\alpha-1)} \varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{2-\beta} (A_2 + \mathrm{Log}\,(i+2))$$
$$+ \sum_{i=0}^{N} \frac{\bar{\ell} a\theta 2^{i(\alpha-1)} \varepsilon^\alpha}{\Psi_\ell(\varepsilon)} (A_2 + \mathrm{Log}\,(i+2)) . \tag{70}$$

Considering these sums separately, we have $\sum_{i=0}^N 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq (N+1)(A_2 + \mathrm{Log}(N+2))$ and $\sum_{i=0}^N 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \leq (N+1)(A_2 + \mathrm{Log}(N+2))$. When $\alpha < 1$, we also have $\sum_{i=0}^N 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq \sum_{i=0}^\infty 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{2}{1-2^{(\alpha-1)(2-\beta)}} \mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)(2-\beta)}}\right) + \frac{1}{1-2^{(\alpha-1)(2-\beta)}} A_2$, and similarly $\sum_{i=0}^N 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{1}{1-2^{(\alpha-1)}} A_2 + \frac{2}{1-2^{(\alpha-1)}} \mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)}}\right)$. Thus, generally $\sum_{i=0}^N 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \lesssim B_2(A_2 + \mathrm{Log}(B_2))$ and $\sum_{i=0}^N 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \lesssim C_2(A_2 + \mathrm{Log}(C_2))$. Plugging this into (70), we find that for an appropriately large numerical constant c, any n satisfying (37) has $n \geq \sum_{j=-\lceil \log_2(\bar{\ell}) \rceil}^{\lfloor \log_2(2/\Gamma_\ell(\varepsilon)) \rfloor} \mathcal{P}(\mathcal{U}_j) u_j$, as required by Corollary 9. □

REFERENCES

[1] K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987. 5.2

[2] J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. 1.1, 2.2

[3] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006. 4

[4] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 4

[5] M.-F. Balcan, S. Hanneke, and J. W. Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010. 5.2, 5.2

[6] P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. 1, 1.1, 2.1, 2.1, 2.1, 2.2, 2.2, 2.2, 2.2, 2.2, 2.4, 3.2, 4, 5.1, 5.1, 5.1, 5.1, 5.1, 5.5

[7] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 1, 2.4

[8] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26*$^{\text{th}}$ *International Conference on Machine Learning*, 2009. 1.1, 3.2, 4, 5.2

[9] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 5.4

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 2.2

[11] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007. 4, 5.2

[12] R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15 (4):1306–1326, 1987. 5.4

[13] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. 2.2

[14] E. Friedman. Active learning for smooth problems. In *Proceedings of the 22*nd *Conference on Learning Theory*, 2009. 5.2, 5.2

[15] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 2.4, 2.4, 3.1, 4, 5.2, 5.3, 5.4, 5.4, 5.4, 5.4, 5.6

[16] E. Giné, V. Koltchinskii, and J. Wellner. Ratio limit theorems for empirical processes. In *Stochastic Inequalities*, pages 249–278. Birkhäuser, 2003. 5.3

[17] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24*th *International Conference on Machine Learning*, 2007. 5.2, 5.2

[18] S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009. 1, 5.2

[19] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011. 1, 5.1, 5.2, 5.2, 5.4, 5.4

[20] S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13:1469–1587, 2012. 4, 4, 5.1, 5.2, 5.2

[21] S. Hanneke and L. Yang. Negative results for active learning with convex losses. In *Proceedings of the* 13^{th} *International Conference on Artificial Intelligence and Statistics*, 2010. 3.2

[22] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992. 5.4

[23] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 1, 2.4, 2.4, 3.1, 4, 5.1, 5.1

[24] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems: Lecture notes. Technical report, Ecole d'ete de Probabilités de Saint-Flour, 2008. A

[25] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010. 1, 1.1, 3.2, 4, 5.2

[26] S. Mahalanabis. A note on active learning for smooth problems. arXiv:*1103.3095*, 2011. 5.2, 5.2

[27] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:

1808–1829, 1999. 2.4, 5.1, 5.1

[28] S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13 (1):67–90, 2012. 1.1

[29] D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2): 780–799, 1987. 5.4

[30] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, Berlin / New York, 1984. 5.4

[31] D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics and American Statistical Association, 1990. 5.4

[32] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems 24*, 2011. 5.2, 5.2, 5.4, 5.4

[33] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 2.4, 5.1, 5.1

[34] A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. 2, 5.3, 5.3

[35] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 2.4, 5.3, 5.3, 5.4, 5.5, A

[36] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971. 5.4

[37] L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011. 5.2, 5.2

[38] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004. 1.1

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVENUE
PITTSBURGH, PA 15213 USA
E-MAIL: shanneke@stat.cmu.edu

MACHINE LEARNING DEPARTMENT
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVENUE
PITTSBURGH, PA 15213 USA
E-MAIL: liuy@cs.cmu.edu