Surrogate Losses in Passive and Active Learning

Steve Hanneke[*] **and Liu Yang**[†]

[*]*Toyota Technological Institute at Chicago*
e-mail: steve.hanneke@gmail.com

[†]*e-mail:* liu.yang0900@outlook.com

Abstract: 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.

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.

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 [e.g., 2, 30–32], 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, using modest computational resources. Although classifiers arrived at in this 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 committed to use a given surrogate loss, and we restrict our attention to just those scenarios in which this heuristic actually *does* work: specifically, where the minimizer of the surrogate risk also minimizes the error rate, and is contained in our function class. 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.

In passive learning, the most common approach to using a surrogate loss is to minimize the empirical surrogate risk on the labeled data. One can then derive guarantees on the error rate of this strategy by bounding the surrogate risk via concentration inequalities, and then converting these guarantees on the surrogate risk into guarantees on the error rate, a technique pioneered by Bartlett, Jordan, and McAuliffe [6] and Zhang [51]. Interestingly, we find that this direct approach is *not* appropriate in the context of active learning: that is, optimizing the surrogate risk to a sufficient extent to guarantee small error rate generally *cannot* yield large improvements over passive learning. While at first this finding might seem quite negative, it leaves open the possibility of methods making use of the surrogate loss in alternative ways, which still guarantee low error rate and computational efficiency, but for which these guarantees arise via a less direct route. Indeed, since we are interested in the surrogate loss only insofar as it helps us to optimize the error rate with computational efficiency, we may even consider methods that provide *no* guarantees on the achieved surrogate risk whatsoever (even in the limit).

In the present work, we propose such an alternative approach to the use of surrogate losses in active learning. The insight leading to this approach 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 location any further, and can therefore focus the label requests elsewhere. Based on this insight, we construct an active learning strategy that optimizes 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. In many cases, these bounds reflect strong improvements over the analogous results for passive learning by minimizing the given surrogate loss. 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 techniques for the study of active learning developed over the past decade [4, 23, 24, 36], 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 [34] and Bartlett, Bousquet, and Mendelson [8] on local 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 [51]. 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 it.

Another related line of research, explored by Audibert and Tsybakov [3], 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 minimax optimality results under smoothness assumptions on the actual regression function. Under similar conditions, Minsker [41] 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 [3]. Minsker's active learning work has also recently been strengthened and extended in [27, 38]. Remarkably, as discussed by Audibert and Tsybakov [3], the rates of convergence obtained in such works are often better than the known results for methods that directly optimize the 0-1 loss, under analogous complexity assumptions on the Bayes optimal classifier (rather than the regression function). As a result, these works raise interesting questions about whether the general analysis 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 might typically be strongest when the surrogate loss is the 0-1 loss itself, even if in some cases stronger results might in principle be achievable from assumptions involving the surrogate loss [as in 3, 41]. As such, in the specific scenarios studied by Minsker [41], 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 [51] than to those of Audibert and Tsybakov [3] and Minsker [41]. That said, we note that several important elements of the design and analysis of the active learning method below are already hinted at to some extent in the work of Minsker [41], albeit in a form that also relies heavily on the assumptions and function class specific to that work; the present work takes the general perspective, developing theory and methods applicable to any function class and surrogate loss function.

Our approach to the design of active learning methods below follows the well-studied strategy of *disagreement-based* active learning, an approach pioneered by Balcan, Beygelzimer, and Langford [4], and further developed by several later works [e.g., 14, 24, 25, 36]. The basic strategy maintains a set V of plausible candidates for the optimal classifier, and requests the labels of samples disagreed-upon by classifiers in V; it periodically updates the set V by eliminating classifiers making an excessive number of mistakes on the requested labels. The analysis of the number of label requests sufficient for this technique to achieve a given error rate in the general case was explored by Hanneke [22, 24], Dasgupta, Hsu, and Monteleoni [14], Koltchinskii [36], and others, and the results are typically expressed in terms of a quantity known as the *disagreement coefficient*. In the present work, we modify the disagreement-based active learning strategy by updating the set V, not based on the number of mistakes, but rather based on the empirical surrogate risk on the queried samples. We derive bounds on the number of label requests this method requires to achieve a given excess error rate, in terms of properties of the surrogate loss. In particular, when the surrogate loss is chosen to be the 0-1 loss itself, this method behaves nearly-identically to previously-studied methods [25, 36], and in this special case, our results match those established in the literature (with some small refinements in the logarithmic factors).

There are several interesting works on active learning methods that optimize a general loss function. Beygelzimer, Dasgupta, and Langford [9] and Koltchinskii [36] have both proposed such 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 for constructing and analyzing passive learning methods based on a surrogate loss. However, as we discuss below, this strategy does not generally lead to the best results for active learning, and often will not be much better than results available for related passive learning methods. Instead, the method we propose does not aim to optimize the surrogate risk overall, but rather optimizes it on a sequence of increasingly-focused subregions of the instance space, and thereby provides a smaller bound on the number of label requests sufficient to guarantee excess error rate ε.

2. Definitions

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space, where $\mathcal{X}$ is called the *instance 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., 48]; when this is not the case, one 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 $\mathrm{sign}(h(\cdot)) = \mathrm{sign}(\eta(\cdot; P) - 1/2)$. For any $\mathcal{H} \subseteq \mathcal{F}^*$, define the *region of sign-disagreement* $\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x))\}$. Additionally, 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 $\mathcal{H}$.

We will use standard big-O notation to express asymptotic dependences. Specifically, for $f, g : (0, \infty) \to [0, \infty)$, we write $f(\varepsilon) = O(g(\varepsilon))$ or $g(\varepsilon) = \Omega(f(\varepsilon))$ if $\limsup_{\varepsilon \to 0} f(\varepsilon)/g(\varepsilon) < \infty$; we write $f(\varepsilon) = \Theta(g(\varepsilon))$ if both $f(\varepsilon) = O(g(\varepsilon))$ and $f(\varepsilon) = \Omega(g(\varepsilon))$, and we write $f(\varepsilon) = o(g(\varepsilon))$ if $\limsup_{\varepsilon \to 0} f(\varepsilon)/g(\varepsilon) = 0$.

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 a given budget n, the algorithm halts and returns a function $\hat{h} \in \mathcal{F}^*$. Formally, this protocol specifies a type of decision rule mapping the random sequence $\mathcal{Z}$ 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*. For simplicity, suppose $|z| < \infty \Rightarrow \ell(z) < \infty$. Define $\bar{\ell} = 1 \vee \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \sup_{h \in \mathcal{F}} \ell(yh(x))$. We will generally suppose $\bar{\ell} < \infty$. In practice, this is more often a constraint on $\mathcal{F}$ and $\mathcal{X}$ 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. 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 $(X, Y) \sim P$. This is the *ℓ-risk* of g under P. When $P = \mathcal{P}_{XY}$, abbreviate this as $\mathrm{R}_\ell(g) = \mathrm{R}_\ell(g; \mathcal{P}_{XY})$.

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 $\ell^\star_-(\eta_0) = \inf_{z\in\bar{\mathbb{R}}:z(2\eta_0-1)\leq 0}(\eta_0\ell(z) + (1-\eta_0)\ell(-z))$. Then the surrogate loss ℓ is said to be *classification-calibrated* if, $\forall\eta_0 \in [0,1]\setminus\{1/2\}$, $\ell^\star_-(\eta_0) > \ell^\star(\eta_0)$. 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, for the remainder of this article 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 infimum 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 $\mathrm{R}_\ell(f^\star_P;P) = \inf_{g\in\mathcal{F}^*}\mathrm{R}_\ell(g;P)$. Furthermore, since ℓ is classification-calibrated, we have $\mathrm{sign}(f^\star_P(x)) = \mathrm{sign}(\eta(x;P)-1/2)$ 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. When $P = \mathcal{P}_{XY}$, we abbreviate by $f^\star = f^\star_{\mathcal{P}_{XY}}$.

All of our main results below rely on the assumption that $f^\star \in \mathcal{F}$. When combined with the fact that ℓ is classification-calibrated, this essentially stands as a formal representation of the informal assumption that the surrogate loss ℓ was chosen wisely: that is, that functions in $\mathcal{F}$ with relatively low surrogate risk necessarily have relatively low error rate. However, it should be noted that this is often a very strong assumption, significantly restricting the allowed distributions $\mathcal{P}_{XY}$. For instance, for many losses ℓ in practical use (e.g., the quadratic loss), when $\mathcal{F}$ is a parametric family, the assumption that $f^\star \in \mathcal{F}$ essentially restricts the allowed functions $\eta(\cdot)$ to also form a parametric family. This fact underscores the need for great care in selecting a surrogate loss when approaching a given learning problem in practice. In principle, one can relax this assumption slightly, at the expense of significantly more-complicated theorem statements, and we include some superficial remarks on this in Appendix F. However, it seems any truly-substantial relaxation would require a significantly different approach.

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 $\mathcal{H} \subseteq \mathcal{F}^*$ 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 define analogous quantities for the

0-1 loss. Define the *distance* $\Delta_P(h,g) = P((x,y) : \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x)))$ and *radius* $\mathrm{radius}(\mathcal{H};P) = \sup_{h \in \mathcal{H}} \Delta_P(h, f_P^\star)$. Also define the ε-minimal set of $\mathcal{H}$ as $\mathcal{H}(\varepsilon;{}_{01},P) = \{h \in \mathcal{H} : \mathrm{er}(h;P) - \inf_{g \in \mathcal{H}} \mathrm{er}(g;P) \leq \varepsilon\}$, 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 $\Delta(h,g) = \Delta_{\mathcal{P}_{XY}}(h,g)$, $\mathrm{radius}(\mathcal{H}) = \mathrm{radius}(\mathcal{H};\mathcal{P}_{XY})$, $\mathcal{H}(\varepsilon;{}_{01}) = \mathcal{H}(\varepsilon;{}_{01},\mathcal{P}_{XY})$, and $\mathrm{B}_{\mathcal{H}}(h,r) = \mathrm{B}_{\mathcal{H},\mathcal{P}_{XY}}(h,r)$; when $\mathcal{H} = \mathcal{F}$, further abbreviate $\mathrm{B}(h,r) = \mathrm{B}_{\mathcal{F}}(h,r)$.

The following definition will enable us to transform guarantees on the excess surrogate risk into guarantees on the excess error rate.

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

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

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

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

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, Γ_ℓ has the property that

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

In fact, Γ_ℓ is defined to be maximal with this property, in that *any* Γ_ℓ' for which (1) is satisfied must have $\Gamma_\ell'(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ for all $\varepsilon \in [0,1]$. For this reason, we will be interested in calculating lower bounds on Γ_ℓ. Bartlett, Jordan, and McAuliffe [6] studied various ways to obtain concrete, calculable 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]; for convenience, also define $\psi_\ell(x)$ for $x \in (1,\infty)$ arbitrarily, subject to maintaining convexity of ψ_ℓ. 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 $\psi_\ell(\mathrm{er}(h) - \mathrm{er}(f^\star)) \leq \mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star)$, so that $\psi_\ell \leq \Gamma_\ell$, 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 argument in the proof of their Theorem 3, one can show that $\forall h \in \mathcal{F}^*$,

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

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

$$\Psi_\ell(\varepsilon) \leq \mathrm{radius}(\mathcal{F}^*(\varepsilon;{}_{01}))\psi_\ell\left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}^*(\varepsilon;{}_{01}))}\right). \tag{2}$$

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 $(0,1)$, this implies $\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h)-\mathrm{er}(f^\star);{}_{01}))\psi_\ell\left(\frac{\mathrm{er}(h)-\mathrm{er}(f^\star)}{2\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h)-\mathrm{er}(f^\star);{}_{01}))}\right) < \Psi_\ell(\varepsilon)$; 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 $\Psi_\ell(\varepsilon) \le \Gamma_\ell(\varepsilon)$, so that $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(f^\star) < \Psi_\ell(\varepsilon) \implies \mathrm{er}(h) - \mathrm{er}(f^\star) < \varepsilon$. 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 $\Gamma_\ell(\varepsilon)$ also hold while replacing $\Gamma_\ell(\varepsilon)$ with any nondecreasing Ψ'_ℓ s.t. $\forall \varepsilon \ge 0$, $\Psi'_\ell(\varepsilon) \le \mathrm{radius}(\mathcal{F}(\varepsilon;{}_{01}))\psi_\ell\left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}(\varepsilon;01))}\right)$, 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 with convexity quantified by a polynomial, as described in the following condition.

Condition 2. *$\mathcal{F}$ is convex, with $\forall x \in \mathcal{X}, \sup_{f\in\mathcal{F}} |f(x)| \le \bar{B}$ 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 $\forall x,y \in [-\bar{B},\bar{B}], |\ell(x)-\ell(y)| \le L d_\ell(x,y)$, and the function*

$$\bar{\delta}_\ell(\varepsilon) = \inf\left(\left\{\frac{1}{2}\ell(x)+\frac{1}{2}\ell(y) - \ell\left(\frac{1}{2}x+\frac{1}{2}y\right) : x,y\in[-\bar{B},\bar{B}], d_\ell(x,y) \ge \varepsilon\right\}\cup\{\infty\}\right)$$

satisfies $\forall \varepsilon \in [0,\infty)$, $\bar{\delta}_\ell(\varepsilon) \ge C_\ell \varepsilon^{r_\ell}$.

In particular, note that if $\mathcal{F}$ is convex, and the functions in $\mathcal{F}$ are uniformly bounded, and ℓ is convex and continuous, then Condition 2 is always satisfied (though possibly with $r_\ell = \infty$) by taking $d_\ell(x,y) = |x-y|/(4\bar{B})$.

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 *quadratic loss* (or squared loss), specified as $\ell(x) = (1-x)^2$, is often used in so-called *plug-in* classifiers [3], 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 $\mathcal{X}\times\mathcal{Y}$, $f^\star_P(\cdot) = 2\eta(\cdot;P)-1$, so that it is straightforward to describe the set of distributions P satisfying the assumption $f^\star_P \in \mathcal{F}$. In classification, this loss is sometimes modified as $\ell(x) = \max\{1-x,0\}^2$, called the *truncated* quadratic loss. Bartlett, Jordan, and McAuliffe [6] show that for the quadratic loss (with or without truncation), $\psi_\ell(x) = x^2$, and Condition 2 is satisfied with $L = 2(\bar{B}+1)$, $C_\ell = 1/4$, $r_\ell = 2$.

Example 2 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

[18]. Bartlett, Jordan, and McAuliffe [6] show that under the exponential loss, $f^\star(x) = \frac{1}{2}\ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$ and $\psi_\ell(x) = 1 - \sqrt{1-x^2}$, which is tightly approximated by $x^2/2$ for small x. They also show this loss satisfies the conditions on ℓ in Condition 2 with $d_\ell(x,y) = |x-y|$, $L = e^{\bar{B}}$, $C_\ell = e^{-\bar{B}}/8$, and $r_\ell = 2$. Note, however, that for noise-free distributions, we would need $f^\star(x) = \pm\infty$, which means most common function classes $\mathcal{F}$ could not be expected to contain $f^\star$ for this loss in the noise-free case.

Example 3 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) [13]. Bartlett, Jordan, and McAuliffe [6] show that for the hinge loss, $f^\star(x) = \operatorname{sign}(\eta(x) - 1/2)$ and $\psi_\ell(x) = |x|$. The hinge loss is Lipschitz continuous, with Lipschitz constant 1. However, for the remaining conditions on ℓ in Condition 2, 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.

3. Methods Based on Optimizing the Surrogate Risk

Perhaps the simplest way to use a surrogate loss is 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 introduce a classic passive learning method based on this strategy, and discuss 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 $h \in \mathcal{F}$. To define this method, we first introduce some notation. For any $m \in \mathbb{N}$, $g : \mathcal{X} \to \bar{\mathbb{R}}$, and $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, we overload the $\mathrm{R}_\ell(g; \cdot)$ notation, defining the *empirical ℓ-risk* as $\mathrm{R}_\ell(g; S) = m^{-1}\sum_{i=1}^m \ell(g(x_i)y_i)$: that is, $\mathrm{R}_\ell(g; S)$ is the ℓ-risk of g under the uniform distribution on S. At times it will be convenient to keep track of the indices for a subsequence of $\mathcal{Z}$, and for this reason we further 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 $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_\ell(h; \mathcal{Z}_m)$. 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 passive learning method for comparison. We review several known performance guarantees for ERM_ℓ below.

3.2. Negative Results for Active Learning

As mentioned, there are several active learning methods designed to optimize a general loss function [9, 36]. 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 $\mathcal{X} = \{x_0, x_1\}$, a fixed $\bar{B} \in (0, \infty)$, and $\mathcal{F}$ as the set of all functions f with $(f(x_0), f(x_1)) \in [-\bar{B}, \bar{B}] \times (0, \bar{B}]$. Let $z \in (0, 1/2)$ be a constant, let $\eta(x_1) = 1/2 + z$, and suppose that ℓ is a classification-calibrated loss with $\bar{\ell} < \infty$ such that for any $\eta(x_0) \in [4/6, 5/6]$, we have $f^\star \in \mathcal{F}$ (the latter condition could equivalently be stated as a constraint on $\bar{B}$). Given a small value $\varepsilon \in (0, z)$, let $\mathcal{P}(\{x_1\}) = \varepsilon/(2z)$, $\mathcal{P}(\{x_0\}) = 1 - \mathcal{P}(\{x_1\})$. For this problem, any function h with $\mathrm{sign}(h(x_1)) = -1$ has $\mathrm{er}(h) - \mathrm{er}(f^\star) \geq \varepsilon$, so that $\Gamma_\ell(\varepsilon) \leq (\varepsilon/(2z))(\ell^\star_-(\eta(x_1)) - \ell^\star(\eta(x_1)))$; since ℓ is classification-calibrated and $\bar{\ell} < \infty$, this implies $\Gamma_\ell(\varepsilon) \leq c\varepsilon$, 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 $\mathbb{E}[\ell(h(X)Y)|X = x_0] - \mathbb{E}[\ell(f^\star(X)Y)|X = x_0] \leq c\varepsilon/\mathcal{P}(\{x_0\}) = O(\varepsilon)$. Existing results of Hanneke and Yang [28] (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 latter guarantee is $\Theta(1/\varepsilon)$. Hanneke and Yang [28] 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 2 with $r_\ell = 2$), $\Theta(1/\varepsilon)$ random labeled samples are sufficient for ERM_ℓ to achieve this same guarantee, so that error bounds based purely on 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.

Below, we provide an active learning algorithm and analysis of its performance which, in the scenario above (with $r_\ell = 2$), guarantees expected excess error rate less than ε, 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 might ask whether there is a method that makes more effective use of ℓ for optimizing the error rate, while maintaining the computational advantages. To explore this question, we propose the following method, which generalizes the methods of Koltchinskii [36] and Hanneke [25]. Results similar to those proven below should also hold for analogous generalizations of the related methods of [4, 9, 14].

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

---

0. $V \leftarrow \mathcal{F}$, $Q \leftarrow \{\}$, $m \leftarrow 1$, $t \leftarrow 0$
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 $\log_2(m) \in \mathbb{N}$
6. $\quad\quad V \leftarrow \left\{h \in V : \mathrm{R}_\ell(h; Q) - \inf_{g \in V} \mathrm{R}_\ell(g; Q) \leq \hat{T}_\ell(V; Q, m)\right\}$
7. $\quad\quad Q \leftarrow \{\}$
8. 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 periodically (Step 6), removing 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 shrinking the sampling region $\mathrm{DIS}(V)$. The actual definition of $\hat{T}_\ell$ sufficient for the results stated below will be specified in Section 6.3 below, based on data-dependent concentration inequalities.

In practice, the set V can be maintained implicitly, simply by keeping track of the constraints (Step 6) 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. Thus, for convex loss functions and convex finite-dimensional classes of function, these steps typically have computationally efficient realizations as convex optimization problems, as long as the $\hat{T}_\ell$ values can also be obtained efficiently.

We include general results on the performance of Algorithm 1 in Section 6 below. For now, we briefly sketch the main ideas of the analysis, in rough outline. For any measurable $\mathcal{U} \subseteq \mathcal{X}$, and 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 $\mathcal{H}_{\mathcal{U},g} = \{h_{\mathcal{U},g} : h \in \mathcal{H}\}$. In the special case $g = f^\star$, we abbreviate these as $h_{\mathcal{U}} = h_{\mathcal{U},f^\star}$ and $\mathcal{H}_{\mathcal{U}} = \{h_{\mathcal{U}} : h \in \mathcal{H}\}$. As mentioned, the idea in the analysis is to argue that Algorithm 1 maintains $f^\star \in V$, while also removing from V any function with relatively large error rate, within a certain number of rounds. More explicitly, upon reaching m satisfying the condition in Step 5, if we denote $\mathcal{L}_m = \{(1 + m/2, Y_{1+m/2}), \ldots, (m, Y_m)\}$, then since every $(m', Y_{m'}) \in \mathcal{L}_m$ is either in Q or else $X_{m'} \notin \mathrm{DIS}(V)$, every $h \in V$ has $(\mathrm{R}_\ell(h; Q) - \inf_{g \in V} \mathrm{R}_\ell(g; Q))|Q| = (\mathrm{R}_\ell(h_{\mathrm{DIS}(V)}; \mathcal{L}_m) - \inf_{g \in V} \mathrm{R}_\ell(g_{\mathrm{DIS}(V)}; \mathcal{L}_m))\frac{m}{2}$.

We therefore define $\hat{T}_\ell(V; Q, m)$ to provide a concentration inequality $\mathrm{R}_\ell(f^\star; \mathcal{L}_m) - \inf_{g \in V} \mathrm{R}_\ell(g_{\mathrm{DIS}(V)}; \mathcal{L}_m) \leq \frac{2|Q|}{m} \hat{T}_\ell(V; Q, m)$, thus maintaining that $f^\star \in V$ in Step 6. This also implies that, if $V_{\mathrm{DIS}(V)} \subseteq [\mathcal{F}](2^{2-j}; \ell)$ upon reaching Step 5 (for some $j \in \mathbb{Z}$), then $V \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01})$. One can then show that, upon reaching m of a certain size u_j (quantified below), the value $\frac{2|Q|}{m} \hat{T}_\ell(V; Q, m)$ will be small enough that, in combination with concentration of $\mathrm{R}_\ell(h_{\mathrm{DIS}(V)}; \mathcal{L}_m)$ values, after the update in Step 6, only functions $h \in V$ with $\mathrm{R}_\ell(h_{\mathrm{DIS}(V)}) - \mathrm{R}_\ell(f^\star) < 2^{-j}$ will remain: that is, after the update, $V_{\mathrm{DIS}(V)} \subseteq [\mathcal{F}](2^{-j}; \ell)$. By induction, upon reaching m of a sufficiently large size u_{j_ε} (quantified below), every $h \in V$ has $\mathrm{R}_\ell(h_{\mathrm{DIS}(V)}) - \mathrm{R}_\ell(f^\star) < \Gamma_\ell(\varepsilon)$, which implies $\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \varepsilon$. This provides a sufficient size of u to obtain excess error rate ε. Next, we note that the algorithm requests a label Y_m only if $X_m \in \mathrm{DIS}(V)$. The above reveals that, if $u_{j-1} < m \leq u_j$, then $V \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01})$, which implies $\mathrm{DIS}(V) \subseteq \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01}))$. Thus, the number of labels the algorithm requests among indices m with $u_{j-1} < m \leq u_j$ is at most the number with $X_m \in \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01}))$, a number which can easily by upper bounded by a simple Chernoff bound. This provides a sufficient size of n for the algorithm to obtain excess error rate ε.

The number of label requests sufficient for Algorithm 1 to obtain excess error rate ε can often (though not always) be significantly smaller than the number of random labeled data points sufficient for ERM_ℓ to achieve the same. This is typically the case when $\mathcal{P}(\mathrm{DIS}(\mathcal{F}(\varepsilon; {}_{01}))) \to 0$ as $\varepsilon \to 0$. When this is the case, the number of labels requested by the algorithm is sublinear in the number of unlabeled samples it processes. Not surprisingly, the magnitude of the improvements of Algorithm 1 over ERM_ℓ can be quantified in terms of the *rate* at which $\mathcal{P}(\mathrm{DIS}(\mathcal{F}(\varepsilon; {}_{01})))$ vanishes as $\varepsilon \to 0$. In the next section, we quantify this rate in terms of a complexity measure known as the disagreement coefficient.

5. Main Results

We provide a general analysis of Algorithm 1 in Section 6.4 below. For now, we summarize a few of the most interesting implications of that analysis, under commonly-studied complexity conditions: namely, VC subgraph classes and entropy conditions. Detailed derivations for all of these results (from the abstract theorems) are included in Section 7 below. Appendix C further includes a brief discussion 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. Throughout, we use the convenient notation $\mathrm{Log}(x) = \max\{\ln(x), 1\}$, defined for all $x \in (0, \infty)$.

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 [40] and Tsybakov [45], is a well-known noise condition, about which there is now an extensive literature [e.g., 6, 24, 25, 34].

Condition 3. *For some $a \in [1,\infty)$ and $\alpha \in [0,1]$, for every $g \in \mathcal{F}^*$,*

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

Condition 3 is equivalently expressed in terms of certain noise conditions [6, 40, 45]. Specifically, satisfying Condition 3 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 3 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 3. 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 replacing (2) with the $\mathcal{F}$-specific Ψ'_ℓ requirement mentioned in Section 2.1, as remarked above.

An analogous condition can be defined for the surrogate loss function, as follows. Essentially-similar notions have been explored by Bartlett, Jordan, and McAuliffe [6] and Koltchinskii [34].

Condition 4. *For some $b \in [1,\infty)$ and $\beta \in [0,1]$, for every $g \in [\mathcal{F}]$,*

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

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 4 can even be satisfied *universally*, in the sense that it holds for a particular value of $\beta > 0$ for *all* distributions. In particular, Bartlett, Jordan, and McAuliffe [6] show that this is the case under Condition 2, as stated in the following lemma (see [6] for the proof).

Lemma 5. *Suppose Condition 2 is satisfied. Let $b = (2C_\ell \bar{d}_\ell^{\min\{r_\ell-2,0\}})^{-\beta}L^2$ and $\beta = \min\{1, \frac{2}{r_\ell}\}$. Then* every *distribution P over $\mathcal{X}\times\mathcal{Y}$ with $f_P^\star \in [\mathcal{F}]$ satisfies Condition 4 with these values of* b *and β.*

Under Condition 3, 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

$$\Psi_\ell(\varepsilon) = a\varepsilon^{\alpha}\psi_\ell\left(\varepsilon^{1-\alpha}/(2a)\right) \tag{3}$$

satisfies $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ [6]. Furthermore, for classification-calibrated ℓ, Ψ_ℓ in (3) is strictly increasing, nonnegative, and continuous on $(0,1)$ [6], and has $\Psi_\ell(0) = 0$; thus, the inverse, defined for $\gamma > 0$ by $\Psi_\ell^{-1}(\gamma) = \inf(\{\varepsilon > 0 : \gamma \leq \Psi_\ell(\varepsilon)\}\cup\{1\})$, 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 $\forall \gamma > 0, \mathcal{E}_\ell(\gamma) \le \Psi_\ell^{-1}(\gamma)$.

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

$$\begin{aligned} \mathcal{G}_{\mathcal{H}} &= \{(x,y) \mapsto \ell(h(x)y) : h \in \mathcal{H}\}, \\ \text{and } \mathcal{G}_{\mathcal{H},P} &= \{(x,y) \mapsto \ell(h(x)y) - \ell(f_P^\star(x)y) : h \in \mathcal{H}\}. \end{aligned} \tag{4}$$

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.

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, 9, 14, 19, 22–25, 36, 44, 50], and the coefficient in the linear function is typically referred to as the *disagreement coefficient*. Specifically, the following definition is due to Hanneke [22, 24]; related quantities have been explored by Alexander [1] and Giné and Koltchinskii [20].

Definition 6. *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 $\theta(r_0) = \theta_{f^\star}(r_0)$.

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, 19, 22, 24, 39], while in other cases, $\theta(\varepsilon)$ may have an interesting dependence on ε [5, 44, 50]. The reader is referred to the works of Hanneke [24, 25] for detailed discussions on the disagreement coefficient.

5.3. VC Subgraph Classes

We begin with results for VC subgraph classes. For a collection $\mathcal{A}$ of sets, a set of points $\{z_1, \ldots, z_k\}$ 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 $\mathcal{A}$ [49]; 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 [29, 43]); to simplify the results below, we adopt the convention that when the VC dimension of this collection is 0, we let $\mathrm{vc}(\mathcal{G}) = 1$. $\mathcal{G}$ is said to be a *VC subgraph* class if $\mathrm{vc}(\mathcal{G}) < \infty$ [47].

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 $\mathrm{vc}(\mathcal{G}_\mathcal{H})$. In some special cases, such as monotonic ℓ, these results can be rephrased directly in terms of $\mathrm{vc}(\mathcal{H})$ if desired [e.g., 17, 29].

Following Giné and Koltchinskii [20], 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 $\tau_\ell(r_0) = \tau_\ell(r_0; \mathcal{F}, \mathcal{P}_{XY})$.

We can now state the following theorem, providing a sample size sufficient for ERM_ℓ to obtain excess error rate ε. This result is implicit in the work of Giné and Koltchinskii [20].

Theorem 7. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (3), 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*

$$m \geq c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\tau_\ell\right) + \mathrm{Log}\left(1/\delta\right)\right), \tag{5}$$

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

As noted by Giné and Koltchinskii [20], in the special case when ℓ is itself the 0-1 loss ($\ell = \mathbb{1}_{[-\infty,0]}$) and $\mathcal{F}$ is a set of $\{-1,+1\}$-valued classifiers, (5) simplifies quite nicely, since then $\|\mathrm{F}(\mathcal{G}_{\mathrm{B}_{\mathcal{F},\mathcal{P}_{XY}}(f^\star, r;\ell), \mathcal{P}_{XY}})\|_{\mathcal{P}_{XY}}^2 = \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star, r\right)\right)\right)$, so that $\tau_\ell(r_0) = \theta(r_0)$; in this case, we also have $\mathrm{vc}(\mathcal{G}_\mathcal{F}) = \mathrm{vc}(\mathcal{F})$ and $\Psi_\ell(\varepsilon) = \varepsilon/2$, and we can take $\beta = \alpha$ and $b = a$, so that it suffices to have

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

where $\theta = \theta\left(a\varepsilon^\alpha\right)$ and $c \in [1, \infty)$ is a universal constant. This is sometimes proportional to the minimax number of samples for passive learning [11, 24, 44].

Next, we turn to the analysis of Algorithm 1 under these same conditions. Suppose $\mathcal{P}_{XY}$ satisfies Conditions 3 and 4, and for $\gamma_0 \geq 0$, define

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

We claim the following theorem, bounding the number of samples (labeled and unlabeled) sufficient for Algorithm 1 to obtain excess error rate ε, under the same conditions as Theorem 7. As mentioned above, the specific definition of $\hat{T}_\ell$ sufficient for this theorem will be formally specified in Section 6.3. Also, the specification of $\hat{\mathfrak{s}}$ will be given in the proof, in Appendix B.

Theorem 8. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (3), for any $\varepsilon \in (0,1)$, letting $\theta = \theta(a\varepsilon^\alpha)$, $\chi_\ell = \chi_\ell(\Psi_\ell(\varepsilon))$, $A_1 = \mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell \bar{\ell}) + \mathrm{Log}(1/\delta)$, $C_1 = \min\left\{\frac{1}{1-2^{(\alpha-1)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, and $B_1 = \min\left\{C_1, \frac{1}{1-2^{(\beta-1)}}\right\}$, if $u, n \in \mathbb{N}$ satisfy*

$$u \geq c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right) A_1, \tag{6}$$

$$n \geq c\theta a\varepsilon^\alpha \left(\frac{b(A_1 + \mathrm{Log}(B_1))B_1}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1 + \mathrm{Log}(C_1))C_1}{\Psi_\ell(\varepsilon)}\right), \tag{7}$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, 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 adopt the convention that $1/0 = \infty$ so that B_1 and C_1 are well-defined even when $\alpha = 1$ or $\beta = 1$. When $\alpha < 1$, the dependence on ε in (7) is $O\left(\theta\varepsilon^\alpha \Psi_\ell(\varepsilon)^{\beta-2}\mathrm{Log}(\chi_\ell)\right)$, while in the case $\alpha = \beta = 1$, it is $O\left(\theta \mathrm{Log}(1/\varepsilon)(\mathrm{Log}(\theta) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)))\right)$. Comparing Theorem 8 to Theorem 7, the conditions on u in (6) and m in (5) are almost identical, aside from a logarithmic factor, so that the total number of data points indicated is roughly the same. However, the number of *labels* indicated by (7) may often be significantly smaller than the condition in (5), multiplying it by roughly $\theta a\varepsilon^\alpha$. This reduction is particularly strong when θ is bounded by a finite constant and α is large. Moreover, this is the same *type* of improvement known to occur when ℓ is itself the 0-1 loss [24]; in particular, in this special case, (7) is sometimes nearly minimax [24, 44]. Regarding the slight difference between (6) and (5) from replacing τ_ℓ by $\chi_\ell\bar{\ell}$, the effect is somewhat mixed, and which of these is smaller may depend on $\mathcal{F}$ and ℓ. For ℓ the 0-1 loss, $\tau_\ell = \chi_\ell\bar{\ell} = \theta(a(\varepsilon/2)^\alpha)$.

In the case when ℓ satisfies Condition 2, we can derive the following sometimes-stronger result with the help of Lemma 5.

Theorem 9. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3, ℓ is classification-calibrated and satisfies Condition 2, $f^\star \in \mathcal{F}$, Ψ_ℓ is as in (3), and b and β are as in Lemma 5, then for any $\varepsilon \in (0,1)$, letting $\theta = \theta(a\varepsilon^\alpha)$ and $A_2 = \mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\left(\bar{\ell}^2/b\right)\left(a\theta\varepsilon^\alpha/\Psi_\ell(\varepsilon)\right)^\beta\right) + \mathrm{Log}\left(1/\delta\right)$, and letting C_1 be as in Theorem 8, if $u, n \in \mathbb{N}$ satisfy*

$$u \geq c\left(\frac{b\left(a\theta\varepsilon^\alpha\right)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right) A_2, \tag{8}$$

$$n \geq c\left(b\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right)(A_2 + \mathrm{Log}(C_1))C_1, \tag{9}$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, 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 constraint on u in (8) has $O\left(\frac{(\theta\varepsilon^{\alpha})^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}}\mathrm{Log}\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right)$ dependence on ε, while the constraint on n in (9) has $O\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{2-\beta}\mathrm{Log}\left(\left(\frac{\theta\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right)$ in the case $\alpha < 1$, or $O\left(\theta^{2-\beta}\mathrm{Log}(1/\varepsilon)\mathrm{Log}\left(\theta^{\beta}\mathrm{Log}(1/\varepsilon)\right)\right)$ in the case $\alpha = 1$. This is noteworthy when θ is small while $\alpha > 0$ and $r_\ell > 2$, for at least two reasons. First, the sufficient size of n in (9) is smaller than that in Theorem 8, multiplying by roughly $(a\theta\varepsilon^{\alpha})^{1-\beta}$. Second, even the sufficient number of *unlabeled* samples in (8) may be smaller than the sufficient number of *labeled* samples for ERM_ℓ from Theorem 7, again multiplying by roughly $(a\theta\varepsilon^{\alpha})^{1-\beta}$. Thus, in the case ℓ satisfies Condition 2 with $r_\ell > 2$, when Theorem 7 is tight, even with 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 (9) 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 losses ℓ of this type for which Theorem 7 is competitive with results for methods directly based on the 0-1 loss. Thus, these improvements in u and n in Theorem 9 may simply indicate that Algorithm 1 is, to some extent, *compensating* for a choice of ℓ that would otherwise lead to suboptimal error rates.

5.4. Entropy Conditions

In this section, we consider characterizations of the complexity of $\mathcal{F}$ in terms of *entropy conditions*. As with the above results, detailed derivations of all of these results are presented in Section 7.3 below, based on the abstract theorems presented in Section 6.4.

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: $(f,g) \mapsto \|f-g\|_P$. Also, 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 $\mathcal{G}$.

The following represent two commonly-studied conditions.

Condition 10. *For some $q \geq 1$, $\rho \in (0,1)$, $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F},\mathcal{P}_{XY}})$, either $\forall \varepsilon > 0$,*

$$\ln \mathcal{N}_{[]}(\varepsilon\|\mathrm{F}\|_{\mathcal{P}_{XY}}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})) \leq q\varepsilon^{-2\rho}, \tag{10}$$

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

$$\ln \mathcal{N}(\varepsilon\|\mathrm{F}\|_{P}, \mathcal{G}_{\mathcal{F}}, L_2(P)) \leq q\varepsilon^{-2\rho}. \tag{11}$$

The following theorem is a classic result on the performance of ERM_ℓ under the above conditions [e.g., 6, 47].

Theorem 11. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, $\mathcal{F}$ and $\mathcal{P}_{XY}$ satisfy Condition 10, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$,* *and Ψ_ℓ is as in (3), 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 $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

Turning to the analogous setting for active learning, we are able to establish the following theorem on the performance of Algorithm 1 under these same conditions.

Theorem 12. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, $\mathcal{F}$ and $\mathcal{P}_{XY}$ satisfy Condition 10, ℓ is classification-calibrated, $f^\star \in \mathcal{F}$, and Ψ_ℓ is as in (3), then for any $\varepsilon \in (0,1)$, letting B_1 and C_1 be as in Theorem 8, $B_2 = \min\left\{B_1, \frac{1}{1-2^{-\rho}}\right\}$, $C_2 = \min\left\{C_1, \frac{1}{1-2^{-\rho}}\right\}$, and abbreviating $\theta = \theta(a\varepsilon^\alpha)$, if $u, n \in \mathbb{N}$ satisfy*

$$u \geq c\frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2}\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}\left(\frac{1}{\delta}\right), \quad (12)$$

$$n \geq c\theta a\varepsilon^\alpha\frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2}\left(\frac{b^{1-\rho}B_2}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}C_2}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\theta a\varepsilon^\alpha\left(\frac{bB_1\mathrm{Log}(B_1/\delta)}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}C_1\mathrm{Log}(C_1/\delta)}{\Psi_\ell(\varepsilon)}\right), \quad (13)$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, 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 constraint on u in (12) is identical (up to constant factors) to the sample size in Theorem 11 sufficient for ERM_ℓ to achieve the same. In contrast, when θ is small, the constraint on n in (13) improves this, multiplying by a factor $\propto \theta a\varepsilon^\alpha$.

As before, when ℓ satisfies Condition 2, we can derive sometimes-stronger results via Lemma 5. In this case, we will distinguish between the cases of (11) and (10), as we find a slightly stronger result for the former. We begin with the following result, under the uniform entropy condition (11).

Theorem 13. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 3, ℓ is classification-calibrated and satisfies Condition 2, $f^\star \in \mathcal{F}$, Ψ_ℓ is as in (3), b and β are as in Lemma 5, and (11) is satisfied with $\mathrm{F} \leq \bar{\ell}$ ($\forall$ finitely discrete P, $\forall \varepsilon > 0$), then*

$\forall \varepsilon \in (0,1)$, *for* C_1 *as in Theorem 8 and* $\theta = \theta(a\varepsilon^\alpha)$, *if*

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

$$n \geq c\left(\frac{q\bar{\ell}^{2\rho}C_1}{(1-\rho)^2}\right)\left(b^{1-\rho}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+\rho}\right) + c\left(b\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right)C_1\mathrm{Log}\left(\frac{C_1}{\delta}\right),$$

then, with arguments ℓ, u, *and* n, *and an appropriate* $\hat{\mathfrak{s}}$ *function, 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* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

Compared to Theorem 12, the constraints for u and n here may have improved dependences on ε, multiplying by $O\left((\theta\varepsilon^\alpha)^{1-\beta(1-\rho)}\right)$. Furthermore, for small θ, these are also smaller than the size of m for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ from Theorem 11.

Next, we turn to the bracketing entropy condition (10). For simplicity, we will only consider the case that (10) is satisfied with $\mathrm{F} = \bar{\ell}$ constant. In this case, we have the following result.

Theorem 14. *For a universal constant* $c \in [1,\infty)$, *if* $\mathcal{P}_{XY}$ *satisfies Condition 3,* ℓ *is classification-calibrated and satisfies Condition 2,* $f^\star \in \mathcal{F}$, Ψ_ℓ *is as in (3),* b *and* β *are as in Lemma 5, and (10) is satisfied with* $\mathrm{F} = \bar{\ell}$, *then* $\forall \varepsilon \in (0,1)$, *letting* C_1 *be as in Theorem 8,* C_2 *be as in Theorem 12, and* $\theta = \theta(a\varepsilon^\alpha)$, *if*

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

$$n \geq c\left(\frac{q\bar{\ell}^{2\rho}C_2}{(1-\rho)^2}\right)\left(\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{\rho}}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+(1-\beta)(1-\rho)} + \frac{\bar{\ell}^{1-\rho}a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)^{1+\rho}}\right) + c\left(b\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right)C_1\mathrm{Log}\left(\frac{C_1}{\delta}\right),$$

then, with arguments ℓ, u, *and* n, *and an appropriate* $\hat{\mathfrak{s}}$ *function, 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* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

Compared to Theorem 12, the dependence on ε in the sizes for both u and n may be smaller here, multiplying by $O\Big((\theta\varepsilon^{\alpha})^{(1-\beta)(1-\rho)}\Big)$, which is sometimes significant, though not quite as dramatic a reduction as we found under (11) in Theorem 13. As with Theorem 13, when $\theta(\varepsilon^{\alpha}) = o(\varepsilon^{-\alpha})$, the sizes of u and n indicated by Theorem 14 are smaller than the results for $\mathrm{ERM}_{\ell}(\mathcal{F}, \mathcal{Z}_m)$ from Theorem 11.

5.5. An Example: Discrete Distributions

As a concrete example applying the above results, we find that Algorithm 1 generally provides some benefits for discrete $\mathcal{P}$ distributions. To describe these benefits quantitatively, consider the special case where $\exists x_1, x_2, \ldots \in \mathcal{X}$ with $\mathcal{P}(\{x_i\}) = \frac{90}{\pi^4 i^4}$, and $\eta(x) \in [0, \nu_0] \cup [1-\nu_0, 1]$ for each $x \in \mathcal{X}$, where $\nu_0 \in [0, 1/2)$ is a constant. Set $\mathcal{F} = \{f \in \mathcal{F}^* : \sup_{x\in\mathcal{X}} |f(x)| \leq 1\}$, and take ℓ to be the quadratic loss (in which case $\bar{\ell} = 4$). In particular, since $f^{\star}(x) = 2\eta(x) - 1 \in [-1, 1]$, the condition $f^{\star} \in \mathcal{F}$ is satisfied in this scenario. We will use Theorem 12 to bound the number of labels sufficient for Algorithm 1 to achieve excess error rate ε. For any $g \in \mathcal{F}^*$, we have $\mathrm{er}(g) - \mathrm{er}(f^{\star}) = \sum_{i\in\mathbb{N}} \mathbb{1}_{\mathrm{DIS}(\{g,f^{\star}\})}(x_i)|1-2\eta(x_i)|\mathcal{P}(\{x_i\}) \geq (1-2\nu_0)\Delta(g, f^{\star})$, so that Condition 3 is satisfied with $\alpha = 1$ and $a = 1/(1-2\nu_0)$. Furthermore, $\mathcal{F}$ is convex, and this ℓ satisfies Condition 2, with $\beta = 1$ and $b = 32$ in Lemma 5. Also, since $\psi_{\ell}(x) = x^2$ here [6], we have that $\Psi_{\ell}(\varepsilon) = \varepsilon^{2-\alpha}/(4a) = (1-2\nu_0)\varepsilon/4$. Additionally, this scenario satisfies (10) in Condition 10 with $q = \frac{7}{\omega}$ and $\rho = \frac{1}{3} + \omega$, for any choice of $\omega \in (0, 1/2]$; we include a simple proof of this fact in Appendix B.1. Finally, we bound $\theta(r_0)$ for $r_0 \in (0, 1]$. For any $r \in (0, 1)$, we have $\mathrm{DIS}(\mathrm{B}(f^{\star}, r)) \cap \{x_i : i \in \mathbb{N}\} = \left\{x_i : \frac{90}{\pi^4 i^4} \leq r\right\}$, so that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, r))) \lesssim \sum_{i \gtrsim r^{-1/4}} i^{-4} \lesssim r^{3/4}$. Therefore, $\theta(r_0) \lesssim r_0^{-1/4}$.

Plugging these values into Theorem 12, and choosing $\omega = \left(\ln\left(\frac{1}{(1-2\nu_0)\varepsilon}\right)\right)^{-1}$, we find that there is a label budget n, sufficient to guarantee $\mathrm{er}(\hat{h}) - \mathrm{er}(f^{\star}) \leq \varepsilon$ with probability at least $1-\delta$ in Algorithm 1, with dependence $\Theta\left(\varepsilon^{-7/12}\mathrm{Log}(1/\varepsilon)\right)$ on ε. For comparison, the corresponding bound for ERM_{ℓ} from Theorem 11 has dependence $\Theta\left(\varepsilon^{-4/3}\mathrm{Log}(1/\varepsilon)\right)$. This is larger than the above bound by a factor $\Theta\left(\varepsilon^{-3/4}\right)$. Furthermore, one can show an $\Omega(\varepsilon^{-4/3})$ *lower bound* on the sample size necessary to obtain ε *minimax* expected excess error rate for passive learning in this scenario. Thus, Algorithm 1 achieves a significant improvement over the guarantees achievable by *all* passive learning methods. The details of this minimax lower bound are included in Appendix B.1.

5.6. An Example: Linear Functions

As another example applying the above results, consider the class of *homogeneous linear functions*. Specifically, fix any $k \in \mathbb{N}$ with $k \geq 5$, $\mathcal{X} = \{x \in \mathbb{R}^k : \|x\| \leq 1\}$, and consider the class $\mathcal{F} = \{x \mapsto w \cdot x : w \in \mathbb{R}^k, \|w\| \leq 1\}$. Take ℓ as the quadratic loss (in which case $\bar{\ell} = 4$). Together with the assumption of $f^{\star} \in \mathcal{F}$, this restricts $\mathcal{P}_{XY}$ to have $\eta(x) = (w \cdot x + 1)/2$ (almost everywhere), for some $w \in \mathbb{R}^k$ with $\|w\| \leq 1$.

Furthermore, this ℓ satisfies Condition 2, with $\beta = 1$ and $b = 32$ in Lemma 5, and has $\Psi_\ell(\varepsilon) = \varepsilon^{2-\alpha}/(4a)$. It is also known that $\mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \lesssim k$ (following from arguments of [16, 29]). Additionally, for this class $\mathcal{F}$, it is known that if $\mathcal{P}$ has a density (with respect to Lebesgue measure), then $\theta(\varepsilon) = o(1/\varepsilon)$ [26]. Together, these facts imply that, if $\mathcal{P}$ has a density, the sufficient size of n in Theorem 9 has dependence on ε that is $o\left(\varepsilon^{\alpha-2}\mathrm{Log}(1/\varepsilon)\right)$. We also note that, by varying $\mathcal{P}$, it is possible to realize any α value in $(0, 1]$ in Condition 3 [see 12, 15].

To exhibit a concrete example, consider the simple scenario of $\mathcal{P}$ uniform on $\{x \in \mathbb{R}^k : \|x\| = 1\}$, and suppose $\mathcal{P}_{XY}$ is such that $f^\star \in \mathcal{F}$. For simplicity, also suppose the $w \in \mathbb{R}^k$ with $f^\star(x) = w \cdot x$ satisfies $\|w\| = 1$. In this case, one can show that Condition 3 is satisfied with $a \propto k^{1/4}$ and $\alpha = 1/2$. For completeness, a proof of this is included in Appendix B.2. It is also known that $\theta(\varepsilon) \leq \pi\sqrt{k}$ for this scenario [22]. Plugging all of this into Theorem 9 reveals that, for Algorithm 1 to achieve excess error rate ε with probability at least $1 - \delta$ (given sufficiently large u), it suffices to have a label budget n of size at least

$$c\frac{k}{\varepsilon}\left(k\mathrm{Log}\left(\frac{k}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

for a universal constant $c > 0$. In contrast, Theorem 7 gives a sufficient sample size for $\mathrm{ERM}_\ell(\mathcal{F}, \cdot)$ proportional to $\frac{k^{1/4}}{\varepsilon^{3/2}}\left(k\mathrm{Log}(k) + \mathrm{Log}(1/\delta)\right)$, which is significantly larger than the above size of n for ε sufficiently small. To our knowledge, it is not presently known what the optimal sample complexity of passive learning is for this scenario, so that in contrast to the previous example, here we can only claim an improvement in the upper bound. We note that Dekel, Gentile, and Sridharan [15] have also studied active learning with this $\mathcal{F}$ and ℓ under the same assumption of $f^\star \in \mathcal{F}$, and established a similar result to the above (with slightly better dependence on k but slightly worse logarithmic factors), via a learning method tailored specifically to this function class.

6. General Theorems

The remainder of the article is devoted to a general analysis of Algorithm 1, from which we derive the more-explicit theorems stated above. The results are formulated analogously to localization arguments common in the literature on empirical risk minimization, but with a slight twist to introduce a relevant subregion to the argument. As such, we begin with a discussion of general localized sample complexity bounds.

6.1. Localized Sample Complexities

The derivation of localized excess risk bounds is essentially motivated as follows. We are interested in bounding the excess ℓ-risk of the $\hat{h}$ returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$. Suppose we have a coarse guarantee $U_\ell(\mathcal{H}, m)$ on this value: that is, $\mathrm{R}_\ell(\hat{h}) - \inf_{h\in\mathcal{H}} \mathrm{R}_\ell(h) \leq U_\ell(\mathcal{H}, m)$. In a sense, this guarantee identifies a set $\mathcal{H}' \subseteq \mathcal{H}$ of functions that a priori may 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 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, now thinking of $\hat{h}$ as the function returned by $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m)$: that is, we calculate $U_\ell(\mathcal{H}', m)$ to determine a further subset $\mathcal{H}'' = \mathcal{H}'(U_\ell(\mathcal{H}', m); \ell) \subseteq \mathcal{H}'$ of functions that we again expect to contain the empirical minimizer $\hat{h}$, 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 $U_\ell(\mathcal{H}, m)$ varies with $\mathcal{H}$.

To formalize this fixed-point argument for $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$, Koltchinskii [34] makes use of the following quantities to define the coarse bound $U_\ell(\mathcal{H}, m)$ [see also 8, 20]. 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 $S \sim P^m$, define

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

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

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

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 [34]. 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 $\mathcal{H}(\varepsilon; \ell, S) = \{h \in \mathcal{H} : \mathrm{R}_\ell(h; S) - \inf_{g\in\mathcal{H}} \mathrm{R}_\ell(g; S) \le \varepsilon\}$; 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, let $\hat{\phi}_\ell(\mathcal{H}; \emptyset, \emptyset) = \hat{\mathrm{D}}_\ell(\mathcal{H}; \emptyset) = 0$, and $\hat{U}_\ell(\mathcal{H}; \emptyset, \emptyset, s) = 752\bar{\ell}s$.

The above U 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 [34].

Lemma 15. *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 $S \sim P^m$ 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.*

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

We typically expect the quantities $\bar{U}$, $\hat{U}$, and $\tilde{U}$ to be roughly within constant factors of each other. Following Koltchinskii [34] and Giné and Koltchinskii [20], 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 17 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. 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 more basic properties of $\mathcal{H}(\gamma';\ell,P)$. See [6, 20, 34, 47], or Section 7.1 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 or noise conditions, this approach can still lead to nearly tight bounds [6, 40, 45].

For our purposes, it is convenient to make these common techniques explicit in the results. This will make the benefits of our proposed method more apparent, while still allowing us to state results in a form abstract enough to encompass the more-specific complexity measures referenced in the theorems of Section 5. Toward this end, we have the following definition (recall the definitions of $h_{\mathcal{U},g}$ and $\mathcal{H}_{\mathcal{U},g}$ from Section 4 above).

Definition 16. *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 $f_P^\star \in \mathcal{H}$.*

$$\text{If } 0 \le \sigma \le \sigma', \mathcal{H} \subseteq \mathcal{H}' \subseteq [\mathcal{F}], \mathcal{U} \subseteq \mathcal{X}, \text{ and } m' \le m,$$
$$\text{then } \mathring{\phi}_\ell(\sigma, \mathcal{H}_{\mathcal{U},f_P^\star}; m, P) \le \mathring{\phi}_\ell(\sigma', \mathcal{H}'; m', P). \tag{14}$$
$$\forall \sigma \ge \mathrm{D}_\ell(\mathcal{H}; P), \phi_\ell(\mathcal{H}; m, P) \le \mathring{\phi}_\ell(\sigma, \mathcal{H}; m, P). \tag{15}$$

For instance, most bounds based on entropy integrals can be made to satisfy this. Section 7.1 states 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 $\mathcal{X} \times \mathcal{Y}$.

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

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 16 implies $\phi_\ell(\mathcal{H}(\gamma; \ell, P); m, P) \le \mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}(\gamma; \ell, P); m, P)$, and furthermore $\mathcal{H}(\gamma; \ell, P) \subseteq \mathcal{H}$ so that $\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}(\gamma; \ell, P); m, P) \le \mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}; m, P)$. Thus,

$$\tilde{U}_\ell(\mathcal{H}(\gamma; \ell, P); P, m, s) \le \mathring{U}_\ell(\mathcal{H}(\gamma; \ell, P), \gamma; P, m, s) \le \mathring{U}_\ell(\mathcal{H}, \gamma; P, m, s). \tag{16}$$

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

$$\mathring{U}_\ell(\mathcal{H}_{\mathcal{U},f_P^\star}, \gamma; P, m, s) \le \mathring{U}_\ell(\mathcal{H}'_{\mathcal{U}',f_P^\star}, \gamma'; P, m, s). \tag{17}$$

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 $\mathrm{D}_\ell(\mathcal{H}_{\mathcal{U},f_P^\star}(\gamma; \ell, P); P) \le \mathrm{D}_\ell(\mathcal{H}_{\mathcal{U}',f_P^\star}(\gamma; \ell, P); P)$, 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 $f_P^\star \in [\mathcal{H}]$, and therefore $\mathrm{D}_\ell([\mathcal{H}_{\mathcal{U},f_P^\star}](\gamma; \ell, P); P) \le \mathrm{D}_\ell([\mathcal{H}_{\mathcal{U}',f_P^\star}](\gamma; \ell, P); P)$.

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

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

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

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

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

$$\mathring{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) \le \max\left\{\ddot{\mathrm{M}}_\ell(\tilde{c}\gamma_1, \gamma_2; \mathcal{H}, P), \dot{\mathrm{M}}_\ell(\tilde{c}\gamma_1, \gamma_2; \mathcal{H}, P, s)\right\}. \tag{18}$$

Also note that we clearly have

$$\dot{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) \le s \cdot \max\left\{\frac{4\mathrm{D}_\ell([\mathcal{H}](\gamma_2; \ell, P); \ell, P)^2}{\gamma_1^2}, \frac{2\bar{\ell}}{\gamma_1}\right\}, \tag{19}$$

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 explicit results for ERM_ℓ. When $f_P^\star \in \mathcal{H}$, (16) implies

$$\bar{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) \le \tilde{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) \le \mathring{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}). \tag{20}$$

6.2. General Analysis of Empirical Risk Minimization

Based on Lemma 15 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 [34] (slightly modified here, following Giné and Koltchinskii [20], to allow for general $\mathfrak{s}$ functions). It will be useful for deriving Theorems 7 and 11. For $\varepsilon > 0$, let $\mathbb{Z}_\varepsilon = \{j \in \mathbb{Z} : 2^j \ge \varepsilon\}$.

Theorem 17. *Fix any function* $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$. *If* $f^\star \in \mathcal{F}$, *then for any* $m \ge \bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s})$, *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* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \le \varepsilon$.

6.3. Specification of $\hat{T}_\ell$ in Algorithm 1

The quantity $\hat{T}_\ell$ in Algorithm 1 can be defined in one of several possible ways. In our present abstract 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 variables 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 $\Xi[Q] = \{\xi'_{i_k}\}_{k=1}^q$, and for $s \geq 1$, define $\hat{U}_\ell(\mathcal{H}; Q, s) = \hat{U}_\ell(\mathcal{H}; S[Q], \Xi[Q], s)$, where $S[Q] = \{(X_{i_1}, y_1), \ldots, (X_{i_q}, y_q)\}$, as previously defined. Then we can define the quantity $\hat{T}_\ell$ in the method above as

$$\hat{T}_\ell(\mathcal{H}; Q, m) = \hat{U}_\ell(\mathcal{H}; Q, \hat{\mathfrak{s}}(m)), \tag{21}$$

for some $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. This definition has the appealing property that it allows us to interpret the update in Step 6 in two complementary ways: as comparing the empirical risks of functions in V under samples from the conditional distribution of (X, Y) given $X \in \mathrm{DIS}(V)$, and as comparing the empirical risks of the functions in $V_{\mathrm{DIS}(V)}$ under samples from the original distribution $\mathcal{P}_{XY}$. Our abstract results below are based on this definition of $\hat{T}_\ell$. This can sometimes be problematic due to the computational challenge of the optimization problems in the definitions of $\hat{\phi}_\ell$ and $\hat{\mathrm{D}}_\ell$. There has been considerable work on calculating and bounding $\hat{\phi}_\ell$ for various classes $\mathcal{F}$ and losses ℓ [e.g., 7, 33], but it is not always feasible. However, the specific theorems stated in Section 5 above continue to hold if we instead take $\hat{T}_\ell$ based on a well-chosen upper bound on the respective $\mathring{U}_\ell$ function, such as those obtained in the derivations of those respective results below; we provide descriptions of such efficiently-computable relaxations, for each of these results, in Appendix D (though in some cases, these bounds have a mild dependence on $\mathcal{P}_{XY}$ via certain parameters of the specific noise conditions considered there).

6.4. General Analysis of Algorithm 1

The following theorem represents our main abstract result. The key steps in its proof were already sketched above in Section 4. The complete proof is included in Appendix A.

Theorem 18. *Fix any function* $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. *Let* $j_\ell = -\lceil \log_2(\bar{\ell}) \rceil$, $u_{j_\ell - 2} = u_{j_\ell - 1} = 1$, *and for each integer* $j \geq j_\ell$, *let* $\mathcal{F}_j = \mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01})_{\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); 01))}$, $\mathcal{U}_j = \mathrm{DIS}(\mathcal{F}_j)$, *and suppose* $u_j \in \mathbb{N}$ *satisfies* $\log_2(u_j) \in \mathbb{N}$ *and*

$$u_j \geq 2\mathring{\mathrm{M}}_\ell(2^{-j-1}, 2^{2-j}; \mathcal{F}_j, \mathcal{P}_{XY}, \hat{\mathfrak{s}}(u_j)) \vee u_{j-1} \vee 2u_{j-2}. \tag{22}$$

Suppose $f^\star \in \mathcal{F}$. *For any* $\varepsilon \in (0, 1)$, $s \in [1, \infty)$, *letting* $j_\varepsilon = \lceil \log_2(1/\Gamma_\ell(\varepsilon)) \rceil$, *if*

$$u \geq u_{j_\varepsilon} \qquad \text{and} \qquad n \geq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j,$$

then, with arguments ℓ, u, *and* n, *Algorithm 1 uses at most* u *unlabeled samples, requests at most* n *labels, and with probability at least* $1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)}$, *returns a function* $\hat{h}$ *with* $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

In defining and calculating the values $\mathring{\mathrm{M}}_\ell$ in Theorem 18, it is sometimes convenient to use the alternative interpretation of Algorithm 1, in terms of sampling the set $S[Q]$ from the conditional distribution given the region of disagreement. Specifically, for any measurable $\mathcal{U} \subseteq \mathcal{X}$ with $\mathcal{P}(\mathcal{U}) > 0$, define the probability measure $\mathcal{P}_{\mathcal{U}}(\cdot) = \mathcal{P}_{XY}(\cdot|\mathcal{U} \times \mathcal{Y})$: that is, $\mathcal{P}_{\mathcal{U}}$ is the conditional distribution of $(X, Y) \sim \mathcal{P}_{XY}$ given that $X \in \mathcal{U}$. Generally, for any probability measure P on $\mathcal{X} \times \mathcal{Y}$, and 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 $\mathcal{H} \subseteq \mathcal{F}^*$, define 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 $\overline{\mathrm{DISF}}(\mathcal{H}) = \mathrm{DISF}(\mathcal{H}) \times \mathcal{Y}$. The following lemma then allows us to replace calculations in terms of $\mathcal{F}_j$ and $\mathcal{P}_{XY}$ with calculations in terms of $\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01})$ and $\mathcal{P}_{\mathrm{DIS}(\mathcal{F}_j)}$. Its proof is included in Appendix A.

Lemma 19. *Let $\mathring{\phi}_\ell$ be any function satisfying Definition 16. Let P be any distribution over $\mathcal{X} \times \mathcal{Y}$. For any $\sigma \geq 0$, $\mathcal{H} \subseteq [\mathcal{F}]$, $m \in \mathbb{N}$, if $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$, define*

$$\mathring{\phi}'_\ell(\sigma, \mathcal{H}; m, P) =$$
$$32\left(\inf_{\substack{\mathcal{U}=\mathcal{U}'\times\mathcal{Y}:\\ \mathcal{U}'\supseteq\mathrm{DISF}(\mathcal{H})}} P(\mathcal{U})\mathring{\phi}_\ell\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil(1/2)P(\mathcal{U})m\rceil, P_{\mathcal{U}}\right) + \frac{\bar{\ell}}{m} + \sigma\sqrt{\frac{1}{m}}\right), \tag{23}$$

and otherwise $\mathring{\phi}'_\ell(\sigma, \mathcal{H}; m, P) = 0$. Then $\mathring{\phi}'_\ell$ also satisfies Definition 16.

Plugging this $\mathring{\phi}'_\ell$ function into Theorem 18 immediately yields the following corollary; the proof is included in Appendix A.

Corollary 20. *Fix any function $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. Let $j_\ell = -\lceil\log_2(\bar{\ell})\rceil$, define $u_{j_\ell-2} = u_{j_\ell-1} = 1$, and for each integer $j \geq j_\ell$, let $\mathcal{F}_j$ and $\mathcal{U}_j$ be as in Theorem 18, and if $\mathcal{P}(\mathcal{U}_j) > 0$, suppose $u_j \in \mathbb{N}$ satisfies $\log_2(u_j) \in \mathbb{N}$ and*

$$u_j \geq 4\mathcal{P}(\mathcal{U}_j)^{-1}\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \hat{\mathfrak{s}}(u_j)\right) \vee u_{j-1} \vee 2u_{j-2}. \tag{24}$$

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

$$u \geq u_{j_\varepsilon} \qquad \text{and} \qquad n \geq s + 2e\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j,$$

then, with arguments ℓ, u, and n, Algorithm 1 uses at most u unlabeled samples, requests at most n labels, and with probability at least $1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)}$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

7. Derivations of the Explicit Results

We are now ready to present derivations of the explicit results of Section 5, based on the general results of the previous section. 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 numerical constant $c \in (0, \infty)$.

7.1. Specification of $\mathring{\phi}_\ell$

We begin by recalling a few well-known bounds on the ϕ_ℓ function, which lead to a more concrete instance of a function $\mathring{\phi}_\ell$ satisfying Definition 16.

Uniform Entropy: The first bound is based on the work of van der Vaart and Wellner [48]; related bounds have been studied by Giné and Koltchinskii [20], Giné, Koltchinskii, and Wellner [21], van der Vaart and Wellner [47], and others. For $\sigma \geq 0$ and $\mathrm{F} \in \mathcal{G}^*$, define the function

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

where Π ranges over all finitely discrete probability measures.

Fix any distribution P on $\mathcal{X} \times \mathcal{Y}$. 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 [48] (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 $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$,

$$\phi_\ell(\mathcal{H}; m, P) \leq \tag{25}$$
$$cJ\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\sigma^2 m}\right).$$

Based on (25), it is straightforward to define a function $\mathring{\phi}_\ell$ that satisfies Definition 16. Specifically, define

$$\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P) =$$
$$\inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})} \inf_{\lambda \geq \sigma} cJ\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\lambda^2 m}\right), \tag{26}$$

for c as in (25). By (25), $\mathring{\phi}_\ell^{(1)}$ satisfies (15). 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, $\mathcal{H} \mapsto \mathcal{N}(\varepsilon, \mathcal{G}_\mathcal{H}, L_2(\Pi))$ is nondecreasing for all Π, 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 $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{N}(\varepsilon, \mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star}}, L_2(\Pi)) \leq \mathcal{N}(\varepsilon, \mathcal{G}_\mathcal{H}, L_2(\Pi))$ for all Π, so that $J(\sigma, \mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star}}, \mathrm{F}) \leq J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F})$. Since $\mathrm{F}(\mathcal{G}_{\mathcal{H}_{\mathcal{U},f_P^\star},P}) \leq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, 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 16, 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 $\sigma \geq 0$, 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 [47] and a triangle inequality imply that for some universal constant $c \in [1, \infty)$, for any $m \in \mathbb{N}$ and $\sigma \geq \mathrm{D}_{\ell}(\mathcal{H}; P)$,

$$\phi_{\ell}(\mathcal{H}; m, P) \leq c J_{[]}\left(\sigma, \mathcal{G}_{\mathcal{H}}, P\right) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}\left(\sigma, \mathcal{G}_{\mathcal{H}}, P\right) \bar{\ell}}{\sigma^2 m} \right). \tag{27}$$

As-is, the right side of (27) nearly satisfies Definition 16 already. Only a small change is needed for the requirement of monotonicity in σ. Specifically, define

$$\mathring{\phi}_{\ell}^{(2)}(\sigma, \mathcal{H}; m, P) = \inf_{\lambda \geq \sigma} c J_{[]}\left(\lambda, \mathcal{G}_{\mathcal{H}}, P\right) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}\left(\lambda, \mathcal{G}_{\mathcal{H}}, P\right) \bar{\ell}}{\lambda^2 m} \right), \tag{28}$$

for c as in (27). Then taking $\mathring{\phi}_{\ell} = \mathring{\phi}_{\ell}^{(2)}$ suffices to satisfy Definition 16.

Since Definition 16 is satisfied for both $\mathring{\phi}_{\ell}^{(1)}$ and $\mathring{\phi}_{\ell}^{(2)}$, it is also satisfied for $\mathring{\phi}_{\ell} = \min\left\{\mathring{\phi}_{\ell}^{(1)}, \mathring{\phi}_{\ell}^{(2)}\right\}$. The remainder of this section takes this as the specification of the $\mathring{\phi}_{\ell}$ function.

7.2. VC Subgraph Classes

The following is a classic result for VC subgraph classes [see e.g., 47], derived from the works of Pollard [42] and Haussler [29].

Lemma 21. *For any $\mathcal{G} \subseteq \mathcal{G}^*$, for any measurable $\mathrm{F} \geq \mathrm{F}(\mathcal{G})$, for any distribution Π such that $\|\mathrm{F}\|_{\Pi} > 0$, for any $\varepsilon \in (0, 1)$,*

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

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

In particular, Lemma 21 implies that any $\mathcal{G} \subseteq \mathcal{G}^*$ has, $\forall \sigma \in (0, 1]$,

$$J(\sigma, \mathcal{G}, \mathrm{F}) \leq \int_0^{\sigma} \sqrt{\ln(eA(\mathcal{G})) + 2\mathrm{vc}(\mathcal{G})\ln(1/\varepsilon)}\,\mathrm{d}\varepsilon \lesssim \sigma\sqrt{\mathrm{vc}(\mathcal{G})\mathrm{Log}(1/\sigma)}. \tag{29}$$

Applying these observations to $J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$ 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 (26) yields the following well-known bound on $\mathring{\phi}_{\ell}^{(1)}$ due to Giné and Koltchinskii [20]. For any $m \in \mathbb{N}$ and $\sigma > 0$,

$$\begin{aligned} &\mathring{\phi}_{\ell}^{(1)}(\sigma, \mathcal{H}; m, P) \\ &\quad \lesssim \inf_{\lambda \geq \sigma} \lambda \sqrt{\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{H}})\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}} + \frac{\mathrm{vc}(\mathcal{G}_{\mathcal{H}})\bar{\ell}\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}. \end{aligned} \tag{30}$$

Specifically, to arrive at (30), we relaxed the $\inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})}$ in (26) 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 (29) remains valid; we also used 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, (30) implies

$$\ddot{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P) \lesssim \inf_{\sigma \geq \mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);P)} \left(\frac{\sigma^2}{\gamma_1^2} + \frac{\bar{\ell}}{\gamma_1}\right) \mathrm{vc}(\mathcal{G}_{\mathcal{H}}) \mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\sigma}\right). \quad (31)$$

For $\lambda > 0$, when $f_P^\star \in \mathcal{H}$ and P satisfies Condition 4, (31) implies that,

$$\sup_{\gamma \geq \lambda} \ddot{\mathrm{M}}_\ell(\gamma/(4\tilde{K}), \gamma; \mathcal{H}(\gamma; \ell, P), P) \lesssim \left(\frac{b}{\lambda^{2-\beta}} + \frac{\bar{\ell}}{\lambda}\right) \mathrm{vc}(\mathcal{G}_{\mathcal{H}}) \mathrm{Log}\left(\tau_\ell\left(b\lambda^\beta; \mathcal{H}, P\right)\right). \quad (32)$$

Combining this observation with (16), (18), (19), (20), and Theorem 17, we arrive at a result for the sample complexity of empirical ℓ-risk minimization with a general VC subgraph class under Conditions 3 and 4. Specifically, for $\mathfrak{s} : (0,\infty)^2 \to [1,\infty)$, when $f^\star \in \mathcal{F}$, (16) implies that

$$\begin{aligned}\bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s}) &\leq \tilde{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s}) \\ &= \sup_{\gamma \geq \Gamma_\ell(\varepsilon)} \tilde{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma)) \\ &\leq \sup_{\gamma \geq \Gamma_\ell(\varepsilon)} \mathring{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma)). \quad (33)\end{aligned}$$

For $\mathcal{P}_{XY}$ satisfying Conditions 3 and 4, applying (18), (19), and (32) to (33), and taking $\mathfrak{s}(\lambda, \gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at Theorem 7 (which is implicit in [20]).

Next, we turn to Theorem 8. Note that $\mathrm{vc}(\mathcal{G}_{\mathcal{F}_j}) \leq \mathrm{vc}(\mathcal{G}_{\mathcal{F}(\mathcal{E}_\ell(2^{2-j});01)}) \leq \mathrm{vc}(\mathcal{G}_{\mathcal{F}})$. Also, $\|\mathrm{F}(\mathcal{G}_{\mathcal{F}_j,\mathcal{P}_{XY}})\|_{\mathcal{P}_{XY}}^2 \leq \bar{\ell}^2 \mathcal{P}\left(\mathrm{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{2-j}\right); {}_{01}\right)\right)\right)$. Thus, for $j_\ell \leq j \leq \lceil \log_2(1/\Psi_\ell(\varepsilon)) \rceil$, (31) implies

$$\ddot{\mathrm{M}}_\ell(2^{-j-2}\tilde{K}^{-1}, 2^{2-j}; \mathcal{F}_j, \mathcal{P}_{XY}) \lesssim \left(b2^{j(2-\beta)} + \bar{\ell}2^j\right) \mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \mathrm{Log}\left(\chi_\ell\left(\Psi_\ell(\varepsilon)\right)\bar{\ell}\right). \quad (34)$$

With a little additional work to define an appropriate $\hat{\mathfrak{s}}$ function and derive closed-form bounds on the summation in Theorem 18, we arrive at Theorem 8. The remaining details appear in Appendix B.

When ℓ satisfies Condition 2, we can derive the sometimes-stronger result in Theorem 9 via Corollary 20. Specifically, combining (31), (18), (19), and Lemma 5, we have that if $f^\star \in \mathcal{F}$ and Condition 2 is satisfied, then for $j \geq j_\ell$ in Corollary 20,

$$\begin{aligned}&\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) \quad (35) \\ &\lesssim \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + 2^j\bar{\ell}\mathcal{P}(\mathcal{U}_j)\right)\left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\bar{\ell}^2 2^{j\beta}\mathcal{P}(\mathcal{U}_j)^\beta/b\right) + s\right),\end{aligned}$$

where b and β are as in Lemma 5. Plugging this into Corollary 20, we arrive at Theorem 9; the remaining details proceed similarly to those of Theorem 8, and a detailed sketch appears in Appendix B.

7.3. Entropy Conditions

Next we turn to problems satisfying entropy conditions. Note that when $\mathcal{F}$ satisfies Condition 10, for $0 \leq \sigma \leq 2\|\mathrm{F}\|_{\mathcal{P}_{XY}}$,

$$\mathring{\phi}_{\ell}(\sigma, \mathcal{F}; m, \mathcal{P}_{XY}) \lesssim \max\left\{ \frac{\sqrt{q}\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{\rho}\sigma^{1-\rho}}{(1-\rho)m^{1/2}}, \frac{\bar{\ell}^{\frac{1-\rho}{1+\rho}} q^{\frac{1}{1+\rho}} \|\mathrm{F}\|_{\mathcal{P}_{XY}}^{\frac{2\rho}{1+\rho}}}{(1-\rho)^{\frac{2}{1+\rho}} m^{\frac{1}{1+\rho}}} \right\}. \tag{36}$$

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 4, then

$$\ddot{\mathrm{M}}_{\ell}(c\gamma, \gamma; \mathcal{F}, \mathcal{P}_{XY}) \lesssim \frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2} \max\left\{ b^{1-\rho}\gamma^{\beta(1-\rho)-2}, \bar{\ell}^{1-\rho}\gamma^{-(1+\rho)} \right\}. \tag{37}$$

Combined with (18), (19), (20), and Theorem 17, taking $\mathfrak{s}(\lambda,\gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at the classic result in Theorem 11 [e.g., 6, 47].

The corresponding result for Algorithm 1, namely Theorem 12, follows by combining (37) with (18), (19), and Theorem 18. The details of the proof follow analogously to that of Theorem 8, and are therefore omitted for brevity.

Next, we turn to deriving the corresponding results stated above under Condition 2. As discussed above, we treat separately the cases of (11) and (10).

First, suppose (11) holds (for all P, ε) with $\mathrm{F} \leq \bar{\ell}$. Following the derivation of (37) above, combined with (19), (18), and Lemma 5, for $j \geq j_{\ell}$ in Corollary 20,

$$\begin{aligned}\mathring{\mathrm{M}}_{\ell}\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) &\lesssim \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + \bar{\ell}2^j\mathcal{P}(\mathcal{U}_j)\right)s \\ &\quad + \frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\left(b^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{1+\rho}\right),\end{aligned}$$

where b and β are from Lemma 5. This immediately leads to Theorem 13 by reasoning analogous to the proof of Theorem 9.

The case (10) can be treated similarly, though the result we obtain (Theorem 14) is slightly weaker. Suppose (10) is satisfied with $\mathrm{F} = \bar{\ell}$ constant. In this case, $\bar{\ell} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F}_j,\mathcal{P}_{\mathcal{U}_j}})$, while $\mathcal{N}_{[]}(\varepsilon\bar{\ell}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{\mathcal{U}_j})) = \mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{XY})) \leq \mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY}))$, so that $\mathcal{F}_j$ and $\mathcal{P}_{\mathcal{U}_j}$ also satisfy (10) with $\mathrm{F} = \bar{\ell}$:

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

Thus, based on (37), (18), (19), and Lemma 5, we have that if $f^{\star} \in \mathcal{F}$ and Condition 2 is satisfied, then for $j \geq j_\ell$ in Corollary 20,

$$\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right) \lesssim \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + \bar{\ell}2^j\mathcal{P}(\mathcal{U}_j)\right) s$$
$$+\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\mathcal{P}(\mathcal{U}_j)^{-\rho}\left(b^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{1+\rho}\right),$$

where b and β are as in Lemma 5. Combining this with Corollary 20 and reasoning analogously to the proof of Theorem 9, we obtain Theorem 14.

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] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2: 343–370, 1988. 1

[3] 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

[4] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 1, 1.1, 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, 3.2, 5.1, 5.1, 5.1, 5.1, 5.1, 5.4, 5.5, 6.1, 7.3, C, F

[7] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(11):463–482, 2002. 6.3

[8] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 1, 6.1

[9] 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

[10] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003. C

[11] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 5.3

[12] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83:71–102, 2011. 5.6

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

[14] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning

algorithm. In *Advances in Neural Information Processing Systems*, 2007. 1.1, 4, 5.2

[15] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13: 2655–2697, 2012. 5.6

[16] R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978. 5.6

[17] R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4):1306–1326, 1987. 5.3

[18] 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

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

[20] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 5.2, 5.3, 5.3, 6.1, 6.1, 6.2, 7.1, 7.2, 7.2, C

[21] E. Giné, V. Koltchinskii, and J. Wellner. Ratio limit theorems for empirical processes. In *Stochastic Inequalities*, pages 249–278. Birkhäuser, 2003. 7.1

[22] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, 2007. 1.1, 5.2, 5.2, 5.6, B.2

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

[24] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39 (1):333–361, 2011. 1, 1.1, 5.1, 5.2, 5.2, 5.3, 5.3

[25] S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13:1469–1587, 2012. 1.1, 4, 5.1, 5.2, 5.2

[26] S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014. 5.6

[27] S. Hanneke. Nonparametric active learning, part 1: Smooth regression functions. *Unpublished Manuscript*, 2016. 1.1

[28] 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

[29] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992. 5.3, 5.6, 7.2

[30] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005. 1

[31] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association for Computing Machinery*, 45(6):983–1006, 1998. 1

[32] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning.

Machine Learning, 17:115–141, 1994. 1
[33] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001. 6.3
[34] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 1, 5.1, 5.1, 6.1, 6.1, 6.2, D, E.2
[35] 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
[36] 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
[37] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011. B.2
[38] A. Locatelli, A. Carpentier, and S. Kpotufe. Adaptivity to noise parameters in nonparametric active learning. In *Proceedings of the 30*$^{\text{th}}$ *Conference on Learning Theory*, 2017. 1.1
[39] S. Mahalanabis. A note on active learning for smooth problems. arXiv*:1103.3095*, 2011. 5.2
[40] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:1808–1829, 1999. 5.1, 5.1, 6.1
[41] S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1):67–90, 2012. 1.1
[42] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984. 7.2
[43] D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Inst. of Math. Stat. and Am. Stat. Assoc., 1990. 5.3
[44] 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, 5.3, 5.3
[45] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 5.1, 5.1, 6.1
[46] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. B.1
[47] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 5.3, 5.4, 6.1, 7.1, 7.1, 7.2, 7.3, A, C
[48] A. W. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. 2, 7.1
[49] 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. 5.3
[50] 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
[51] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004. 1, 1.1

Appendix A: Main Proofs

This appendix includes the proofs of the main abstract results from Section 6.

Proof of Theorem 18. Fix any $\varepsilon \in (0,1)$, $s \in [1,\infty)$, values u_j satisfying (22), and consider running Algorithm 1 with values of u and n satisfying the conditions specified in Theorem 18. 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 15.

Let S denote the set of values of m obtained in Algorithm 1 for which $\log_2(m) \in \mathbb{N}$. For each $m \in S$, let $V^{(m)}$ and Q_m denote the values of V and Q (respectively) upon reaching Step 5 on the round that Algorithm 1 obtains that value of m, and let $\tilde{V}^{(m)}$ denote the value of V upon completing Step 6 on that round; also denote $D_m = \mathrm{DIS}(V^{(m)})$ and $\mathcal{L}_m = \{(1+m/2, Y_{1+m/2}), \ldots, (m, Y_m)\}$, and define $\tilde{V}^{(1)} = \mathcal{F}$ and $D_1 = \mathrm{DIS}(\mathcal{F})$.

Consider any $m \in S$, and note that $\forall h, g \in V^{(m)}$,

$$(|Q_m| \vee 1)\left(\mathrm{R}_\ell(h; Q_m) - \mathrm{R}_\ell(g; Q_m)\right) = \frac{m}{2}\left(\mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_\ell(g_{D_m}; \mathcal{L}_m)\right), \quad (38)$$

and furthermore that

$$(|Q_m| \vee 1)\hat{U}_\ell(V^{(m)}; Q_m, \hat{\mathfrak{s}}(m)) = \frac{m}{2}\hat{U}_\ell(V^{(m)}_{D_m}; \mathcal{L}_m, \hat{\mathfrak{s}}(m)). \quad (39)$$

Applying Lemma 15 under the conditional distribution given $V^{(m)}$, combined with the law of total probability, we have that, for every $m \in \mathbb{N}$ with $\log_2(m) \in \mathbb{N}$, on an event of probability at least $1 - 6e^{-\hat{\mathfrak{s}}(m)}$, if $f^\star \in V^{(m)}$ and $m \in S$, then letting $\hat{U}_m = \hat{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{L}_m, \hat{\mathfrak{s}}(m)\right)$, every $h_{D_m} \in V^{(m)}_{D_m}$ has

$$\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(f^\star) < \mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_\ell(f^\star; \mathcal{L}_m) + \hat{U}_m, \quad (40)$$

$$\mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \min_{g_{D_m} \in V^{(m)}_{D_m}} \mathrm{R}_\ell(g_{D_m}; \mathcal{L}_m) < \mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(f^\star) + \hat{U}_m, \quad (41)$$

and furthermore

$$\hat{U}_m < \tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right). \quad (42)$$

By a union bound, on an event of probability at least $1 - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)}$, for every $m \in S$ with $m \le u_{j_\varepsilon}$ and $f^\star \in V^{(m)}$, the inequalities (40), (41), and (42) hold. Call this event E.

In particular, note that on the event E, for any $m \in S$ with $m \le u_{j_\varepsilon}$ and $f^\star \in V^{(m)}$, since $f^\star_{D_m} = f^\star$, (38), (41), and (39) imply

$$(|Q_m| \vee 1)\left(\mathrm{R}_\ell(f^\star; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m)\right) = \frac{m}{2}\left(\mathrm{R}_\ell(f^\star; \mathcal{L}_m) - \inf_{g_{D_m} \in V^{(m)}_{D_m}} \mathrm{R}_\ell(g_{D_m}; Q_m)\right) < \frac{m}{2}\hat{U}_m = (|Q_m| \vee 1)\hat{U}_\ell(V^{(m)}; Q_m, \hat{\mathfrak{s}}(m)),$$

so that $f^\star \in \tilde{V}^{(m)}$ as well. Since $f^\star \in V^{(2)}$, and every $m \in S$ with $m > 2$ has $V^{(m)} = \tilde{V}^{(m/2)}$, by induction we have that, on the event E, every $m \in S$ with $m \leq u_{j_\varepsilon}$ has $f^\star \in V^{(m)}$ and $f^\star \in \tilde{V}^{(m)}$; this also implies that (40), (41), and (42) all hold for these values of m on the event E.

We next prove by induction that, on the event E, $\forall j \in \{j_\ell - 2, j_\ell - 1, j_\ell, \ldots, j_\varepsilon\}$, if $u_j \in S \cup \{1\}$, then $\tilde{V}^{(u_j)}_{D_{u_j}} \subseteq [\mathcal{F}](2^{-j}; \ell)$ and $\tilde{V}^{(u_j)} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{-j}); {\scriptstyle 01}\right)$. This claim is trivially satisfied for $j \in \{j_\ell - 2, j_\ell - 1\}$, since in that case $[\mathcal{F}](2^{-j}; \ell) = [\mathcal{F}] \supseteq \tilde{V}^{(u_j)}_{D_{u_j}}$ and $\mathcal{F}(\mathcal{E}_\ell(2^{-j}); {\scriptstyle 01}) = \mathcal{F}$, so that these values can serve as our base case. Now take as an inductive hypothesis that, for some $j \in \{j_\ell, \ldots, j_\varepsilon\}$, if $u_{j-2} \in S \cup \{1\}$, then on the event E, $\tilde{V}^{(u_{j-2})}_{D_{u_{j-2}}} \subseteq [\mathcal{F}](2^{2-j}; \ell)$ and $\tilde{V}^{(u_{j-2})} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{2-j}); {\scriptstyle 01}\right)$, and suppose the event E occurs. If $u_j \notin S$, the claim is trivially satisfied; otherwise, suppose $u_j \in S$, which further implies $u_{j-2} \in S \cup \{1\}$. Since $u_j \leq u_{j_\varepsilon}$, for any $h \in \tilde{V}^{(u_j)}$, (40) implies

$$\frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(f^\star)\right) < \frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}; \mathcal{L}_{u_j}) - \mathrm{R}_\ell(f^\star; \mathcal{L}_{u_j}) + \hat{U}_{u_j}\right).$$

Since we have already established that $f^\star \in V^{(u_j)}$, (38) and (39) imply

$$\frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}; \mathcal{L}_{u_j}) - \mathrm{R}_\ell(f^\star; \mathcal{L}_{u_j}) + \hat{U}_{u_j}\right) = (|Q_{u_j}| \vee 1)\left(\mathrm{R}_\ell(h; Q_{u_j}) - \mathrm{R}_\ell(f^\star; Q_{u_j}) + \hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right).$$

The definition of $\tilde{V}^{(u_j)}$ from Step 6 implies

$$(|Q_{u_j}| \vee 1)\left(\mathrm{R}_\ell(h; Q_{u_j}) - \mathrm{R}_\ell(f^\star; Q_{u_j}) + \hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right) \leq (|Q_{u_j}| \vee 1)\left(2\hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right).$$

By (39) and (42),

$$(|Q_{u_j}| \vee 1)\left(2\hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right) = u_j \hat{U}_{u_j} < u_j \tilde{U}_\ell\left(V^{(u_j)}_{D_{u_j}}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right).$$

Altogether, we have that, $\forall h \in \tilde{V}^{(u_j)}$,

$$\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(f^\star) < 2\tilde{U}_\ell\left(V^{(u_j)}_{D_{u_j}}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right). \tag{43}$$

By definition of $\mathring{\mathrm{M}}_\ell$, monotonicity of $m \mapsto \mathring{U}_\ell(\cdot,\cdot;\cdot,m,\cdot)$, and the condition on u_j in (22), we know that

$$\mathring{U}_\ell\left(\mathcal{F}_j, 2^{2-j}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

The fact that $u_j \geq 2u_{j-2}$, combined with the inductive hypothesis, implies

$$V^{(u_j)} \subseteq \tilde{V}^{(u_{j-2})} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{2-j}); {\scriptstyle 01}\right).$$

This also implies $D_{u_j} \subseteq \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {\scriptstyle 01}))$. Combined with (17), these imply

$$\mathring{U}_\ell\left(V^{(u_j)}_{D_{u_j}}, 2^{2-j}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

Together with (16), this implies

$$\tilde{U}_\ell\left(V^{(u_j)}_{D_{u_j}}(2^{2-j};\ell); \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

The inductive hypothesis implies $V^{(u_j)}_{D_{u_j}} = V^{(u_j)}_{D_{u_j}}(2^{2-j};\ell)$, which means

$$\tilde{U}_\ell\left(V^{(u_j)}_{D_{u_j}}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

Plugging this into (43) implies, $\forall h \in \tilde{V}^{(u_j)}$,

$$\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(f^\star) < 2^{-j}. \tag{44}$$

In particular, since $f^\star \in \mathcal{F}$, we always have $\tilde{V}^{(u_j)}_{D_{u_j}} \subseteq [\mathcal{F}]$, so that (44) establishes that $\tilde{V}^{(u_j)}_{D_{u_j}} \subseteq [\mathcal{F}](2^{-j};\ell)$. Furthermore, since $f^\star \in V^{(u_j)}$ on E, $\mathrm{sign}(h_{D_{u_j}}) = \mathrm{sign}(h)$ for every $h \in \tilde{V}^{(u_j)}$, so that every $h \in \tilde{V}^{(u_j)}$ has $\mathrm{er}(h) = \mathrm{er}(h_{D_{u_j}})$, and therefore (by definition of $\mathcal{E}_\ell(\cdot)$), (44) implies

$$\mathrm{er}(h) - \mathrm{er}(f^\star) = \mathrm{er}(h_{D_{u_j}}) - \mathrm{er}(f^\star) \leq \mathcal{E}_\ell\left(2^{-j}\right).$$

This implies $\tilde{V}^{(u_j)} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{-j}); {\scriptstyle 01}\right)$, which completes the inductive proof. This implies that, on the event E, if $u_{j_\varepsilon} \in S$, then (by monotonicity of $\mathcal{E}_\ell(\cdot)$ and the fact that $\mathcal{E}_\ell(\Gamma_\ell(\varepsilon)) \leq \varepsilon$)

$$\tilde{V}^{(u_{j_\varepsilon})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{-j_\varepsilon}); {\scriptstyle 01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(\Gamma_\ell(\varepsilon)); {\scriptstyle 01}) \subseteq \mathcal{F}(\varepsilon; {\scriptstyle 01}).$$

In particular, since the update in Step 6 always keeps at least one element in V, the function $\hat{h}$ in Step 8 exists, and has $\hat{h} \in \tilde{V}^{(u_{j_\varepsilon})}$ (if $u_{j_\varepsilon} \in S$). Thus, on the event E, if $u_{j_\varepsilon} \in S$, then $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. Therefore, since $u \geq u_{j_\varepsilon}$, to complete the proof it suffices to show that taking n of the size indicated in the theorem statement suffices to guarantee $u_{j_\varepsilon} \in S$, on an event (which includes E) having at least the stated probability.

Note that for any $j \in \{j_\ell, \ldots, j_\varepsilon\}$ with $u_{j-1} \in S \cup \{1\}$, every $m \in \{u_{j-1} + 1, \ldots, u_j\} \cap S$ has $V^{(m)} \subseteq \tilde{V}^{(u_{j-1})}$; furthermore, we showed above that on the event E, if $u_{j-1} \in S$, then $\tilde{V}^{(u_{j-1})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01})$, so that $\mathrm{DIS}(V^{(m)}) \subseteq \mathrm{DIS}(\tilde{V}^{(u_{j-1})}) \subseteq \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01})) \subseteq \mathcal{U}_j$. Thus, on the event E, to guarantee $u_{j_\varepsilon} \in S$, it suffices to have

$$n \geq \sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathbb{1}_{\mathcal{U}_j}(X_m).$$

Noting that this is a sum of independent Bernoulli random variables, a Chernoff bound implies that on an event E' of probability at least $1 - 2^{-s}$,

$$\begin{aligned}\sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathbb{1}_{\mathcal{U}_j}(X_m) &\leq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathcal{P}(\mathcal{U}_j) \\ &= s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)(u_j - u_{j-1}) \leq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j.\end{aligned}$$

Thus, for n satisfying the condition in the theorem statement, on the event $E \cap E'$, we have $u_{j_\varepsilon} \in S$, and therefore (as proven above) $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. Finally, a union bound implies that the event $E \cap E'$ has probability at least

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)},$$

as required. □

Proof of Lemma 19. 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 (15). Otherwise, suppose $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$. By the classic symmetrization inequality [e.g., 47, Lemma 2.3.1],

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

where $S \sim P^m$ 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

$$\mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; S, \Xi_{[m]})\right|\right] = \mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; S \cap \mathcal{U}, \Xi_{[|S\cap\mathcal{U}|]})\right| \frac{|S \cap \mathcal{U}|}{m}\right], \tag{45}$$

where $\Xi_{[q]} = \{\xi_1, \ldots, \xi_q\}$ for any $q \in \{0, \ldots, m\}$. By the classic desymmetrization inequality [see e.g., 35], applied under the conditional distribution given $|S \cap \mathcal{U}|$, the right hand side of (45) is at most

$$\mathbb{E}\left[2\phi_\ell(\mathcal{H}; |S \cap \mathcal{U}|, P_{\mathcal{U}}) \frac{|S \cap \mathcal{U}|}{m}\right] + \sup_{h,g \in \mathcal{H}} |\mathrm{R}_\ell(h; P_{\mathcal{U}}) - \mathrm{R}_\ell(g; P_{\mathcal{U}})| \frac{\mathbb{E}\left[\sqrt{|S \cap \mathcal{U}|}\right]}{m}. \tag{46}$$

By Jensen's inequality, the second term in (46) 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 $|S \cap \mathcal{U}|$, the first term in (46) is at most

$$\mathbb{E}\left[2\phi_\ell(\mathcal{H}; |S \cap \mathcal{U}|, P_{\mathcal{U}})\frac{|S \cap \mathcal{U}|}{m}\mathbb{1}\left[|S \cap \mathcal{U}| \geq (1/2)P(\mathcal{U})m\right]\right] \\ + 2\bar{\ell}P(\mathcal{U})\mathbb{P}\left(|S \cap \mathcal{U}| < (1/2)P(\mathcal{U})m\right). \quad (47)$$

Since $|S \cap \mathcal{U}| \geq (1/2)P(\mathcal{U})m \Rightarrow |S \cap \mathcal{U}| \geq \lceil (1/2)P(\mathcal{U})m \rceil$, and $\phi_\ell(\mathcal{H}; q, P_{\mathcal{U}})$ is nonincreasing in q, the first term in (47) is at most

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

while a Chernoff bound implies the second term in (47) is at most

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

Plugging back into (46), we have

$$\phi_\ell(\mathcal{H}; m, P) \leq 4\phi_\ell(\mathcal{H}; \lceil (1/2)P(\mathcal{U})m \rceil, P_{\mathcal{U}})P(\mathcal{U}) + \frac{32\bar{\ell}}{m} + 2\mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{1}{m}}. \quad (48)$$

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 $f^\star_P \in \mathcal{H}$, (15) implies

$$\phi_\ell(\mathcal{H}; \lceil (1/2)P(\mathcal{U})m \rceil, P_{\mathcal{U}}) \leq \mathring{\phi}_\ell\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil (1/2)P(\mathcal{U})m \rceil, P_{\mathcal{U}}\right). \quad (49)$$

Combining (48) with (49), we see that $\mathring{\phi}'_\ell$ satisfies the condition (15) of Definition 16.

Furthermore, by the fact that $\mathring{\phi}_\ell$ satisfies (14) of Definition 16, 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 (14) of Definition 16. 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 20. Let $\mathring{\phi}'_\ell$ be as in Lemma 19, and define for any $m \in \mathbb{N}$, $s \in [1, \infty)$, $\zeta \in [0, \infty]$, and $\mathcal{H} \subseteq [\mathcal{F}]$,

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

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 $m \geq 2\mathcal{P}(\mathcal{U})^{-1}$,

$$\mathring{U}'_\ell(\mathcal{H}, \zeta; \mathcal{P}_{XY}, m, s) \leq 33\mathcal{P}(\mathcal{U})\mathring{U}_\ell(\mathcal{H}, \zeta/\mathcal{P}(\mathcal{U}); \mathcal{P}_{\mathcal{U}}, \lceil (1/2)\mathcal{P}(\mathcal{U})m\rceil, s). \tag{50}$$

In particular, for $j \geq j_\ell$, 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

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

suffices to make the right side of (50) (with $s = \hat{\mathfrak{s}}(2m)$ and $\zeta = 2^{2-j}$) at most 2^{-j-1}; in particular, this means taking u_j equal to $2m \vee u_{j-1} \vee 2u_{j-2}$ for any such m (with $\log_2(m) \in \mathbb{N}$) suffices to satisfy (22) (with the $\mathring{\mathrm{M}}_\ell$ in (22) defined with respect to the $\mathring{\phi}'_\ell$ function); monotonicity of $\zeta \mapsto \mathring{\mathrm{M}}_\ell\left(\zeta, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \hat{\mathfrak{s}}(2m)\right)$ implies (24) is a sufficient condition for this. In the special case where $\mathcal{P}(\mathcal{U}_j) = 0$, $\mathring{U}'_\ell(\mathcal{F}_j, 2^{2-j}; \mathcal{P}_{XY}, m, s) = \tilde{K}\frac{\bar{\ell}s}{m}$, so that taking $u_j \geq \tilde{K}\bar{\ell}\hat{\mathfrak{s}}(u_j)2^{j+2} \vee u_{j-1} \vee 2u_{j-1}$ suffices to satisfy (22) (again, with the $\mathring{\mathrm{M}}_\ell$ in (22) defined in terms of $\mathring{\phi}'_\ell$). Plugging these values into Theorem 18 completes the proof. □

Appendix B: Proofs of Results in Section 5

This appendix includes the remaining details of the proof of Theorem 8, to complete the derivations from Section 7.2, and also presents the remaining essential details for the proof of Theorem 9.

Proof of Theorem 8. Let $\tilde{j}_\varepsilon = \lceil \log_2(1/\Psi_\ell(\varepsilon))\rceil$. For $j_\ell \leq j \leq \tilde{j}_\varepsilon$, define $s_j = \mathrm{Log}\left(\frac{48\left(2+\tilde{j}_\varepsilon-j\right)^2}{\delta}\right)$, and let $u_j = 2^{\lceil \log_2(u'_j)\rceil}$, where

$$u'_j = c'\left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(\mathrm{vc}\left(\mathcal{G}_{\mathcal{F}}\right)\mathrm{Log}\left(\chi_\ell\bar{\ell}\right) + s_j\right), \tag{51}$$

for an appropriate universal constant $c' \in [1, \infty)$. A bit of calculus reveals that for $j_\ell + 2 \leq j \leq \tilde{j}_\varepsilon$, $u'_j \geq u'_{j-1}$ and $u'_j \geq 2u'_{j-2}$, so that $u_j \geq u_{j-1}$ and $u_j \geq 2u_{j-2}$ as well; this is also trivially satisfied for $j \in \{j_\ell, j_\ell + 1\}$ if we take $u_{j-2} = 1$ in these cases (as in Theorem 18). Combining this fact with (34), (18), and (19), we find that, for an appropriate choice of the constant c', these u_j satisfy (22) when we define $\hat{\mathfrak{s}}$ such that, for every $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, $\forall m \in \{2u_{j-1}, \ldots, u_j\}$ with $\log_2(m) \in \mathbb{N}$,

$$\hat{\mathfrak{s}}(m) = \mathrm{Log}\left(\frac{12\log_2\left(4u_j/m\right)^2\left(2 + \tilde{j}_\varepsilon - j\right)^2}{\delta}\right).$$

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

Next, note that, since $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ and u_j is nondecreasing in j,

$$u_{j_\varepsilon} \leq u_{\tilde{j}_\varepsilon} \leq 26c'\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\mathrm{vc}\left(\mathcal{G}_{\mathcal{F}}\right)\mathrm{Log}\left(\chi_\ell\bar{\ell}\right) + \mathrm{Log}(1/\delta)\right),$$

so that, for any $c \geq 26c'$, we have $u \geq u_{i_\varepsilon}$, as required by Theorem 18.

For $\mathcal{U}_j$ as in Theorem 18, note that by Condition 3 and the definition of θ,

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

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

$$\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}\left(\mathcal{U}_j\right) u_j \leq \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \mathcal{P}\left(\mathcal{U}_j\right) u_j$$
$$\lesssim \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} a\theta\Psi_\ell^{-1}\left(2^{2-j}\right)^\alpha \left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(A_1 + \mathrm{Log}\left(2 + \tilde{j}_\varepsilon - j\right)\right). \tag{52}$$

We can change the order of summation in the above expression by letting $i = \tilde{j}_\varepsilon - j$ and summing from 0 to $N = j_\varepsilon - j_\ell$. In particular, since $2^{\tilde{j}_\varepsilon} \leq 2/\Psi_\ell(\varepsilon)$, (52) is at most

$$\sum_{i=0}^{N} a\theta\Psi_\ell^{-1}\left(2^{2-\tilde{j}_\varepsilon}2^i\right)^\alpha \left(\frac{4b2^{i(\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{2\bar{\ell}2^{-i}}{\Psi_\ell(\varepsilon)}\right)\left(A_1 + \mathrm{Log}(i+2)\right). \tag{53}$$

Since $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing on $(0,\infty)$, we have $\Psi_\ell^{-1}\left(2^{2-\tilde{j}_\varepsilon}2^i\right) \leq 2^{i+2}\Psi_\ell^{-1}\left(2^{-\tilde{j}_\varepsilon}\right)$, and since Ψ_ℓ^{-1} is increasing, this latter expression is at most $2^{i+2}\Psi_\ell^{-1}\left(\Psi_\ell(\varepsilon)\right) = 2^{i+2}\varepsilon$. Thus, (53) is at most

$$16a\theta\varepsilon^\alpha \sum_{i=0}^{N}\left(\frac{b2^{i(\alpha+\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}2^{i(\alpha-1)}}{\Psi_\ell(\varepsilon)}\right)\left(A_1 + \mathrm{Log}(i+2)\right). \tag{54}$$

In general, $\mathrm{Log}(i+2) \leq \mathrm{Log}(N+2)$, so that $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)}\left(A_1 + \mathrm{Log}(i+2)\right) \leq (A_1 + \mathrm{Log}(N+2))(N+1)$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)}\left(A_1 + \mathrm{Log}(i+2)\right) \leq (A_1 + \mathrm{Log}(N+2))(N+1)$. When $\alpha+\beta < 2$ holds, 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 furthermore $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \sum_{i=0}^{\infty} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \frac{2}{1-2^{(\alpha+\beta-2)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha+\beta-2)}}\right)$. 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 $\alpha = \beta = 1$), and noting that $\frac{1}{1-2^{(\alpha+\beta-2)}} / \min\left\{\frac{1}{1-2^{(\alpha-1)}}, \frac{1}{1-2^{(\beta-1)}}\right\} \in [1/2, 1]$, we find that (54) 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 (7) has

$$n \geq s + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j,$$

as required by Theorem 18.

Finally, we need to show the success probability from Theorem 18 is at least $1-\delta$, for $\hat{\mathfrak{s}}$ and s as above. Toward this end, note that

$$\begin{aligned}
\sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)} &\leq \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{i=\log_2(u_{j-1})+1}^{\log_2(u_j)} \frac{\delta}{2\left(2+\log_2(u_j)-i\right)^2 \left(2+\tilde{j}_\varepsilon - j\right)^2} \\
&= \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{t=0}^{\log_2(u_j/u_{j-1})-1} \frac{\delta}{2(2+t)^2\left(2+\tilde{j}_\varepsilon - j\right)^2} \\
&< \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \frac{\delta}{2\left(2+\tilde{j}_\varepsilon - j\right)^2} < \sum_{t=0}^{\infty} \frac{\delta}{2(2+t)^2} < \delta/2.
\end{aligned}$$

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

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)} \geq 1-\delta.$$

Therefore, Theorem 18 implies the stated result. $\square$

We note that the values $\hat{\mathfrak{s}}(m)$ used in the proof of Theorem 8 have a direct dependence on the parameters b, β, a, α, and χ_ℓ. 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 $\hat{\mathfrak{s}}(m) = \mathrm{Log}\left(\frac{12\log_2(2m)^2}{\delta}\right)$ instead, which only leads to an increase by a $\log\log$ factor: specifically, replacing the factor of A_1 in (6), and the factors $(A_1 + \mathrm{Log}(B_1))$ and $(A_1 + \mathrm{Log}(C_1))$ in (7), 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 8 without having direct access to the values b, β, a, α, and χ_ℓ in the algorithm.

Proof Sketch of Theorem 9. The proof follows analogously to the proof of Theorem 8, with the exception that now, for each integer j with $j_\ell \leq j \leq \tilde{j}_\varepsilon$, we replace the definition of u'_j from (51) with the following definition. Letting

$$c_j = \mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\left(\bar{\ell}^2/b\right)\left(a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha\right)^\beta\right),$$

define

$$u'_j = c'\left(b2^{j(2-\beta)}\left(a\theta\Psi_\ell^{-1}(2^{2-j})^\alpha\right)^{1-\beta} + \bar{\ell}2^j\right)(c_j + s_j),$$

where $c' \in [1, \infty)$ is an appropriate universal constant, and s_j is as in the proof of Theorem 8. With this substitution in place, the values u_j and s, and function $\hat{\mathfrak{s}}$, are then defined as in the proof of Theorem 8. Since $x \mapsto x\Psi_\ell^{-1}(1/x)$ is nondecreasing, a bit of calculus reveals $u_j \geq u_{j-1}$ and $u_j \geq 2u_{j-2}$. Combined with (35), (19), (18), and Lemma 5, this implies we can choose the constant c' so that these u_j satisfy (24). By an identical argument to that used in Theorem 8, we have

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)} \geq 1 - \delta.$$

It remains only to show that any values of u and n satisfying (8) and (9), respectively, necessarily also satisfy the respective conditions for u and n in Corollary 20.

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

$$u_{j_\varepsilon} \leq u_{\tilde{j}_\varepsilon} \lesssim \left(\frac{b\,(a\theta\varepsilon^\alpha)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) A_2.$$

Thus, for an appropriate choice of c, any u satisfying (8) has $u \geq u_{j_\varepsilon}$, as required by Corollary 20.

Finally, note that for $\mathcal{U}_j$ as in Theorem 18, and $i_j = \tilde{j}_\varepsilon - j$,

$$\begin{aligned}
\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j &\leq \sum_{j=j_\ell}^{j_\varepsilon} a\theta\Psi_\ell^{-1}(2^{2-j})^\alpha u_j \\
&\lesssim \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} b\left(a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha\right)^{2-\beta} \left(A_2 + \mathrm{Log}\,(i_j + 2)\right) \\
&\qquad + \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \bar{\ell} a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha \left(A_2 + \mathrm{Log}\,(i_j + 2)\right).
\end{aligned}$$

By changing the order of summation, now summing over values of i_j from 0 to $N = \tilde{j}_\varepsilon - j_\ell \leq \log_2(4\bar{\ell}/\Psi_\ell(\varepsilon))$, and noting $2^{\tilde{j}_\varepsilon} \leq 2/\Psi_\ell(\varepsilon)$, and $\Psi_\ell^{-1}(2^{-\tilde{j}_\varepsilon} 2^{2+i}) \leq 2^{2+i}\varepsilon$ for $i \geq 0$, this last expression is

$$\begin{aligned}
&\lesssim \sum_{i=0}^{N} b\left(\frac{a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} \left(A_2 + \mathrm{Log}\,(i+2)\right) \qquad (55) \\
&\quad + \sum_{i=0}^{N} \frac{\bar{\ell} a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \left(A_2 + \mathrm{Log}\,(i+2)\right).
\end{aligned}$$

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 have $\sum_{i=0}^{N} 2^{i(\alpha-1)(2-\beta)}(A_2+\mathrm{Log}(i+2))$

$\leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{2}{1-2^{(\alpha-1)(2-\beta)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)(2-\beta)}}\right) + \frac{A_2}{1-2^{(\alpha-1)(2-\beta)}}$, and $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{2}{1-2^{(\alpha-1)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)}}\right) + \frac{A_2}{1-2^{(\alpha-1)}}$. Thus, noting that $\frac{1}{1-2^{(\alpha-1)(2-\beta)}}/\frac{1}{1-2^{(\alpha-1)}} \in [1/2, 1]$, we generally have $\sum_{i=0}^{N} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \lesssim C_1(A_2 + \mathrm{Log}(C_1))$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \lesssim C_1(A_2 + \mathrm{Log}(C_1))$. Plugging this into (55), we find that for an appropriately large numerical constant c, any n satisfying (9) has $n \geq \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j$, as required by Corollary 20. □

We note that, as in Theorem 8, the values $\hat{\mathfrak{s}}$ used to obtain Theorem 9 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 8, we can obtain only slightly worse results by instead taking $\hat{\mathfrak{s}}(m) = \mathrm{Log}\left(\frac{12\log_2(2m)^2}{\delta}\right)$, which again only leads to an increase by a $\log\log$ factor: replacing the factor of A_2 in (8), and the factor of $(A_2 + \mathrm{Log}(C_1))$ in (9), 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 9 is always available, without requiring direct dependence on these quantities.

B.1. Derivations for Section 5.5

For completeness, we include here derivations of quantities appearing in the example given in Section 5.5. We begin with the claim that, for any $\omega \in (0, 1/2]$, (10) is satisfied in Condition 10 with the values $q = \frac{7}{\omega}$ and $\rho = \frac{1}{3} + \omega$. Specifically, for a given $\varepsilon > 0$, let $i_\varepsilon = \left\lceil \frac{3}{\varepsilon^{2/3}} \right\rceil$, and let $\mathcal{G}_\varepsilon$ be the set of functions g in $\mathcal{G}^*$ with $g(x,y) \in \{j\varepsilon/\sqrt{2} : j \in \{0, \ldots, \lceil 4\sqrt{2}/\varepsilon \rceil - 1\}\}$ for each $x \in \{x_i : 1 \leq i \leq i_\varepsilon\}$ and $y \in \mathcal{Y}$, and $g(x,y) = 0$ for every $x \in \mathcal{X} \setminus \{x_i : 1 \leq i \leq i_\varepsilon\}$ and $y \in \mathcal{Y}$. For each $g \in \mathcal{G}_\varepsilon$, let g' be the function in $\mathcal{G}^*$ with $g'(x,y) = g(x,y) + \varepsilon/\sqrt{2}$ for each $x \in \{x_i : 1 \leq i \leq i_\varepsilon\}$ and $y \in \mathcal{Y}$, and $g'(x,y) = 4$ for each $x \in \mathcal{X} \setminus \{x_i : 1 \leq i \leq i_\varepsilon\}$ and $y \in \mathcal{Y}$. Note that $\bigcup_{g \in \mathcal{G}_\varepsilon}[g, g']$ contains all functions g in $\mathcal{G}^*$ having $0 \leq g(x,y) \leq 4$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$; in particular, this implies it contains $\mathcal{G}_{\mathcal{F}}$. Furthermore, for each $g \in \mathcal{G}_\varepsilon$, $\|g - g'\|^2_{\mathcal{P}_{XY}} = \sum_{i=1}^{i_\varepsilon} \frac{\varepsilon^2}{2}\mathcal{P}(\{x_i\}) + \sum_{i=i_\varepsilon+1}^{\infty} 16\mathcal{P}(\{x_i\}) \leq \frac{\varepsilon^2}{2} + \frac{16 \cdot 90}{\pi^4}\int_{i_\varepsilon}^{\infty}\frac{1}{x^4}\mathrm{d}x = \frac{\varepsilon^2}{2} + \frac{16 \cdot 30}{\pi^4}\frac{1}{i_\varepsilon^3} \leq \frac{\varepsilon^2}{2} + \frac{16 \cdot 30}{27\pi^4}\varepsilon^2 < \varepsilon^2$, so that $[g, g']$ is an ε-bracket under $L_2(\mathcal{P}_{XY})$. Therefore, $\mathcal{N}_{[]}\left(\varepsilon, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})\right) \leq |\mathcal{G}_\varepsilon| = \lceil 4\sqrt{2}/\varepsilon \rceil^{2i_\varepsilon}$, so that (taking $F = \bar{\ell} = 4$, constant, in Condition 10) $\ln \mathcal{N}_{[]}\left(4\varepsilon, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})\right) \leq 2\left\lceil \frac{3}{(4\varepsilon)^{2/3}} \right\rceil \ln\left(\left\lceil \frac{\sqrt{2}}{\varepsilon} \right\rceil\right)$. Since $\ln(x) \leq t x^{1/t}$ for any $x, t \geq 1$, this is at most $\frac{7}{\omega}\varepsilon^{-2\left(\frac{1}{3}+\omega\right)}$ when $\varepsilon \in (0,1)$, for any value $\omega \in (0, 1/2]$. This is trivially also an upper bound on $\ln \mathcal{N}_{[]}\left(4\varepsilon, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})\right)$ for all $\varepsilon \geq 1$ (since $\mathcal{N}_{[]}\left(4\varepsilon, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})\right) = 1$ in that case). Thus, (10) is satisfied with $q = \frac{7}{\omega}$ and $\rho = \frac{1}{3} + \omega$, for any choice of $\omega \in (0, 1/2]$, as claimed.

Next, we present a proof of the claimed $\Omega(\varepsilon^{-4/3})$ lower bound on the sample size required to obtain an ε bound on the minimax expected excess error rate of passive learning methods in the example scenario. We approach this with the classic technique of Assouad (see e.g., [46]). Specifically, fix any $\varepsilon \in (0, (1-2\nu_0)/64)$, and fix a sample size $m \in \mathbb{N}$ with $m \leq 2^{-13}(1-2\nu_0)^{1/3}\varepsilon^{-4/3}$. Let $j_0 = \left\lfloor \left(\frac{72}{107\pi^4}\right)^{1/4} \left(\frac{1-2\nu_0}{\varepsilon}\right)^{1/3} \right\rfloor$, $j_1 = \left\lfloor \frac{1}{2^{4/3}} \left(\frac{1-2\nu_0}{\varepsilon}\right)^{1/3} \right\rfloor$, and $k = j_1 - j_0 + 1$. In particular, a simple calculation reveals $k \geq \frac{27}{250}\left(\frac{1-2\nu_0}{\varepsilon}\right)^{1/3}$. Now for any binary vector $v = (v_1, \ldots, v_k) \in \{0,1\}^k$, define P_v as the probability measure on $\mathcal{X} \times \mathcal{Y}$ with marginal $\mathcal{P}$ on $\mathcal{X}$ (as specified in the construction), $\eta(x_i; P_v) = 1$ for $i \in \mathbb{N} \setminus \{j_0, \ldots, j_1\}$, and $\eta(x_i; P_v) = \nu_0 + (1-2\nu_0)v_{i-j_0+1}$ for $i \in \{j_0, \ldots, j_1\}$. Then note that for any $v, v' \in \{0,1\}^k$ with $\|v - v'\|_1 = 1$, the total variation distance $\|P_v - P_{v'}\|$ between the corresponding distributions is at most $\frac{90}{\pi^4 j_0^4}(1-2\nu_0)$. This further implies $\|P_v^m - P_{v'}^m\| \leq m\|P_v - P_{v'}\| \leq 2^{-13}\frac{90}{\pi^4 j_0^4}\left(\frac{1-2\nu_0}{\varepsilon}\right)^{4/3} < \frac{1}{2}$. Therefore, Theorem 2.12(ii) of [46] implies that, for any estimator $\hat{v} : (\mathcal{X} \times \mathcal{Y})^m \to \{0,1\}^k$ (possibly randomized), there exists a choice $v \in \{0,1\}^k$ such that, defining $\mathcal{P}_{XY} = P_v$, we have $\mathbb{E}\left[\|\hat{v}(\mathcal{Z}_m) - v\|_1\right] \geq \frac{k}{4} \geq \frac{27}{1000}\left(\frac{1-2\nu_0}{\varepsilon}\right)^{1/3}$. In particular, for any passive learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{F}^*$, we can define a vector $\hat{v}$ based on the returned function $\hat{f}$ from $\mathcal{A}$ by letting $\hat{v}_i = (\mathrm{sign}(\hat{f}(x_{i+j_0-1})) + 1)/2$ for each $i \in \{1, \ldots, k\}$. Then we note that for any $v \in \{0,1\}^k$, if $\mathcal{P}_{XY} = P_v$, then $\mathrm{er}(\hat{f}) - \mathrm{er}(f^\star) \geq \frac{90}{\pi^4 j_1^4}(1-2\nu_0)\|\hat{v} - v\|_1$. Thus, there exists a choice of $v \in \{0,1\}^k$ such that, defining $\mathcal{P}_{XY} = P_v$, we have that for $\hat{f} = \mathcal{A}(\mathcal{Z}_m)$, $\mathbb{E}\left[\mathrm{er}(\hat{f}) - \mathrm{er}(f^\star)\right] \geq \frac{90}{\pi^4 j_1^4}(1-2\nu_0) \cdot \frac{27}{1000}\left(\frac{1-2\nu_0}{\varepsilon}\right)^{1/3} > \varepsilon$. Thus, since these P_v distributions satisfy the description of the construction in Section 5.5, we see that to guarantee expected excess error rate at most ε for all $\mathcal{P}_{XY}$ fitting the description in the construction, any passive learning method would require the sample size m for its input labeled data set to be greater than $2^{-13}(1-2\nu_0)^{1/3}\varepsilon^{-4/3} = \Omega(\varepsilon^{-4/3})$, as claimed. In particular, this agrees with the dependence on ε derived for ERM_ℓ in Section 5.5 (up to a logarithmic factor). In contrast, the analysis of Algorithm 1 in Section 5.5 reveals that (by choosing $\delta = \varepsilon/2$), Algorithm 1 can achieve $\mathbb{E}[\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star)] \leq \varepsilon$ for all such $\mathcal{P}_{XY}$ with a number of label requests n having only $O(\varepsilon^{-7/12}\mathrm{Log}(1/\varepsilon))$ dependence on ε, a significant decrease compared to the $\Omega(\varepsilon^{-4/3})$ lower bound we have just established for all passive learning methods.

B.2. Derivations for Section 5.6

For completeness, we include here a derivation of the parameters a and α for which the distributions $\mathcal{P}_{XY}$ in the example in Section 5.6 satisfy Condition 3. Specifically, as in Section 5.6, let ℓ be the quadratic loss, fix an integer $k \geq 5$, suppose $\mathcal{P}$ is uniform on $\{x \in \mathbb{R}^k : \|x\| = 1\}$, and suppose $\mathcal{P}_{XY}$ is such that $f^\star(x) = w^* \cdot x$ for some $w^* \in \mathbb{R}^k$ with $\|w^*\| = 1$. In particular, for this choice of ℓ, this implies $\eta(x) = (w^* \cdot x + 1)/2$. For any $f \in \mathcal{F}^*$, $\mathrm{er}(f) - \mathrm{er}(f^\star) = \mathbb{E}\left[|1-2\eta(X)| \middle| X \in \mathrm{DIS}(\{f, f^\star\})\right]\Delta(f, f^\star)$, for $X \sim \mathcal{P}$. Therefore, among functions $f \in \mathcal{F}^*$ with a given value p of $\Delta(f, f^\star)$, the functions with minimal $\mathrm{er}(f) - \mathrm{er}(f^\star)$ are those that minimize $\mathbb{E}\left[|2\eta(X) - 1| \middle| X \in \mathrm{DIS}(\{f, f^\star\})\right]$ subject to $\mathcal{P}(\mathrm{DIS}(\{f, f^\star\})) = p$; since $|2\eta(x) - 1| = |w^* \cdot x|$ is increasing in $|w^* \cdot x|$ and $t \mapsto \mathcal{P}(x : |w^* \cdot x| \leq t)$ is continuous, any $f \in \mathcal{F}^*$ of minimal $\mathrm{er}(f) - \mathrm{er}(f^\star)$ subject to $\Delta(f, f^\star) = p$ has $\mathrm{DIS}(\{f, f^\star\}) = \{x : |w^* \cdot x| \leq \gamma_p\}$ (up to probability zero differences) for some $\gamma_p \in [0,1]$ chosen so that $\mathcal{P}(x : |w^* \cdot x| \leq$

$\gamma_p) = p$; in particular, the minimum value of $\mathrm{er}(f) - \mathrm{er}(f^\star)$ among such functions f is $\mathbb{E}[|w^* \cdot X|\mathbb{1}[|w^* \cdot X| \leq \gamma_p]]$. Fix such a function f_p with $\mathrm{DIS}(\{f_p, f^\star\}) = \{x : |w^* \cdot x| \leq \gamma_p\}$.

For $X \sim \mathcal{P}$, one can show that the $[0,1]$-valued random variable $|w^* \cdot X|$ has density function $g(t) = \frac{2\Gamma(k/2)}{\sqrt{\pi}\Gamma((k-1)/2)}(1-t^2)^{\frac{k-3}{2}}$, where Γ is the usual gamma function (see [37] for a derivation of the CDF, from which this g can be derived). Thus,

$$\begin{aligned}\mathbb{E}[|w^* \cdot X|\mathbb{1}[|w^* \cdot X| \leq \gamma_p]] &= \int_0^{\gamma_p} \frac{2\Gamma(k/2)}{\sqrt{\pi}\Gamma((k-1)/2)} t(1-t^2)^{\frac{k-3}{2}} \mathrm{d}t \\ &= \frac{2\Gamma(k/2)}{\sqrt{\pi}\Gamma((k-1)/2)} \frac{1}{k-1} \left(1 - (1-\gamma_p^2)^{\frac{k-1}{2}}\right).\end{aligned}$$

When $\gamma_p \leq \frac{1}{\sqrt{k-3}}$, some basic calculus reveals $1 - (1-\gamma_p^2)^{\frac{k-1}{2}} \geq \gamma_p^2 \frac{k-1}{2e}$. Since one can also verify that $\frac{2\Gamma(k/2)}{\sqrt{\pi}\Gamma((k-1)/2)} \geq \sqrt{k/3}$, we have that if p is such that $\gamma_p \leq \frac{1}{\sqrt{k-3}}$, then $\mathrm{er}(f_p) - \mathrm{er}(f^\star) \geq \frac{\sqrt{k}\gamma_p^2}{2e\sqrt{3}}$. It also holds that $\Delta(f_p, f^\star) = \mathcal{P}(x : |w^* \cdot x| \leq \gamma_p) \leq \sqrt{k}\gamma_p$ [see e.g., 22]. Together, we have that if $\gamma_p \leq \frac{1}{\sqrt{k-3}}$, then $\Delta(f_p, f^\star) \leq \sqrt{k}\gamma_p = \sqrt{2e}(3k)^{1/4}\left(\frac{\sqrt{k}\gamma_p^2}{2e\sqrt{3}}\right)^{1/2} \leq \sqrt{2e}(3k)^{1/4}\left(\mathrm{er}(f_p) - \mathrm{er}(f^\star)\right)^{1/2}$.

Noting that γ_p is continuous in p, with $\gamma_0 = 0$ and $\gamma_1 = 1$, the intermediate value theorem implies $\exists p_* \in [0,1]$ with $\gamma_{p_*} = \frac{1}{\sqrt{k-3}}$. Since $\sqrt{2e}(3k)^{1/4}\left(\frac{\sqrt{k}\gamma_{p_*}^2}{2e\sqrt{3}}\right)^{1/2} = \sqrt{\frac{k}{k-3}} > 1$, we have $\sqrt{2e}(3k)^{1/4}\left(\mathrm{er}(f_{p_*}) - \mathrm{er}(f^\star)\right)^{1/2} > 1$. Now for any p with $\gamma_p > \frac{1}{\sqrt{k-3}}$, we have $\mathrm{DIS}(\{f_p, f^\star\}) \supseteq \mathrm{DIS}(\{f_{p_*}, f^\star\})$, which implies $\mathrm{er}(f_p) \geq \mathrm{er}(f_{p_*})$. Therefore, $\sqrt{2e}(3k)^{1/4}\left(\mathrm{er}(f_p) - \mathrm{er}(f^\star)\right)^{1/2} > 1 \geq \Delta(f_p, f^\star)$. Thus, we have established that $\Delta(f_p, f^\star) \leq \sqrt{2e}(3k)^{1/4}\left(\mathrm{er}(f_p) - \mathrm{er}(f^\star)\right)^{1/2}$ for *every* $p \in [0,1]$. Since, for every $p \in [0,1]$, f_p was chosen to minimize $\mathrm{er}(f_p) - \mathrm{er}(f^\star)$ subject to $\Delta(f_p, f^\star) = p$, we have $\Delta(f, f^\star) \leq \sqrt{2e}(3k)^{1/4}\left(\mathrm{er}(f) - \mathrm{er}(f^\star)\right)^{1/2}$ for *every* $f \in \mathcal{F}^*$: that is, that Condition 3 holds with $a = \sqrt{2e}(3k)^{1/4}$ and $\alpha = 1/2$.

Appendix C: Remarks on VC Major and VC Hull Classes

In addition to VC Subgraph classes, and scenarios satisfying general entropy conditions, 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 [20], 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 $\varepsilon > 0$,

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

This implies that for $\mathcal{F}$ a VC major class, and ℓ classification-calibrated and either nonincreasing or Lipschitz on $[-\sup_{h\in\mathcal{F}}\sup_{x\in\mathcal{X}}|h(x)|, \sup_{h\in\mathcal{F}}\sup_{x\in\mathcal{X}}|h(x)|]$, if $f^\star \in \mathcal{F}$ and $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, then the conditions of Theorem 18 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 2, with β as in Lemma 5, the conditions of Corollary 20 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 $n = \tilde{O}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta/2}\right)$. When θ is small, these values of n (and indeed u) compare favorably to the value of $m = \tilde{O}\left(\Psi_\ell(\varepsilon)^{\beta/2-2}\right)$, derived analogously from Theorem 17, sufficient for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ to achieve the same [see 20].

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 $\theta(0) \le 2$ 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

$$\mathcal{F} = B\mathrm{conv}(\mathcal{C}) = \left\{ x \mapsto B\sum_j \lambda_j h_j(x) : \sum_j |\lambda_j| \le 1, h_j \in \mathcal{C} \right\}$$

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, using established bounds on the uniform covering numbers of VC hull classes [see 47, Corollary 2.6.12], and noting that for any VC hull class $\mathcal{F}$ with envelope function F, and any $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{F}_\mathcal{U}$ is also a VC hull class, with envelope function $\mathrm{F}\mathbb{1}_\mathcal{U}$. Specifically, one can use these observations to derive the following results. For a VC hull class $\mathcal{F} = B\mathrm{conv}(\mathcal{C})$, if ℓ is classification-calibrated and Lipschitz on $[-\sup_{h\in\mathcal{F}}\sup_{x\in\mathcal{X}}|h(x)|, \sup_{h\in\mathcal{F}}\sup_{x\in\mathcal{X}}|h(x)|]$, $f^\star \in \mathcal{F}$, and $\mathcal{P}_{XY}$ satisfies Condition 3 and Condition 4, then letting $d = 2\mathrm{vc}(\mathcal{C})$, the conditions of Theorem 18 can be satisfied with the probability bound having value 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 2, with β as in Lemma 5, the conditions of Corollary 20 can be satisfied with the probability 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 $n = \tilde{O}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\frac{2\beta}{d+2}}\right)$. Compare these to the value $m = \tilde{O}\left(\Psi_\ell(\varepsilon)^{\frac{2\beta}{d+2}-2}\right)$, derived analogously from Theorem 17, sufficient for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ to achieve the same general guarantee [see also 6, 10]. However, it is not clear whether these results for active learning with VC hull classes have any practical implications, since we do not know of any scenarios where this sufficient value of m reflects a *tight* analysis of $\mathrm{ERM}_\ell(\mathcal{F},\cdot)$ while simultaneously being significantly larger than either of the above sufficient n values.

Appendix D: Computationally Efficient Updates

As mentioned in Section 6.3, though convenient in the sense that it offers a completely abstract and unified approach, the choice of $\hat{T}_\ell(V;Q,m)$ given by (21) may often make Algorithm 1 computationally inefficient. However, for each of the applications studied in this work, we can relax this $\hat{T}_\ell$ function to a computationally-accessible value, which will then allow the algorithm to be efficient under convexity conditions on the loss and class of functions.

In particular, in the application to VC Subgraph classes, Theorem 8 remains valid if we instead define $\hat{T}_\ell$ as follows. If we let $V^{(m)}$ and Q_m denote the sets V and Q upon reaching Step 5 for any given value of m with $\log_2(m) \in \mathbb{N}$ realized in Algorithm 1, then consider defining $\hat{T}_\ell$ in Step 6 inductively by letting

$$\hat{\gamma}_{m/2} = \frac{8(|Q_{m/2}| \vee 1)}{m}\left(\hat{T}_\ell(V^{(m/2)};Q_{m/2},m/2) \wedge \bar{\ell}\right)$$

(or $\hat{\gamma}_{m/2} = \bar{\ell}$ if $m = 2$), and taking (with a slight abuse of notation to allow $\hat{T}_\ell$ to depend on sets $V^{(m')}$ and $Q_{m'}$ with $m' < m$)

$$\begin{aligned}&\hat{T}_\ell(V^{(m)};Q_m,m) = \\ &\quad c_0\frac{m/2}{|Q_m| \vee 1}\left(\sqrt{\hat{\gamma}_{m/2}^{\beta}\frac{b}{m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_m| + \hat{\mathfrak{s}}(m))}{mb\hat{\gamma}_{m/2}^{\beta}}\right) + \hat{\mathfrak{s}}(m)\right)}\right.\\ &\qquad\qquad \left. + \frac{\bar{\ell}}{m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_m| + \hat{\mathfrak{s}}(m))}{mb\hat{\gamma}_{m/2}^{\beta}}\right) + \hat{\mathfrak{s}}(m)\right)\right), \qquad (56)\end{aligned}$$

for an appropriate universal constant c_0. This value is essentially derived by bounding $\frac{m/2}{|Q|\vee 1}\tilde{U}_\ell(V_{\mathrm{DIS}(V)};\mathcal{P}_{XY},m/2,\hat{\mathfrak{s}}(m))$ (which is a bound on (21) by Lemma 15), based on (30) and Condition 4 (and a Chernoff bound to argue $|Q_m| \approx \mathcal{P}(\mathrm{DIS}(V))m/2)$; since the sample sizes derived for u and n in Theorem 8 are based on these relaxations anyway, they remain sufficient (with slight changes to the constant factors) for these relaxed $\hat{T}_\ell$ values. We include a more detailed proof that these values of $\hat{T}_\ell$ suffice to achieve Theorem 8 in Appendix E.1. Note that we have introduced a dependence on b and β in (56). These values would indeed be available for some applications, such as when they are derived from Lemma 5 when Condition 2 is satisfied; however, in other cases, there may be more-favorable values of b and β than given by Lemma 5, dependent on the specific $\mathcal{P}_{XY}$ distribution, and in these cases direct observation of these values might not be available. Thus, there remains an interesting open question of whether there exists a function $\hat{T}_\ell(V;Q,m)$, which is efficiently computable (under convexity assumptions) and yet preserves the validity of Theorem 8.

In the special case where Condition 2 is satisfied, it is also possible to define a value for $\hat{T}_\ell$ that is computationally accessible, and preserves the validity of Theorem 9. Specifically, consider instead defining $\hat{T}_\ell$ in Step 6 as

$$\hat{T}_\ell(V;Q,m) \\ = \bar{\ell} \wedge c_0 \max\left\{ \begin{array}{l} \left(\frac{b}{|Q|\vee 1}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{|Q|}{b\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^{\frac{\beta}{2-\beta}}\right)+\hat{\mathfrak{s}}(m)\right)\right)^{\frac{1}{2-\beta}} \\ \frac{\bar{\ell}}{|Q|\vee 1}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{|Q|}{\bar{\ell}\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^{\beta}\right)+\hat{\mathfrak{s}}(m)\right) \end{array}\right. , \tag{57}$$

for b and β as in Lemma 5, and for an appropriate universal constant c_0. This value is essentially derived (following 34) by using Lemma 15 under the conditional distribution $\mathcal{P}_{\mathrm{DIS}(V)}$, in conjunction with a localization technique similar to that employed in the derivation of Theorem 17. Appendix E.2 includes a proof that the conclusions of Theorem 9 remain valid for this specification of $\hat{T}_\ell$ in place of (21). That these conclusions remain valid for this bound on excess conditional risks should not be too surprising, since Theorem 9 is itself proven by considering concentration under the conditional distributions $\mathcal{P}_{\mathcal{U}_j}$ via Corollary 20. Note that, unlike the analogous result for Theorem 8 based on (56) above, in this case all of the quantities in $\hat{T}_\ell(V;Q,m)$ are directly observable (in particular, b and β), aside from any possible dependence arising in the specification of $\hat{\mathfrak{s}}$.

It is also possible to define computationally tractable values of $\hat{T}_\ell(V;Q,m)$ in scenarios satisfying the entropy conditions (Condition 10), while preserving the validity of Theorem 12. This substitution can be derived analogously to (56) above, this time leading to the definition

$$\hat{T}_\ell\left(V^{(m)};Q_m,m\right) = \\ c_0\frac{m/2}{|Q_m|\vee 1}\left(\max\left\{\frac{\sqrt{q}\|\mathrm{F}\|^{\rho}_{\mathcal{P}_{XY}}\left(b\hat{\gamma}^{\beta}_{m/2}\right)^{\frac{1-\rho}{2}}}{(1-\rho)m^{1/2}}, \frac{\bar{\ell}^{\frac{1-\rho}{1+\rho}}q^{\frac{1}{1+\rho}}\|\mathrm{F}\|^{\frac{2\rho}{1+\rho}}_{\mathcal{P}_{XY}}}{(1-\rho)^{\frac{2}{1+\rho}}m^{\frac{1}{1+\rho}}}\right\} \right. \\ \left. +\sqrt{b\hat{\gamma}^{\beta}_{m/2}\frac{\hat{\mathfrak{s}}(m)}{m}}+\frac{\bar{\ell}\hat{\mathfrak{s}}(m)}{m}\right), \tag{58}$$

where $\hat{\gamma}_{m/2}$ is defined (inductively) as above, and c_0 is an appropriately large universal constant. By essentially the same argument used for (56) (see Appendix E.1), one can show that using (58) in place of (21) preserves the validity of Theorem 12; for brevity, the details are omitted.

In the case that Condition 2 and (11) are satisfied, it is possible to define a computationally accessible quantity $\hat{T}_\ell(V;Q,m)$, while preserving the validity of Theorem 13. Specifically, following the same reasoning used to arrive at (57), except using (36) instead of (30), we find that while replacing (21) with the definition

$$\hat{T}_\ell\left(V;Q,m\right)= \bar{\ell}\wedge c_0\left(\max\left\{\left(\frac{q\bar{\ell}^{2\rho}b^{1-\rho}}{(1-\rho)^2(|Q|\vee 1)}\right)^{\frac{1}{2-\beta(1-\rho)}},\frac{\bar{\ell}q^{\frac{1}{1+\rho}}}{(1-\rho)^{\frac{2}{1+\rho}}(|Q|\vee 1)^{\frac{1}{1+\rho}}}\right\} + \left(\frac{b\hat{\mathfrak{s}}(m)}{|Q|\vee 1}\right)^{\frac{1}{2-\beta}}+\frac{\bar{\ell}\hat{\mathfrak{s}}(m)}{|Q|\vee 1}\right), \quad (59)$$

for b and β as in Lemma 5 and for an appropriate universal constant c_0, the conclusions of Theorem 13 remain valid. The proof follows similarly to the proof (in Appendix E.2) that (57) preserves the validity of Theorem 9, and is omitted for brevity.

Finally, in the case that Condition 2 and (10) are satisfied, we can again derive an efficiently computable value of $\hat{T}_\ell(V;Q,m)$, which in this case preserves the validity of Theorem 14. Specifically, noting that the reasoning preceding Theorem 14 also implies $\ln\mathcal{N}_{[]}\left(\varepsilon\bar{\ell},\mathcal{G}_V,L_2(\mathcal{P}_{\mathrm{DIS}(V)})\right)\leq q\mathcal{P}(\mathrm{DIS}(V))^{-\rho}\varepsilon^{-2\rho}$, and following the reasoning leading to (59) while replacing q with $q\mathcal{P}(\mathrm{DIS}(V))^{-\rho}$, combined with a Chernoff bound to argue $\mathcal{P}(\mathrm{DIS}(V))\approx 2|Q|/m$ in the algorithm, we find that Theorem 14 remains valid after replacing (21) with the definition

$$\hat{T}_\ell(V;Q,m)= \bar{\ell}\wedge c_0\left(\max\left\{\left(\frac{qm^{\rho}\bar{\ell}^{2\rho}b^{1-\rho}}{(1-\rho)^2(|Q|\vee 1)^{1+\rho}}\right)^{\frac{1}{2-\beta(1-\rho)}},\frac{\bar{\ell}q^{\frac{1}{1+\rho}}m^{\frac{\rho}{1+\rho}}}{(1-\rho)^{\frac{2}{1+\rho}}(|Q|\vee 1)}\right\} + \left(\frac{b\hat{\mathfrak{s}}(m)}{|Q|\vee 1}\right)^{\frac{1}{2-\beta}}+\frac{\bar{\ell}\hat{\mathfrak{s}}(m)}{|Q|\vee 1}\right),$$

for an appropriate universal constant c_0, and where b and β are as in Lemma 5. The proof is essentially similar to that given for (57) in Appendix E.2, and is omitted for brevity.

Appendix E: Proofs for Efficiently Computable Updates

Here we include more detailed proofs of the arguments leading to computationally efficient variants of Algorithm 1, for which the specific results proven in this work for the given applications remain valid. Specifically, we focus on the application to VC Subgraph classes here; the applications to scenarios satisfying the entropy conditions follow analogously. Throughout this section, we adopt the notational conventions introduced in the proof of Theorem 18 (e.g., $V^{(m)}$, $\tilde{V}^{(m)}$, Q_m, $\mathcal{L}_m$, S), except in each instance here these are defined in the context of applying Algorithm 1 with the respective stated variant of $\hat{T}_\ell$.

E.1. *Proof of Theorem 8 under* (56)

We begin by showing that if we specify $\hat{T}_\ell(V; Q, m)$ as in (56), the conclusions of Theorem 8 remain valid. Fix any $\hat{\mathfrak{s}}$ function (to be specified below), and fix any value of $\varepsilon \in (0,1)$. First note that, for any m with $\log_2(m) \in \mathbb{N}$, by a Chernoff bound and the law of total probability, on an event E''_m of probability at least $1 - 2^{1-\hat{\mathfrak{s}}(m)}$, if $m \in S$, then

$$(1/2)m\mathcal{P}(D_m) - \sqrt{\hat{\mathfrak{s}}(m)m\mathcal{P}(D_m)} \leq |Q_m| \leq \hat{\mathfrak{s}}(m) + em\mathcal{P}(D_m). \tag{60}$$

Also recall that, for any m with $\log_2(m) \in \mathbb{N}$, by Lemma 15 and the law of total probability, on an event E_m of probability at least $1 - 6e^{-\hat{\mathfrak{s}}(m)}$, if $m \in S$ and $f^\star \in V^{(m)}$, then

$$\begin{aligned}
&(|Q_m| \vee 1)\left(\mathrm{R}_\ell(f^\star; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m)\right) \\
&\qquad = \frac{m}{2}\left(\mathrm{R}_\ell(f^\star; \mathcal{L}_m) - \inf_{g_{D_m} \in V^{(m)}_{D_m}} \mathrm{R}_\ell(g_{D_m}; \mathcal{L}_m)\right) \\
&\qquad\qquad < \frac{m}{2}\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right)
\end{aligned} \tag{61}$$

and $\forall h \in \tilde{V}^{(m)}$,

$$\begin{aligned}
&\frac{m}{2}\left(\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(f^\star)\right) \\
&< \frac{m}{2}\left(\mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_\ell(f^\star; \mathcal{L}_m) + \tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right) \\
&= |Q_m|\left(\mathrm{R}_\ell(h; Q_m) - \mathrm{R}_\ell(f^\star; Q_m)\right) + \frac{m}{2}\left(\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right) \\
&\leq (|Q_m| \vee 1)\hat{T}_\ell\left(V^{(m)}; Q_m, m\right) + \frac{m}{2}\left(\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right).
\end{aligned} \tag{62}$$

Fix a value $i_\varepsilon \in \mathbb{N}$ (an appropriate value for which will be determined below), and let $\chi_\ell = \chi_\ell(\Psi_\ell(\varepsilon))$. For $m \in \mathbb{N}$ with $\log_2(m) \in \mathbb{N}$, let

$$\begin{aligned}
\tilde{T}_\ell(m) = c_2\left(\frac{b}{m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell\bar{\ell}) + \hat{\mathfrak{s}}(m)\right)\right)^{\frac{1}{2-\beta}} \\
+ c_2\frac{\bar{\ell}}{m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell\bar{\ell}) + \hat{\mathfrak{s}}(m)\right),
\end{aligned}$$

for an appropriate universal constant $c_2 \in [1,\infty)$ (to be determined below); for completeness, also define $\tilde{T}_\ell(1) = \bar{\ell}$. We will now prove by induction that, for an appropriate value of the constant c_0 in (56), for any m' with $\log_2(m') \in \{1, \ldots, i_\varepsilon\}$, on the event $\bigcap_{i=1}^{\log_2(m')-1} E_{2^i} \cap E''_{2^{i+1}}$, if $m' \in S$, then $f^\star \in V^{(m')}$,

$$V^{(m')}_{D_{m'}} \subseteq [\mathcal{F}](\hat{\gamma}_{m'/2}; \ell) \subseteq [\mathcal{F}](2\tilde{T}_\ell(m'/2) \vee \Psi_\ell(\varepsilon); \ell),$$

$$V^{(m')} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_{m'/2}); {\scriptstyle 01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m'/2) \vee \Psi_\ell(\varepsilon)); {\scriptstyle 01}),$$

$$\tilde{U}_\ell\left(V^{(m')}_{D_{m'}}; \mathcal{P}_{XY}, m'/2, \hat{\mathfrak{s}}(m')\right) \wedge \bar{\ell} \leq \frac{|Q_{m'}| \vee 1}{m'/2}\left(\hat{T}_\ell\left(V^{(m')}; Q_{m'}, m'\right) \wedge \bar{\ell}\right),$$

and if $\hat{\gamma}_{m'/2} \geq \Psi_\ell(\varepsilon)$,

$$\frac{|Q_{m'}| \vee 1}{m'/2}\left(\hat{T}_\ell\left(V^{(m')}; Q_{m'}, m'\right) \wedge \bar{\ell}\right) \leq \tilde{T}_\ell(m').$$

As a base case for this inductive argument, we note that for $m' = 2$, we have (by definition) $\hat{\gamma}_{m'/2} = \bar{\ell}$, and furthermore (if $c_0 \wedge c_2 \geq 2$) $\hat{T}_\ell(V^{(2)}; Q_2, 2) \geq \bar{\ell}$ and $\tilde{T}_\ell(1) \geq \bar{\ell}$, so that the claimed inclusions and inequalities trivially hold. Now, for the inductive step, take as an inductive hypothesis that the claim is satisfied for $m' = m$ for some $m \in \mathbb{N}$ with $\log_2(m) \in \{1, \ldots, i_\varepsilon - 1\}$. Suppose the event $\bigcap_{i=1}^{\log_2(m)} E_{2^i} \cap E''_{2^{i+1}}$ occurs, and that $2m \in S$. By the inductive hypothesis, combined with (61) and the fact that $(|Q_m| \vee 1)\mathrm{R}_\ell(f^\star; Q_m) \leq (m/2)\bar{\ell}$, we have

$$\begin{aligned}(|Q_m| \vee 1)&\left(\mathrm{R}_\ell(f^\star; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m)\right) \\ &\leq \frac{m}{2}\left(\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right) \leq (|Q_m| \vee 1)\hat{T}_\ell\left(V^{(m)}; Q_m, m\right).\end{aligned}$$

Therefore, $f^\star \in \tilde{V}^{(m)}$ as well, which implies $f^\star \in V^{(2m)} = \tilde{V}^{(m)}$. Furthermore, by (62), the inductive hypothesis, and the definition of $\tilde{V}^{(m)}$ from Step 6, $\forall h \in V^{(2m)} = \tilde{V}^{(m)}$,

$$\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(f^\star) < 2\frac{|Q_m| \vee 1}{m/2}\left(\hat{T}_\ell\left(V^{(m)}; Q_m, m\right) \wedge \bar{\ell}\right),$$

and if $\hat{\gamma}_{m/2} \geq \Psi_\ell(\varepsilon)$, then this is at most $2\tilde{T}_\ell(m)$.

Since $\hat{\gamma}_m = 2\frac{|Q_m| \vee 1}{m/2}\left(\hat{T}_\ell\left(V^{(m)}; Q_m, m\right) \wedge \bar{\ell}\right)$, and $\mathrm{R}_\ell(h_{D_{2m}}) \leq \mathrm{R}_\ell(h_{D_m})$ for every $h \in V^{(2m)}$, we have $V^{(2m)}_{D_{2m}} \subseteq [\mathcal{F}](\hat{\gamma}_m; \ell) \subseteq [\mathcal{F}](2\tilde{T}_\ell(m) \vee \Psi_\ell(\varepsilon); \ell)$. By definition of $\mathcal{E}_\ell(\cdot)$, we also have $\mathrm{er}(h_{D_{2m}}) - \mathrm{er}(f^\star) \leq \mathcal{E}_\ell(\hat{\gamma}_m)$ for every $h \in V^{(2m)}$; since $f^\star \in V^{(2m)}$, we have $\mathrm{sign}(h_{D_{2m}}) = \mathrm{sign}(h)$, so that $\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \mathcal{E}_\ell(\hat{\gamma}_m)$ as well: that is, $V^{(2m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m); {\scriptstyle 01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m) \vee \Psi_\ell(\varepsilon)); {\scriptstyle 01})$. Combining these facts with (15), (30), Condition 4, monotonicity of $\mathrm{vc}(\mathcal{G}_{\mathcal{H}_\mathcal{U}})$ in both $\mathcal{U}$ and $\mathcal{H}$, and the fact that $\|\mathrm{F}(\mathcal{G}_{V^{(2m)}_{D_{2m}}, \mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} \leq \bar{\ell}^2 \mathcal{P}(D_{2m})$, we have that

$$\begin{aligned}\tilde{U}_\ell\left(V^{(2m)}_{D_{2m}}; \mathcal{P}_{XY}, m, \hat{\mathfrak{s}}(2m)\right) \leq c_1 &\sqrt{b\hat{\gamma}_m^\beta \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}\mathcal{P}(D_{2m})}{b\hat{\gamma}_m^\beta}\right) + \hat{\mathfrak{s}}(2m)}{m}} \\ &+ c_1\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}\mathcal{P}(D_{2m})}{b\hat{\gamma}_m^\beta}\right) + \hat{\mathfrak{s}}(2m)}{m}, \quad (63)\end{aligned}$$

for some universal constant $c_1 \in [1,\infty)$. By (60), we have $\mathcal{P}(D_{2m}) \leq \frac{3}{m}(|Q_{2m}| + \hat{\mathfrak{s}}(2m))$, so that the right hand side of (63) is at most

$$c_1\sqrt{b\hat{\gamma}_m^{\beta}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}6(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{m}} + c_1\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}6(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{m}$$
$$\leq 8c_1\sqrt{b\hat{\gamma}_m^{\beta}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{2m}} + 8c_1\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{2m}.$$

Thus, if we take $c_0 = 8c_1$ in the definition of $\hat{T}_\ell$ in (56), then we have

$$\tilde{U}_\ell\left(V_{D_{2m}}^{(2m)};\mathcal{P}_{XY},m,\hat{\mathfrak{s}}(2m)\right)\wedge\bar{\ell} \leq \frac{|Q_{2m}|\vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)};Q_{2m},2m\right)\wedge\bar{\ell}\right).$$

Furthermore, (60) implies $|Q_{2m}| \leq \hat{\mathfrak{s}}(2m) + 2em\mathcal{P}(D_{2m})$. In particular, if $\hat{\mathfrak{s}}(2m) > 2em\mathcal{P}(D_{2m})$, then

$$\frac{|Q_{2m}|\vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)};Q_{2m},2m\right)\wedge\bar{\ell}\right) \leq \frac{\hat{\mathfrak{s}}(2m)+2em\mathcal{P}(D_{2m})}{m}\bar{\ell} \leq \frac{2\hat{\mathfrak{s}}(2m)\bar{\ell}}{m},$$

and taking any $c_2 \geq 4$ guarantees this last quantity is at most $\tilde{T}_\ell(2m)$. On the other hand, if $\hat{\mathfrak{s}}(2m) \leq 2em\mathcal{P}(D_{2m})$, then $|Q_{2m}| \leq 4em\mathcal{P}(D_{2m})$, and we have already established that $V^{(2m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m);_{01})$, so that

$$\frac{|Q_{2m}|\vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)};Q_{2m},2m\right)\wedge\bar{\ell}\right)$$
$$\leq 8c_1\sqrt{b\hat{\gamma}_m^{\beta}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}3e\mathcal{P}(\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m);_{01})))}{b\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{2m}} + 8c_1\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\frac{\bar{\ell}3e\mathcal{P}(\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m);_{01})))}{b\hat{\gamma}_m^{\beta}}\right)+\hat{\mathfrak{s}}(2m)}{2m}. \quad (64)$$

If $\hat{\gamma}_m \geq \Psi_\ell(\varepsilon)$, then this is at most

$$8c_1\left(\sqrt{b\hat{\gamma}_m^{\beta}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(3e\chi_\ell\bar{\ell}\right)+\hat{\mathfrak{s}}(2m)}{2m}}+\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(3e\chi_\ell\bar{\ell}\right)+\hat{\mathfrak{s}}(2m)}{2m}\right)$$
$$\leq 48c_1\left(\sqrt{b\hat{\gamma}_m^{\beta}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\chi_\ell\bar{\ell}\right)+\hat{\mathfrak{s}}(2m)}{2m}}+\bar{\ell}\frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}\left(\chi_\ell\bar{\ell}\right)+\hat{\mathfrak{s}}(2m)}{2m}\right).$$

For brevity, let $K = \frac{\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}(\chi_\ell \bar{\ell}) + \hat{\mathfrak{s}}(2m)}{2m}$. As argued above, $\hat{\gamma}_m \leq 2\tilde{T}_\ell(m)$, so that the right hand side of the above inequality is at most

$$48\sqrt{2}c_1 \left(\sqrt{b\tilde{T}_\ell(m)^\beta K} + \bar{\ell}K \right).$$

Then since $\hat{\mathfrak{s}}(m) \leq 2\hat{\mathfrak{s}}(2m)$, the above expression is at most

$$48 \cdot 4c_1\sqrt{c_2} \left(\sqrt{b\left((bK)^{\frac{1}{2-\beta}} \vee \bar{\ell}K \right)^\beta K} + \bar{\ell}K \right). \tag{65}$$

If $\bar{\ell}K \leq (bK)^{\frac{1}{2-\beta}}$, then (65) is equal

$$48 \cdot 4c_1\sqrt{c_2} \left((bK)^{\frac{1}{2-\beta}} + \bar{\ell}K \right).$$

On the other hand, if $\bar{\ell}K > (bK)^{\frac{1}{2-\beta}}$, then (65) is equal

$$48 \cdot 4c_1\sqrt{c_2} \left(\sqrt{bK(\bar{\ell}K)^\beta} + \bar{\ell}K \right) \\ < 48 \cdot 4c_1\sqrt{c_2} \left(\sqrt{(\bar{\ell}K)^{2-\beta}(\bar{\ell}K)^\beta} + \bar{\ell}K \right) = 48 \cdot 8c_1\sqrt{c_2}\bar{\ell}K.$$

In all of the above cases, taking $c_2 = 9 \cdot 2^{14}c_1^2$ in the definition of $\tilde{T}_\ell$ yields

$$\frac{|Q_{2m}| \vee 1}{m} \left(\hat{T}_\ell \left(V^{(2m)}; Q_{2m}, 2m \right) \wedge \bar{\ell} \right) \leq \tilde{T}_\ell(2m).$$

This completes the inductive step, so that we have proven that the claim holds for all m' with $\log_2(m') \in \{1, \ldots, i_\varepsilon\}$.

Let $j_\ell = -\lceil \log_2(\bar{\ell}) \rceil$, $\tilde{j}_\varepsilon = \lceil \log_2(1/\Psi_\ell(\varepsilon)) \rceil$, and for each $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, let $s_j = \log_2\left(\frac{144(2+\tilde{j}_\varepsilon - j)^2}{\delta} \right)$, define

$$m'_j = 32c_2^2 \left(b2^{j(2-\beta)} + \bar{\ell}2^j \right) \left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}})\mathrm{Log}(\chi_\ell\bar{\ell}) + s_j \right),$$

and let $m_j = 2^{\lceil \log_2(m'_j) \rceil}$. Also define $m_{j_\ell - 1} = 1$. Using this notation, we can now define the relevant values of the $\hat{\mathfrak{s}}$ function as follows. For each $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, and each $m \in \{m_{j-1}+1, \ldots, m_j\}$ with $\log_2(m) \in \mathbb{N}$, define

$$\hat{\mathfrak{s}}(m) = \log_2\left(\frac{16\log_2(4m_j/m)^2(2+\tilde{j}_\varepsilon - j)^2}{\delta} \right).$$

In particular, taking $i_\varepsilon = \log_2(m_{\tilde{j}_\varepsilon})$, we have that $2\tilde{T}_\ell(2^{i_\varepsilon - 1}) \leq \Psi_\ell(\varepsilon)$, so that on the event $\bigcap_{i=1}^{i_\varepsilon - 1} E_{2^i} \cap E''_{2^{i+1}}$, if we have $2^{i_\varepsilon} \in S$, then $\hat{h} \in V^{(2^{i_\varepsilon})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(2^{i_\varepsilon - 1}) \vee \Psi_\ell(\varepsilon)); {}_{01}) = \mathcal{F}(\mathcal{E}_\ell(\Psi_\ell(\varepsilon)); {}_{01}) \subseteq \mathcal{F}(\Psi_\ell^{-1}(\Psi_\ell(\varepsilon)); {}_{01}) = \mathcal{F}(\varepsilon; {}_{01})$, so that $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$.

Furthermore, we established above that, on the event $\bigcap_{i=1}^{i_\varepsilon-1} E_{2^i} \cap E''_{2^{i+1}}$, for every $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$ with $m_j \in S$, and every $m \in \{m_{j-1}+1, \ldots, m_j\}$ with $\log_2(m) \in \mathbb{N}$, $V^{(m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m/2) \vee \Psi_\ell(\varepsilon)); {}_{01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m_{j-1}) \vee \Psi_\ell(\varepsilon)); {}_{01})$. Noting that $2\tilde{T}_\ell(m_{j-1}) \le 2^{1-j}$, we have

$$\sum_{m \in S : m \le m_{\tilde{j}_\varepsilon}} |Q_m| \le \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{m=m_{j-1}+1}^{m_j} \mathbb{1}_{\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01}))}(X_m).$$

A Chernoff bound implies that, on an event E' of probability at least $1-\delta/2$, the right hand side of the above inequality is at most

$$\begin{aligned}\log_2(2/\delta) + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} (m_j - m_{j-1}) \mathcal{P}(\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01}))) \\ \le \log_2(2/\delta) + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} m_j \mathcal{P}(\mathrm{DIS}(\mathcal{F}(\Psi_\ell^{-1}(2^{1-j}); {}_{01}))).\end{aligned}$$

By essentially the same reasoning used in the proof of Theorem 8, the right hand side of this inequality is

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

Since

$$m_{\tilde{j}_\varepsilon} \lesssim \left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) A_1,$$

the conditions on u and n stated in Theorem 8 (with an appropriate constant c) suffice to guarantee $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \le \varepsilon$ on the event $E' \cap \bigcap_{i=1}^{i_\varepsilon-1} E_{2^i} \cap E''_{2^{i+1}}$. Finally, the proof is completed by noting that a union bound implies the event $E' \cap \bigcap_{i=1}^{i_\varepsilon-1} E_{2^i} \cap E''_{2^{i+1}}$ has probability at least

$$\begin{aligned} & 1 - \frac{\delta}{2} - \sum_{i=1}^{i_\varepsilon-1} 2^{1-\hat{\mathfrak{s}}(2^{i+1})} + 6e^{-\hat{\mathfrak{s}}(2^i)} \\ & \ge 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{i=\log_2(m_{j-1})+1}^{\log_2(m_j)} \frac{\delta}{2(2+\log_2(m_j)-i)^2(2+\tilde{j}_\varepsilon-j)^2} \\ & \ge 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{k=0}^{\infty} \frac{\delta}{2(2+k)^2(2+\tilde{j}_\varepsilon-j)^2} \\ & \ge 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \frac{\delta}{2(2+\tilde{j}_\varepsilon-j)^2} \ge 1 - \frac{\delta}{2} - \sum_{t=0}^{\infty} \frac{\delta}{2(2+t)^2} \ge 1-\delta. \end{aligned}$$

Note that, as in Theorem 8, the function $\hat{\mathfrak{s}}$ in this proof has a direct dependence on a, α, and χ_ℓ, in addition to b and β. As before, with an alternative definition of $\hat{\mathfrak{s}}$, similar to that mentioned in the discussion following the proof of Theorem 8, it is possible to remove this dependence, at the expense of the same logarithmic factors mentioned above.

E.2. Proof of Theorem 9 under (57)

Next, consider the conditions of Theorem 9, and suppose the definition of $\hat{T}_\ell$ from (57) is used in Step 6. For simplicity, we let $V^{(m)}$ and Q_m be defined (though arbitrarily) even when $m \notin S$. Fix a function $\hat{\mathfrak{s}}$ (to be specified below) and any value of $\varepsilon \in (0,1)$. We will prove by induction that there exist events $\hat{E}_{m'}$, for values m' with $\log_2(m') \in \mathbb{N}$, each with respective probability at least $1 - 12e^{-\hat{\mathfrak{s}}(m')}$ such that, for every m with $\log_2(m) \in \mathbb{N}$, on $\bigcap_{i=1}^{\log_2(m)} \hat{E}_{2^i}$, if $m \in S$, we have that $f^\star \in \tilde{V}^{(m)}$ and $\tilde{V}^{(m)} \subseteq V^{(m)}\left(4\hat{T}_m; \ell, \mathcal{P}_{D_m}\right)$, where $\hat{T}_m = \hat{T}_\ell\left(V^{(m)}; Q_m, m\right)$. This claim is trivially satisfied for $m = 2$, since $\hat{T}_2 = \bar{\ell}$, so this will serve as our base case in the inductive proof. Now fix any $m > 2$ with $\log_2(m) \in \mathbb{N}$, and take as an inductive hypothesis that there exist events $\hat{E}_{m'}$ for each $m' < m$ with $\log_2(m') \in \mathbb{N}$, such that, on $\bigcap_{i=1}^{\log_2(m)-1} \hat{E}_{2^i}$, if $m/2 \in S$, then $f^\star \in \tilde{V}^{(m/2)}$. Note that, since $V^{(m)} = \tilde{V}^{(m/2)}$ (if $m \in S$), we have that $f^\star \in V^{(m)}$ on $\bigcap_{i=1}^{\log_2(m)-1} \hat{E}_{2^i}$ by the inductive hypothesis.

For any $T > 0$, let $\mathfrak{s}(T,\gamma) = \mathrm{Log}\left(\frac{\gamma}{T}\right) + \hat{\mathfrak{s}}(m)$. Note that (16), (18), (19), Lemma 5, (31), and monotonicity of $\mathcal{H} \mapsto \mathrm{vc}(\mathcal{G}_\mathcal{H})$ imply that, if $f^\star \in V^{(m)} \subseteq \mathcal{F}$, then

$$\sup_{\gamma \ge T} \tilde{M}_\ell\left(\gamma/8, \gamma; V^{(m)}, \mathcal{P}_{D_m}, \mathfrak{s}(T,\gamma)\right) \le \bar{c}\left(\frac{b}{T^{2-\beta}} + \frac{\bar{\ell}}{T}\right)\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{bT^\beta}\right) + \hat{\mathfrak{s}}(m)\right), \quad (66)$$

for an appropriate finite universal constant $\bar{c} \ge 1$. If $m \in S$ and $\hat{T}_m = \bar{\ell}$, then we trivially have $\mathrm{R}_\ell(f^\star; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \le \hat{T}_m$, so that $f^\star \in \tilde{V}^{(m)}$, and furthermore $\tilde{V}^{(m)} = V^{(m)} = V^{(m)}\left(4\hat{T}_m; \ell, \mathcal{P}_{D_m}\right)$. Otherwise, if $m \in S$ and $\hat{T}_m < \bar{\ell}$, we have that

$$|Q_m| \ge \max\begin{cases} \left(\frac{c_0}{\hat{T}_m}\right)^{2-\beta} b\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{|Q_m|}{b\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^{\frac{\beta}{2-\beta}}\right) + \hat{\mathfrak{s}}(m)\right) \\ \frac{c_0\bar{\ell}}{\hat{T}_m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{|Q_m|}{\bar{\ell}\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^\beta\right) + \hat{\mathfrak{s}}(m)\right) \end{cases},$$

which implies

$$\begin{aligned} |Q_m| &\ge \max\left\{\left(\frac{c_0}{\hat{T}_m}\right)^{2-\beta} b, \frac{c_0\bar{\ell}}{\hat{T}_m}\right\}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b\hat{T}_m^\beta}\right) + \hat{\mathfrak{s}}(m)\right) \\ &\ge \frac{c_0}{2}\left(\frac{b}{\hat{T}_m^{2-\beta}} + \frac{\bar{\ell}}{\hat{T}_m}\right)\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b\hat{T}_m^\beta}\right) + \hat{\mathfrak{s}}(m)\right). \end{aligned}$$

Combined with (66), this implies that if we take $c_0 \geq 2\bar{c}$, and if $f^\star \in V^{(m)} \subseteq \mathcal{F}$, then

$$|Q_m| \geq \sup_{\gamma \geq \hat{T}_m} \tilde{M}_\ell \left(\gamma/8, \gamma; V^{(m)}, \mathcal{P}_{D_m}, \mathfrak{s}(\hat{T}_m, \gamma)\right). \tag{67}$$

We now follow the derivation of localized risk bounds by Koltchinskii [34]. Specifically, applying Lemma 15 under the conditional distribution given $V^{(m)}$ and $|Q_m|$, combined with the law of total probability, there is an event E''_m of conditional probability at least $1 - 6\sum_{j \in \mathbb{Z}_{\hat{T}_m}} e^{-\mathfrak{s}(\hat{T}_m, 2^j)}$ (given $V^{(m)}$ and $|Q_m|$), such that on E''_m, if $m \in S$, $f^\star \in V^{(m)}$, and $\hat{T}_m < \bar{\ell}$ (so that (67) holds), then $\forall j \in \mathbb{Z}_{\hat{T}_m}$, the following claims hold for every $h \in V^{(m)}\left(2^j; \ell, \mathcal{P}_{D_m}\right)$.

$$\mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq \mathrm{R}_\ell(h; Q_m) - \mathrm{R}_\ell(f^\star; Q_m) + 2^{j-3}, \tag{68}$$

$$\mathrm{R}_\ell(h; Q_m) - \inf_{g \in V^{(m)}(2^j; \ell, \mathcal{P}_{D_m})} \mathrm{R}_\ell(g; Q_m) \leq \mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) + 2^{j-3}. \tag{69}$$

Since $\sum_{j \in \mathbb{Z}_{\hat{T}_m}} e^{-\mathfrak{s}(\hat{T}_m, 2^j)} = e^{-\hat{\mathfrak{s}}(m)} \sum_{j \in \mathbb{Z}_{\hat{T}_m}} 2^{-j}\hat{T}_m \leq 2e^{-\hat{\mathfrak{s}}(m)}$, the law of total probability implies that there exists an event $\hat{E}_m$ of probability at least $1 - 12e^{-\hat{\mathfrak{s}}(m)}$, on which this implication holds. In particular, for any $h_0 \in V^{(m)}$ with $\mathrm{R}_\ell(h_0; Q_m) - \mathrm{R}_\ell(f^\star; Q_m) \leq 0$, (68) implies that for any $j \in \mathbb{Z}_{\hat{T}_m}$, if $\mathrm{R}_\ell(h_0; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq 2^j$, then $\mathrm{R}_\ell(h_0; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq 2^{j-3}$; this inductively implies that $\mathrm{R}_\ell(h_0; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq \hat{T}_m$, so that (69) can more simply be stated as: $\forall h \in V^{(m)}\left(2^j; \ell, \mathcal{P}_{D_m}\right)$,

$$\mathrm{R}_\ell(h; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \leq \mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) + 2^{j-3}.$$

Furthermore, this implies

$$\mathrm{R}_\ell(f^\star; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \leq \hat{T}_m, \tag{70}$$

so that $f^\star \in \tilde{V}^{(m)}$ in this case as well. Also, (68) and the fact that $f^\star \in V^{(m)}$ further imply that for any $h \in V^{(m)}$ with $\mathrm{R}_\ell(h; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \leq \hat{T}_m$, for any $j \in \mathbb{Z}_{4\hat{T}_m}$, if $\mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq 2^j$, then $\mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq \hat{T}_m + 2^{j-3} \leq 2^{j-2} + 2^{j-3} \leq 2^{j-1}$; this inductively implies that any such h has $\mathrm{R}_\ell(h; \mathcal{P}_{D_m}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{D_m}) \leq 4\hat{T}_m$. In particular, by definition of $\tilde{V}^{(m)}$, this implies $\tilde{V}^{(m)} \subseteq V^{(m)}\left(4\hat{T}_m; \ell, \mathcal{P}_{D_m}\right)$. Since the inductive hypothesis implies $f^\star \in V^{(m)}$ on $\bigcap_{i=1}^{\log_2(m)-1} \hat{E}_{2^i}$ if $m \in S$, we have that on $\bigcap_{i=1}^{\log_2(m)} \hat{E}_{2^i}$, if $m \in S$, then $f^\star \in \tilde{V}^{(m)}$ and $\tilde{V}^{(m)} \subseteq V^{(m)}\left(4\hat{T}_m; \ell, \mathcal{P}_{D_m}\right)$, which extends the inductive hypothesis. By the principle of induction, we have established this claim for every m with $\log_2(m) \in \mathbb{N}$.

Let $\hat{j}_\varepsilon = \left\lceil \log_2(\bar{\ell}/\Psi_\ell(\varepsilon)) \right\rceil$. For each $j \in \mathbb{N} \cup \{0\}$, let $\varepsilon_j = \bar{\ell}2^{-j}$, $p_j = \mathcal{P}\left(\mathrm{DIS}\left(\mathcal{F}\left(\Psi_\ell^{-1}(\varepsilon_j); {}_{01}\right)\right)\right)$, and $s_j = \log_2\left(\frac{192(2+\hat{j}_\varepsilon - j)^2}{\delta}\right)$. Let $m_0 = 1$, and for each $j \in \mathbb{N}$, define

$$m'_j = c' \left(\frac{b p_{j-1}^{1-\beta}}{\varepsilon_j^{2-\beta}} + \frac{\bar{\ell}}{\varepsilon_j} \right) \left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \mathrm{Log}\left(\frac{\bar{\ell}^2 (c')^{\beta} p_{j-1}^{\beta}}{b \varepsilon_j^{\beta}} \right) + s_j \right),$$

for an appropriate universal constant $c' \in [1, \infty)$ (specified below), and let $m_j = \max\left\{ 2m_{j-1}, 2^{1+\lceil \log_2(m'_j) \rceil} \right\}$. Also, for every $j \in \mathbb{N}$ and $m \in \{2m_{j-1}, \ldots, m_j\}$, define

$$\hat{\mathfrak{s}}(m) = \log_2\left(\frac{48 \log_2(4m_j/m)^2 (2 + \hat{j}_{\varepsilon} - j)^2}{\delta} \right).$$

In particular, this definition implies $\hat{\mathfrak{s}}(m_j) = s_j$.

We next prove by induction that there are events $\hat{E}'_j$, for $j \in \mathbb{N} \cup \{0\}$, each with respective probability at least $1 - 2^{-s_j}$, such that for every $j \in \mathbb{N} \cup \{0\}$, on $\bigcap_{i=1}^{\log_2(m_j)} \hat{E}_{2^i} \cap \bigcap_{j'=0}^{j} \hat{E}'_{j'}$, if $m_j \in S \cup \{1\}$, then $\tilde{V}^{(m_j)} \subseteq \mathcal{F}\left(\Psi_{\ell}^{-1}(\varepsilon_j); {\scriptstyle 01}\right)$. This claim is trivially satisfied for $j = 0$, which therefore serves as the base case for this inductive proof. Now fix any $j \in \mathbb{N}$, and take as an inductive hypothesis that there exist events $\hat{E}'_{j'}$, as above, for all $j' < j$, such that on $\bigcap_{i=1}^{\log_2(m_{j-1})} \hat{E}_{2^i} \cap \bigcap_{j'=0}^{j-1} \hat{E}'_{j'}$, if $m_{j-1} \in S$, then $\tilde{V}^{(m_{j-1})} \subseteq \mathcal{F}\left(\Psi_{\ell}^{-1}(\varepsilon_{j-1}); {\scriptstyle 01}\right)$. By the above, we have that on $\bigcap_{i=1}^{\log_2(m_j)} \hat{E}_{2^i}$, if $m_j \in S$, then $f^{\star} \in \tilde{V}^{(m_j)} \subseteq V^{(m_j)}\left(4\hat{T}_{m_j}; \ell, \mathcal{P}_{D_{m_j}}\right)$. In particular, this implies that every $h \in \tilde{V}^{(m_j)}$ has

$$\begin{aligned} \mathrm{R}_{\ell}(h_{D_{m_j}}; \mathcal{P}_{XY}) - \mathrm{R}_{\ell}(f^{\star}; \mathcal{P}_{XY}) &= \Big(\mathrm{R}_{\ell}(h; \mathcal{P}_{D_{m_j}}) - \mathrm{R}_{\ell}(f^{\star}; \mathcal{P}_{D_{m_j}}) \Big) \mathcal{P}(D_{m_j}) \\ &\leq 4\hat{T}_{m_j} \mathcal{P}(D_{m_j}). \end{aligned} \tag{71}$$

By a Chernoff bound and the law of total probability, on an event $\hat{E}'_j$ of probability at least $1 - 2^{-s_j}$, if $m_j \in S$,

$$(1/2) m_j \mathcal{P}(D_{m_j}) - \sqrt{s_j m_j \mathcal{P}(D_m)} \leq |Q_{m_j}|. \tag{72}$$

If $m_j \in S$ and $\mathcal{P}(D_{m_j}) \leq \frac{16 s_j}{m_j}$, then $4\hat{T}_{m_j}\mathcal{P}(D_{m_j}) \leq \frac{64 \bar{\ell} s_j}{m_j} \leq \frac{32 \varepsilon_j}{c'}$, so that with any $c' \geq 32$, (71) would give $\mathrm{R}_{\ell}(h_{D_{m_j}}; \mathcal{P}_{XY}) - \mathrm{R}_{\ell}(f^{\star}; \mathcal{P}_{XY}) \leq \varepsilon_j$. Otherwise, (72) implies that on $\hat{E}'_j$, if $m_j \in S$ and $\mathcal{P}(D_{m_j}) > \frac{16 s_j}{m_j}$, then $|Q_{m_j}| \geq (1/4) m_j \mathcal{P}(D_{m_j})$. In this latter case, we have

$$\begin{aligned} &4\hat{T}_{m_j}\mathcal{P}(D_{m_j}) \leq \\ &16 c_0 \max\left\{ \begin{array}{l} \mathcal{P}(D_{m_j})^{\frac{1-\beta}{2-\beta}} \left(\frac{b}{m_j} \left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \mathrm{Log}\left(\frac{\bar{\ell}^2}{b} \left(\frac{m_j \mathcal{P}(D_{m_j})}{4 b \mathrm{vc}(\mathcal{G}_{\mathcal{F}})} \right)^{\frac{\beta}{2-\beta}} \right) + s_j \right) \right)^{\frac{1}{2-\beta}} \\ \frac{\bar{\ell}}{m_j} \left(\mathrm{vc}(\mathcal{G}_{\mathcal{F}}) \mathrm{Log}\left(\frac{\bar{\ell}^2}{b} \left(\frac{m_j \mathcal{P}(D_{m_j})}{4 \bar{\ell} \mathrm{vc}(\mathcal{G}_{\mathcal{F}})} \right)^{\beta} \right) + s_j \right) \end{array} \right. . \end{aligned} \tag{73}$$

Since $m_j \geq 2m_{j-1}$, by the inductive hypothesis, on $\bigcap_{i=1}^{\log_2(m_{j-1})} \hat{E}_{2^i} \cap \bigcap_{j'=0}^{j-1} \hat{E}'_{j'}$, if $m_j \in S$, we have $V^{(m_j)} \subseteq \tilde{V}^{(m_{j-1})} \subseteq \mathcal{F}\left(\Psi_\ell^{-1}(\varepsilon_{j-1}); {}_{01}\right)$, which implies $\mathcal{P}(D_{m_j}) \leq \mathcal{P}\left(\mathrm{DIS}\left(\mathcal{F}\left(\Psi_\ell^{-1}\left(\varepsilon_{j-1}\right); {}_{01}\right)\right)\right) = p_{j-1}$. In this case, the right hand side of (73) is at most

$$16c_0 \max \begin{cases} p_{j-1}^{\frac{1-\beta}{2-\beta}} \left(\frac{b}{m_j}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{m_j p_{j-1}}{4b\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^{\frac{\beta}{2-\beta}}\right) + s_j\right)\right)^{\frac{1}{2-\beta}} \\ \frac{\bar{\ell}}{m_j}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{m_j p_{j-1}}{4\bar{\ell}\mathrm{vc}(\mathcal{G}_\mathcal{F})}\right)^{\beta}\right) + s_j\right) \end{cases}.$$

The value of m'_j was defined to make this value at most ε_j, with any value of $c' \geq 16c_0$. Altogether, we have that on $\bigcap_{i=1}^{\log_2(m_j)} \hat{E}_{2^i} \cap \bigcap_{j'=0}^{j} \hat{E}'_{j'}$, if $m_j \in S$, then every $h \in \tilde{V}^{(m_j)}$ has $\mathrm{R}_\ell(h_{D_{m_j}}; \mathcal{P}_{XY}) - \mathrm{R}_\ell(f^\star; \mathcal{P}_{XY}) \leq \varepsilon_j$; in particular, this also implies every $h \in \tilde{V}^{(m_j)}$ has $\mathrm{er}(h_{D_{m_j}}) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}(\varepsilon_j)$. Since we have already proven that $f^\star \in V^{(m_j)}$ on this event, and since $\tilde{V}^{(m_j)} \subseteq V^{(m)}$, we have that every $h \in \tilde{V}^{(m)}$ has $\mathrm{er}(h) = \mathrm{er}(h_{D_m})$, which therefore implies $\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}(\varepsilon_j)$: that is, $\tilde{V}^{(m_j)} \subseteq \mathcal{F}\left(\Psi_\ell^{-1}(\varepsilon_j); {}_{01}\right)$. This completes the inductive proof.

The above result implies that, on $\bigcap_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} \hat{E}_{2^i} \cap \bigcap_{j=0}^{\hat{j}_\varepsilon} \hat{E}'_j$, if $m_{\hat{j}_\varepsilon} \in S$, then $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}(\varepsilon_{\hat{j}_\varepsilon}) \leq \Psi_\ell^{-1}\left(\Psi_\ell(\varepsilon)\right) = \varepsilon$. In particular, we are guaranteed to have $m_{\hat{j}_\varepsilon} \in S$ as long as $u \geq m_{\hat{j}_\varepsilon}$ and

$$n > \sum_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} \sum_{m=2^{i-1}+1}^{\min\left\{2^i, \max S\right\}} \mathbb{1}_{\mathrm{DIS}\left(\tilde{V}^{(2^{i-1})}\right)}(X_m). \tag{74}$$

By monotonicity of $m \mapsto \mathrm{DIS}\left(\tilde{V}^{(m)}\right)$, the right hand side of (74) is at most

$$\sum_{j=0}^{\hat{j}_\varepsilon} \sum_{m=m_{j-1}+1}^{\min\{m_j, \max S\}} \mathbb{1}_{\mathrm{DIS}\left(\tilde{V}^{(m_{j-1})}\right)}(X_m).$$

Furthermore, on $\bigcap_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} \hat{E}_{2^i} \cap \bigcap_{j=0}^{\hat{j}_\varepsilon} \hat{E}'_j$, the above result implies this is at most

$$\sum_{j=1}^{\hat{j}_\varepsilon} \sum_{m=m_{j-1}+1}^{\min\{m_j, \max S\}} \mathbb{1}_{\mathrm{DIS}\left(\mathcal{F}\left(\Psi_\ell^{-1}(\varepsilon_{j-1}); {}_{01}\right)\right)}(X_m)$$
$$\leq \sum_{j=1}^{\hat{j}_\varepsilon} \sum_{m=m_{j-1}+1}^{m_j} \mathbb{1}_{\mathrm{DIS}\left(\mathcal{F}\left(\Psi_\ell^{-1}(\varepsilon_{j-1}); {}_{01}\right)\right)}(X_m).$$

By a Chernoff bound, on an event $\hat{E}''$ of probability at least $1 - \delta/2$, the right hand side of the above is at most

$$\log_2(2/\delta) + \sum_{j=1}^{\hat{j}_\varepsilon} (m_j - m_{j-1}) p_{j-1}. \tag{75}$$

Since $\varepsilon_{j-1} \geq \bar{\ell}2^{1-\hat{j}_\varepsilon} \geq \Psi_\ell(\varepsilon)$, and therefore

$$p_{j-1} \leq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star, a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha\right)\right)\right)$$
$$\leq \theta\left(a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha\right) a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha \leq \theta\left(a\varepsilon^\alpha\right) a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha,$$

letting $\hat{c}_j = \mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{c'\theta a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha}{\varepsilon_j}\right)^\beta\right)$, we have that

$$2^{1+\lceil\log_2(m_j')\rceil} \leq 4c'\left(\frac{b}{\varepsilon_j}\left(\frac{\theta a\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha}{\varepsilon_j}\right)^{1-\beta} + \frac{\bar{\ell}}{\varepsilon_j}\right)\left(\hat{c}_j + s_j\right). \tag{76}$$

Since $\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha/\varepsilon_j$ is nondecreasing in j, the right hand side of (76) at least doubles when j is increased by one, so that by induction we have that the right hand side of (76) is also an upper bound on m_j. This fact also implies that $\hat{c}_j + s_j$ is at most

$$\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{2c'\theta a\Psi_\ell^{-1}(2\Psi_\ell(\varepsilon))^\alpha}{\Psi_\ell(\varepsilon)}\right)^\beta\right) + \mathrm{Log}\left(\frac{192}{\delta}\right) + 2\mathrm{Log}\left(2 + \hat{j}_\varepsilon - j\right),$$

and the fact that $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing implies this is at most

$$\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}^2}{b}\left(\frac{4c'\theta a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^\beta\right) + \mathrm{Log}\left(\frac{192}{\delta}\right) + 2\mathrm{Log}\left(2 + \hat{j}_\varepsilon - j\right)$$
$$\leq c''\left(A_2 + \mathrm{Log}\left(2 + \hat{j}_\varepsilon - j\right)\right).$$

where $c'' = \ln(768ec')$. Furthermore,

$$\frac{\Psi_\ell^{-1}(\varepsilon_{j-1})^\alpha}{\varepsilon_j} = 2\frac{\Psi_\ell^{-1}(2^{(\hat{j}_\varepsilon - j)}\varepsilon_{\hat{j}_\varepsilon - 1})^\alpha}{2^{(\hat{j}_\varepsilon - j)}\varepsilon_{\hat{j}_\varepsilon - 1}}$$
$$\leq 2\frac{\Psi_\ell^{-1}(2^{(\hat{j}_\varepsilon - j)}\Psi_\ell(\varepsilon))^\alpha}{2^{(\hat{j}_\varepsilon - j)}\Psi_\ell(\varepsilon)} \leq 2^{1+(\hat{j}_\varepsilon - j)(\alpha-1)}\frac{\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}.$$

Applying these inequalities to bound $m_j p_{j-1}$, and reversing the order of summation (now summing over $i = \hat{j}_\varepsilon - j$), we have that

$$\sum_{j=1}^{\hat{j}_\varepsilon} m_j p_{j-1} \leq 16c'c'' \sum_{i=0}^{\hat{j}_\varepsilon - 1} b\left(\frac{a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta}\left(A_2 + \mathrm{Log}(i+2)\right)$$
$$+ 16c'c'' \sum_{i=0}^{\hat{j}_\varepsilon - 1} \frac{\bar{\ell}a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\left(A_2 + \mathrm{Log}(i+2)\right).$$

Note that this is of the same form as (55) in the proof of Theorem 9, so that following that proof, the right hand side above is at most

$$144c'c''\left(b(A_2 + \mathrm{Log}(C_1)C_1\left(\frac{\theta a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}(A_2 + \mathrm{Log}(C_1))C_1\left(\frac{\theta a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right).$$

Therefore, since $\log_2(2/\delta) \leq 3A_2$, (75) is less than

$$147c'c''\left(b(A_2 + \mathrm{Log}(C_1)C_1\left(\frac{\theta a\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}(A_2 + \mathrm{Log}(C_1))C_1\left(\frac{\theta a\varepsilon^{\alpha}}{\Psi_\ell(\varepsilon)}\right)\right).$$

The above inequalities also imply that

$$m_{\hat{j}_\varepsilon} \leq 32c'c''\left(\frac{b\,(\theta a\varepsilon^{\alpha})^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)A_2.$$

Thus, taking $c = 147c'c''$ in the statement of Theorem 9 suffices to guarantee that, for any u and n satisfying the given size constraints, $u \geq m_{\hat{j}_\varepsilon}$, and on the event $\bigcap_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} \hat{E}_{2^i} \cap \bigcap_{j=0}^{\hat{j}_\varepsilon} \hat{E}'_j \cap \hat{E}''$, (74) is satisfied, which (as discussed above) implies $\mathrm{er}(\hat{h}) - \mathrm{er}(f^{\star}) \leq \varepsilon$ on this event. We complete the proof by noting that, by a union bound, the event $\bigcap_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} \hat{E}_{2^i} \cap \bigcap_{j=0}^{\hat{j}_\varepsilon} \hat{E}'_j \cap \hat{E}''$ has probability at least

$$1 - \sum_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} 12e^{-\hat{\mathfrak{s}}(2^i)} - \sum_{j=0}^{\hat{j}_\varepsilon} 2^{-s_j} - \frac{\delta}{2},$$

which is greater than $1 - \delta$, since

$$\sum_{i=1}^{\log_2(m_{\hat{j}_\varepsilon})} 12e^{-\hat{\mathfrak{s}}(2^i)} \leq \sum_{j=1}^{\hat{j}_\varepsilon} \sum_{i=\log_2(m_{j-1})+1}^{\log_2(m_j)} \frac{\delta}{4\log(4m_j/2^i)^2(2+\hat{j}_\varepsilon - j)^2}$$

$$\leq \sum_{j=1}^{\hat{j}_\varepsilon}\sum_{k=0}^{\infty} \frac{\delta}{4(2+k)^2(2+\hat{j}_\varepsilon - j)^2} \leq \sum_{j=1}^{\hat{j}_\varepsilon} \frac{\delta}{4(2+\hat{j}_\varepsilon - j)^2} \leq \sum_{k=0}^{\infty}\frac{\delta}{4(2+k)^2} \leq \frac{\delta}{4},$$

and $\sum_{j=0}^{\hat{j}_\varepsilon} 2^{-s_j} \leq \sum_{j=0}^{\hat{j}_\varepsilon} \frac{\delta}{192(2+\hat{j}_\varepsilon - j)^2} \leq \sum_{k=0}^{\infty} \frac{\delta}{192(2+k)^2} \leq \frac{\delta}{192}$.

Appendix F: Remarks on the Assumption that $f^{\star} \in \mathcal{F}$

We conclude with some remarks on the assumption that $f^{\star} \in \mathcal{F}$ (used throughout this article). As noted in Section 2.1, this assumption is often *very* strong. While the specific assumption that $f^{\star} \in \mathcal{F}$ adds a certain elegance to the theory developed in this work, one natural question is to what extent it can be relaxed without changing the essence of the approach considered here. For instance, in passive learning, one can generalize the abstract results on empirical risk minimization (stated in Theorem 17) to hold under the weaker condition that $\mathrm{argmin}_{h\in\mathcal{F}} \mathrm{R}_\ell(h) = \mathrm{argmin}_{h\in\mathcal{F}} \mathrm{er}(h)$. However, this simple relaxation appears insufficient for the approach to active learning considered here. Specifically, for our analysis, we would require that an error minimizer $\mathrm{argmin}_{h\in\mathcal{F}} \mathrm{er}(h)$ also be an (approximate) minimizer of $\mathrm{R}_\ell(h;P)$ in $\mathcal{F}$, not merely for $P = \mathcal{P}_{XY}$, but also for certain conditional distributions $\mathcal{P}_{XY}(\cdot|\mathrm{DIS}(V)\times\mathcal{Y})$, for sets $V \subseteq \mathcal{F}$ arising in the algorithm. In principle, the results in this work can be generalized to provide guarantees when this condition (suitably formalized) is satisfied. However, the statements of the results become considerably more involved, and moreover we do not know of concise, general, *a priori* conditions on $\mathcal{F}$, ℓ, and $\mathcal{P}_{XY}$, under which this property will hold. Beyond this, it appears our analysis does not easily extend to the important problem of active learning with surrogate losses in the *general* case, where results would presumably need to be expressed in terms of the approximation loss $\inf_{f \in \mathcal{F}} \mathrm{R}_{\ell}(f) - \mathrm{R}_{\ell}(f^{\star})$ or related quantities (as observed for passive learning [6]). It seems such a generalization would require a significantly different approach.