Beyond Disagreement-based Agnostic Active Learning

Chicheng Zhang
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
chz038@eng.ucsd.edu

Kamalika Chaudhuri
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
kamalika@cs.ucsd.edu

Abstract

We study agnostic active learning, where the goal is to learn a classifier in a pre-specified hypothesis class interactively with as few label queries as possible, while making no assumptions on the true function generating the labels. The main algorithm for this problem is *disagreement-based active learning*, which has a high label requirement. Thus a major challenge is to find an algorithm which achieves better label complexity, is consistent in an agnostic setting, and applies to general classification problems.

In this paper, we provide such an algorithm. Our solution is based on two novel contributions; first, a reduction from consistent active learning to confidence-rated prediction with guaranteed error, and second, a novel confidence-rated predictor.

1 Introduction

In this paper, we study *active learning* of classifiers in an agnostic setting, where no assumptions are made on the true function that generates the labels. The learner has access to a large pool of unlabelled examples, and can interactively request labels for a small subset of these; the goal is to learn an accurate classifier in a pre-specified class with as few label queries as possible. Specifically, we are given a hypothesis class $\mathcal{H}$ and a target ϵ, and our aim is to find a binary classifier in $\mathcal{H}$ whose error is at most ϵ more than that of the best classifier in $\mathcal{H}$, while minimizing the number of requested labels.

There has been a large body of previous work on active learning; see the surveys by [10, 28] for overviews. The main challenge in active learning is ensuring consistency in the agnostic setting while still maintaining low label complexity. In particular, a very natural approach to active learning is to view it as a generalization of binary search [17, 9, 27]. While this strategy has been extended to several different noise models [23, 27, 26], it is generally inconsistent in the agnostic case [11].

The primary algorithm for agnostic active learning is called *disagreement-based active learning*. The main idea is as follows. A set V_k of possible risk minimizers is maintained with time, and the label of an example x is queried if there exist two hypotheses h_1 and h_2 in V_k such that $h_1(x) \neq h_2(x)$. This algorithm is consistent in the agnostic setting [7, 2, 12, 18, 5, 19, 6, 24]; however, due to the conservative label query policy, its label requirement is high. A line of work due to [3, 4, 1] have provided algorithms that achieve better label complexity for linear classification on the uniform distribution over the unit sphere as well as log-concave distributions; however, their algorithms are limited to these specific cases, and it is unclear how to apply them more generally.

Thus, a major challenge in the agnostic active learning literature has been to find a general active learning strategy that applies to any hypothesis class and data distribution, is consistent in the agnostic case, and has a better label requirement than disagreement based active learning. This has been mentioned as an open problem by several works, such as [2, 10, 4].

In this paper, we provide such an algorithm. Our solution is based on two key contributions, which may be of independent interest. The first is a general connection between *confidence-rated predictors* and active learning. A confidence-rated predictor is one that is allowed to abstain from prediction on occasion, and as a result, can guarantee a target prediction error. Given a confidence-rated predictor with guaranteed error, we show how to use it to construct an active label query algorithm consistent in the agnostic setting. Our second key contribution is a novel confidence-rated predictor with guaranteed error that applies to any general classification problem. We show that our predictor is *optimal* in the realizable case, in the sense that it has the lowest abstention rate out of all predictors that guarantee a certain error. Moreover, we show how to extend our predictor to the agnostic setting.

Combining the label query algorithm with our novel confidence-rated predictor, we get a general active learning algorithm consistent in the agnostic setting. We provide a characterization of the label complexity of our algorithm, and show that this is better than disagreement-based active learning in general. Finally, we show that for linear classification with respect to the uniform distribution and log-concave distributions, our bounds reduce to those of [3, 4].

2 Algorithm

2.1 The Setting

We study active learning for binary classification. Examples belong to an instance space $\mathcal{X}$, and their labels lie in a label space $\mathcal{Y} = \{-1, 1\}$; labelled examples are drawn from an underlying data distribution D on $\mathcal{X} \times \mathcal{Y}$. We use $D_{\mathcal{X}}$ to denote the marginal on D on $\mathcal{X}$, and $D_{Y|X}$ to denote the conditional distribution on $Y|X = x$ induced by D. Our algorithm has access to examples through two oracles – an example oracle $\mathcal{U}$ which returns an unlabelled example $x \in \mathcal{X}$ drawn from $D_{\mathcal{X}}$ and a labelling oracle $\mathcal{O}$ which returns the label y of an input $x \in \mathcal{X}$ drawn from $D_{Y|X}$.

Given a hypothesis class $\mathcal{H}$ of VC dimension d, the error of any $h \in \mathcal{H}$ with respect to a data distribution Π over $\mathcal{X} \times \mathcal{Y}$ is defined as $\text{err}_{\Pi}(h) = \mathbb{P}_{(x,y)\sim\Pi}(h(x) \neq y)$. We define: $h^*(\Pi) = \text{argmin}_{h\in\mathcal{H}}\text{err}_{\Pi}(h)$, $\nu^*(\Pi) = \text{err}_{\Pi}(h^*(\Pi))$. For a set S, we abuse notation and use S to also denote the uniform distribution over the elements of S. We define $\mathbb{P}_{\Pi}(\cdot) := \mathbb{P}_{(x,y)\sim\Pi}(\cdot)$, $\mathbb{E}_{\Pi}(\cdot) := \mathbb{E}_{(x,y)\sim\Pi}(\cdot)$.

Given access to examples from a data distribution D through an example oracle $\mathcal{U}$ and a labeling oracle $\mathcal{O}$, we aim to provide a classifier $\hat{h} \in \mathcal{H}$ such that with probability $\geq 1 - \delta$, $\text{err}_D(\hat{h}) \leq \nu^*(D) + \epsilon$, for some target values of ϵ and δ; this is achieved in an adaptive manner by making as few queries to the labelling oracle $\mathcal{O}$ as possible. When $\nu^*(D) = 0$, we are said to be in the *realizable case*; in the more general *agnostic* case, we make no assumptions on the labels, and thus $\nu^*(D)$ can be positive.

Previous approaches to agnostic active learning have frequently used the notion of *disagreements*. The disagreement between two hypotheses h_1 and h_2 with respect to a data distribution Π is the fraction of examples according to Π to which h_1 and h_2 assign different labels; formally: $\rho_{\Pi}(h_1, h_2) = \mathbb{P}_{(x,y)\sim\Pi}(h_1(x) \neq h_2(x))$. Observe that a data distribution Π induces a pseudo-metric ρ_{Π} on the elements of $\mathcal{H}$; this is called the disagreement metric. For any r and any $h \in \mathcal{H}$, define $B_{\Pi}(h, r)$ to be the disagreement ball of radius r around h with respect to the data distribution Π. Formally: $B_{\Pi}(h, r) = \{h' \in \mathcal{H} : \rho_{\Pi}(h, h') \leq r\}$.

For notational simplicity, we assume that the hypothesis space is "dense" with repsect to the data distribution D, in the sense that $\forall r > 0$, $\sup_{h\in B_D(h^*(D),r)} \rho_D(h, h^*(D)) = r$. Our analysis will still apply without the denseness assumption, but will be significantly more messy. Finally, given a set of hypotheses $V \subseteq \mathcal{H}$, the *disagreement region* of V is the set of all examples x such that there exist two hypotheses $h_1, h_2 \in V$ for which $h_1(x) \neq h_2(x)$.

This paper establishes a connection between active learning and confidence-rated predictors with guaranteed error. A confidence-rated predictor is a prediction algorithm that is occasionally allowed to abstain from classification. We will consider such predictors in the transductive setting. Given a set V of candidate hypotheses, an error guarantee η, and a set U of unlabelled examples, a confidence-rated predictor P either assigns a label or abstains from prediction on each unlabelled

$x \in U$. The labels are assigned with the guarantee that the expected disagreement[1] between the label assigned by P and any $h \in V$ is $\leq \eta$. Specifically,

$$\text{for all } h \in V, \quad \mathbb{P}_{x \sim U}(h(x) \neq P(x), P(x) \neq 0) \leq \eta \tag{1}$$

This ensures that if some $h^* \in V$ is the true risk minimizer, then, the labels predicted by P on U do not differ very much from those predicted by h^*. The performance of a confidence-rated predictor which has a guarantee such as in Equation (1) is measured by its *coverage*, or the probability of non-abstention $\mathbb{P}_{x \sim U}(P(x) \neq 0)$; higher coverage implies better performance.

2.2 Main Algorithm

Our active learning algorithm proceeds in epochs, where the goal of epoch k is to achieve excess generalization error $\epsilon_k = \epsilon 2^{k_0 - k + 1}$, by querying a fresh batch of labels. The algorithm maintains a candidate set V_k that is guaranteed to contain the true risk minimizer.

The critical decision at each epoch is how to select a subset of unlabelled examples whose labels should be queried. We make this decision using a confidence-rated predictor P. At epoch k, we run P with candidate hypothesis set $V = V_k$ and error guarantee $\eta = \epsilon_k/64$. Whenever P abstains, we query the label of the example. The number of labels m_k queried is adjusted so that it is enough to achieve excess generalization error ϵ_{k+1}.

An outline is described in Algorithm 1; we next discuss each individual component in detail.

Algorithm 1 Active Learning Algorithm: Outline

```
Inputs: Example oracle $\mathcal{U}$, Labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$ of VC dimension $d$,
  confidence-rated predictor $P$, target excess error $\epsilon$ and target confidence $\delta$.
Set $k_0 = \lceil \log 1/\epsilon \rceil$. Initialize candidate set $V_1 = \mathcal{H}$.
for $k = 1, 2, ..k_0$ do
    Set $\epsilon_k = \epsilon 2^{k_0 - k + 1}$, $\delta_k = \frac{\delta}{2(k_0 - k + 1)^2}$.
    Call $\mathcal{U}$ to generate a fresh unlabelled sample $U_k = \{z_{k,1}, ..., z_{k,n_k}\}$ of size $n_k =
      192(\frac{512}{\epsilon_k})^2(d \ln 192(\frac{512}{\epsilon_k})^2 + \ln \frac{288}{\delta_k})$.
    Run confidence-rated predictor $P$ with inpuy $V = V_k$, $U = U_k$ and error guarantee
      $\eta = \epsilon_k/64$ to get abstention probabilities $\gamma_{k,1}, \ldots, \gamma_{k,n_k}$ on the examples in $U_k$. These
      probabilities induce a distribution $\Gamma_k$ on $U_k$. Let $\phi_k = \mathbb{P}_{x \sim U_k}(P(x) = 0) = \frac{1}{n_k} \sum_{i=1}^{n_k} \gamma_{k,i}$.
    if in the Realizable Case then
        Let $m_k = \frac{1536\phi_k}{\epsilon_k}(d \ln \frac{1536\phi_k}{\epsilon_k} + \ln \frac{48}{\delta_k})$. Draw $m_k$ i.i.d examples from $\Gamma_k$ and query
          $\mathcal{O}$ for labels of these examples to get a labelled data set $S_k$. Update $V_{k+1}$ using $S_k$:
          $V_{k+1} := \{h \in V_k : h(x) = y, \text{ for all } (x, y) \in S_k\}$.
    else
        In the non-realizable case, use Algorithm 2 with inputs hypothesis set $V_k$, distribution
          $\Gamma_k$, target excess error $\frac{\epsilon_k}{8\phi_k}$, target confidence $\frac{\delta_k}{2}$, and the labeling oracle $\mathcal{O}$ to get a new
          hypothesis set $V_{k+1}$.
return an arbitrary $\hat{h} \in V_{k_0+1}$.
```

Candidate Sets. At epoch k, we maintain a set V_k of candidate hypotheses guaranteed to contain the true risk minimizer $h^*(D)$ (w.h.p). In the realizable case, we use a version space as our candidate set. The version space with respect to a set S of labelled examples is the set of all $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all $(x_i, y_i) \in S$.

Lemma 1. *Suppose we run Algorithm 1 in the realizable case with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P, target excess error ϵ and target confidence δ. Then, with probability* 1, $h^*(D) \in V_k$, *for all* $k = 1, 2, \ldots, k_0 + 1$.

In the non-realizable case, the version space is usually empty; we use instead a $(1 - \alpha)$-confidence set for the true risk minimizer. Given a set S of n labelled examples, let $C(S) \subseteq \mathcal{H}$ be a function of

[1] where the expectation is with respect to the random choices made by P

S; $C(S)$ is said to be a $(1-\alpha)$-confidence set for the true risk minimizer if for all data distributions Δ over $\mathcal{X} \times \mathcal{Y}$,

$$\mathbb{P}_{S \sim \Delta^n}[h^*(\Delta) \in C(S)] \geq 1 - \alpha,$$

Recall that $h^*(\Delta) = \text{argmin}_{h \in \mathcal{H}} \text{err}_\Delta(h)$. In the non-realizable case, our candidate sets are $(1-\alpha)$-confidence sets for $h^*(D)$, for $\alpha = \delta$. The precise setting of V_k is explained in Algorithm 2.

Lemma 2. *Suppose we run Algorithm 1 in the non-realizable case with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P, target excess error ϵ and target confidence δ. Then with probability $1-\delta$, $h^*(D) \in V_k$, for all $k = 1, 2, \ldots, k_0 + 1$.*

Label Query. We next discuss our label query procedure – which examples should we query labels for, and how many labels should we query at each epoch?

Which Labels to Query? Our goal is to query the labels of the most informative examples. To choose these examples while still maintaining consistency, we use a confidence-rated predictor P with guaranteed error. The inputs to the predictor are our candidate hypothesis set V_k which contains (w.h.p) the true risk minimizer, a fresh set U_k of unlabelled examples, and an error guarantee $\eta = \epsilon_k/64$. For notation simplicity, assume the elements in U_k are distinct. The output is a sequence of abstention probabilities $\{\gamma_{k,1}, \gamma_{k,2}, \ldots, \gamma_{k,n_k}\}$, for each example in U_k. It induces a distribution Γ_k over U_k, from which we independently draw examples for label queries.

How Many Labels to Query? The goal of epoch k is to achieve excess generalization error ϵ_k. To achieve this, passive learning requires $\tilde{O}(d/\epsilon_k)$ labelled examples[2] in the realizable case, and $\tilde{O}(d(\nu^*(D) + \epsilon_k)/\epsilon_k^2)$ examples in the agnostic case. A key observation in this paper is that in order to achieve excess generalization error ϵ_k on D, it suffices to achieve a much larger excess generalization error $O(\epsilon_k/\phi_k)$ on the data distribution induced by Γ_k and $D_{Y|X}$, where ϕ_k is the fraction of examples on which the confidence-rated predictor abstains.

In the realizable case, we achieve this by sampling $m_k = \frac{1536\phi_k}{\epsilon_k}(d \ln \frac{1536\phi_k}{\epsilon_k} + \ln \frac{48}{\delta_k})$ i.i.d examples from Γ_k, and querying their labels to get a labelled dataset S_k. Observe that as ϕ_k is the abstention probability of P with guaranteed error $\leq \epsilon_k/64$, it is generally smaller than the measure of the disagreement region of the version space; this key fact results in improved label complexity over disagreement-based active learning. This sampling procedure has the following property:

Lemma 3. *Suppose we run Algorithm 1 in the realizable case with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P, target excess error ϵ and target confidence δ. Then with probability $1-\delta$, for all $k = 1, 2, \ldots, k_0 + 1$, and for all $h \in V_k$, $err_D(h) \leq \epsilon_k$. In particular, the $\hat{h}$ returned at the end of the algorithm satisfies $err_D(\hat{h}) \leq \epsilon$.*

The agnostic case has an added complication – in practice, the value of ν^* is not known ahead of time. Inspired by [24], we use a *doubling procedure*(stated in Algorithm 2) which adaptively finds the number m_k of labelled examples to be queried and queries them. The following two lemmas illustrate its properties – that it is consistent, and that it does not use too many label queries.

Lemma 4. *Suppose we run Algorithm 2 with inputs hypothesis set V, example distribution Δ, labelling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon}$ and target confidence $\tilde{\delta}$. Let $\tilde{\Delta}$ be the joint distribution on $\mathcal{X} \times \mathcal{Y}$ induced by Δ and $D_{Y|X}$. Then there exists an event $\tilde{E}$, $\mathbb{P}(\tilde{E}) \geq 1 - \tilde{\delta}$, such that on $\tilde{E}$, (1) Algorithm 2 halts and (2) the set V_{j_0} has the following properties:*

(2.1) If for $h \in \mathcal{H}$, $err_{\tilde{\Delta}}(h) - err_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \tilde{\epsilon}/2$, then $h \in V_{j_0}$.

(2.2) On the other hand, if $h \in V_{j_0}$, then $err_{\tilde{\Delta}}(h) - err_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \tilde{\epsilon}$.

When event $\tilde{E}$ happens, we say Algorithm 2 succeeds.

Lemma 5. *Suppose we run Algorithm 2 with inputs hypothesis set V, example distribution Δ, labelling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon}$ and target confidence $\tilde{\delta}$. There exists some absolute constant $c_1 > 0$, such that on the event that Algorithm 2 succeeds, $n_{j_0} \leq c_1((d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})\frac{\nu^*(\tilde{\Delta}) + \tilde{\epsilon}}{\tilde{\epsilon}^2})$. Thus the total number of labels queried is $\sum_{j=1}^{j_0} n_j \leq 2n_{j_0} \leq 2c_1((d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})\frac{\nu^*(\tilde{\Delta}) + \tilde{\epsilon}}{\tilde{\epsilon}^2})$.*

[2] $\tilde{O}(\cdot)$ hides logarithmic factors

A naive approach (see Algorithm 4 in the Appendix) which uses an additive VC bound gives a sample complexity of $O((d\ln(1/\tilde{\epsilon}) + \ln(1/\tilde{\delta}))\tilde{\epsilon}^{-2})$; Algorithm 2 gives a better sample complexity.

The following lemma is a consequence of our label query procedure in the non-realizable case.

Lemma 6. *Suppose we run Algorithm 1 in the non-realizable case with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P, target excess error ϵ and target confidence δ. Then with probability $1-\delta$, for all $k = 1, 2, \ldots, k_0+1$, and for all $h \in V_k$, $err_D(h) \leq err_D(h^*(D)) + \epsilon_k$. In particular, the $\hat{h}$ returned at the end of the algorithm satisfies $err_D(\hat{h}) \leq err_D(h^*(D)) + \epsilon$.*

Algorithm 2 An Adaptive Algorithm for Label Query Given Target Excess Error

```
**Inputs:** Hypothesis set $V$ of VC dimension $d$, Example distribution $\Delta$, Labeling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon}$, target confidence $\tilde{\delta}$.
**for** $j = 1, 2, \ldots$ **do**
    Draw $n_j = 2^j$ i.i.d examples from $\Delta$; query their labels from $\mathcal{O}$ to get a labelled dataset $S_j$. Denote $\tilde{\delta}_j := \tilde{\delta}/(j(j+1))$.
    Train an ERM classifier $\hat{h}_j \in V$ over $S_j$.
    Define the set $V_j$ as follows:
        $$V_j = \left\{ h \in V : \text{err}_{S_j}(h) \leq \text{err}_{S_j}(\hat{h}_j) + \frac{\tilde{\epsilon}}{2} + \sigma(n_j, \tilde{\delta}_j) + \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)} \right\}$$
    Where $\sigma(n, \delta) := \frac{16}{n}(2d\ln\frac{2en}{d} + \ln\frac{24}{\delta})$.
    **if** $\sup_{h \in V_j}(\sigma(n_j, \tilde{\delta}_j) + \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)}) \leq \frac{\tilde{\epsilon}}{6}$ **then**
        $j_0 = j$, **break**
**return** $V_{j_0}$.
```

2.3 Confidence-Rated Predictor

Our active learning algorithm uses a confidence-rated predictor with guaranteed error to make its label query decisions. In this section, we provide a novel confidence-rated predictor with guaranteed error. This predictor has optimal coverage in the realizable case, and may be of independent interest. The predictor P receives as input a set $V \subseteq \mathcal{H}$ of hypotheses (which is likely to contain the true risk minimizer), an error guarantee η, and a set of U of unlabelled examples. We consider a *soft prediction algorithm*; so, for each example in U, the predictor P outputs three probabilities that add up to 1 – the probability of predicting 1, -1 and 0. This output is subject to the constraint that the expected disagreement[3] between the ± 1 labels assigned by P and those assigned by any $h \in V$ is at most η, and the goal is to maximize the coverage, or the expected fraction of non-abstentions.

Our key insight is that this problem can be written as a linear program, which is described in Algorithm 3. There are three variables, ξ_i, ζ_i and γ_i, for each unlabelled $z_i \in U$; there are the probabilities with which we predict 1, -1 and 0 on z_i respectively. Constraint (2) ensures that the expected disagreement between the label predicted and any $h \in V$ is no more than η, while the LP objective maximizes the coverage under these constraints. Observe that the LP is always feasible. Although the LP has infinitely many constraints, the number of constraints in Equation (2) distinguishable by U_k is at most $(em/d)^d$, where d is the VC dimension of the hypothesis class $\mathcal{H}$.

The performance of a confidence-rated predictor is measured by its error and coverage. The error of a confidence-rated predictor is the probability with which it predicts the wrong label on an example, while the coverage is its probability of non-abstention. We can show the following guarantee on the performance of the predictor in Algorithm 3.

Theorem 1. *In the realizable case, if the hypothesis set V is the version space with respect to a training set, then $\mathbb{P}_{x\sim U}(P(x) \neq h^*(x), P(x) \neq 0) \leq \eta$. In the non-realizable case, if the hypothesis set V is an $(1-\alpha)$-confidence set for the true risk minimizer h^*, then, w.p $\geq 1-\alpha$, $\mathbb{P}_{x\sim U}(P(x) \neq y, P(x) \neq 0) \leq \mathbb{P}_{x\sim U}(h^*(x) \neq y) + \eta$.*

[3] where the expectation is taken over the random choices made by P

Algorithm 3 Confidence-rated Predictor

```
Inputs: hypothesis set $V$, unlabelled data $U = \{z_1, \dots, z_m\}$, error bound $\eta$.
Solve the linear program:
    $\min \sum_{i=1}^{m} \gamma_i$
    subject to: $\forall i,\ \xi_i + \zeta_i + \gamma_i = 1$
        $\forall h \in V,\ \sum_{i:h(z_i)=1} \zeta_i + \sum_{i:h(z_i)=-1} \xi_i \leq \eta m$        (2)
        $\forall i,\ \xi_i, \zeta_i, \gamma_i \geq 0$
For each $z_i \in U$, output probabilities for predicting $1$, $-1$ and $0$: $\xi_i$, $\zeta_i$, and $\gamma_i$.
```

In the realizable case, we can also show that our confidence rated predictor has optimal coverage. Observe that we cannot directly show optimality in the non-realizable case, as the performance depends on the exact choice of the $(1 - \alpha)$-confidence set.

Theorem 2. *In the realizable case, suppose that the hypothesis set V is the version space with respect to a training set. If P' is any confidence rated predictor with error guarantee η, and if P is the predictor in Algorithm 3, then, the coverage of P is at least much as the coverage of P'.*

3 Performance Guarantees

An essential property of any active learning algorithm is consistency – that it converges to the true risk minimizer given enough labelled examples. We observe that our algorithm is consistent provided we use *any* confidence-rated predictor P with guaranteed error as a subroutine. The consistency of our algorithm is a consequence of Lemmas 3 and 6 and is shown in Theorem 3.

Theorem 3 (Consistency)**.** *Suppose we run Algorithm 1 with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P, target excess error ϵ and target confidence δ. Then with probability $1 - \delta$, the classifier $\hat{h}$ returned by Algorithm 1 satisfies $err_D(\hat{h}) - err_D(h^*(D)) \leq \epsilon$.*

We now establish a label complexity bound for our algorithm; however, this label complexity bound applies only if we use the predictor described in Algorithm 3 as a subroutine.

For any hypothesis set V, data distribution D, and η, define $\mathbf{\Phi}_D(V, \eta)$ to be the minimum abstention probability of a confidence-rated predictor which guarantees that the disagreement between its predicted labels and any $h \in V$ under $D_{\mathcal{X}}$ is at most η.

Formally, $\mathbf{\Phi}_D(V, \eta) = \min\{\mathbb{E}_D \gamma(x) : \mathbb{E}_D[I(h(x) = +1)\zeta(x) + I(h(x) = -1)\xi(x)] \leq \eta \text{ for all } h \in V, \gamma(x) + \xi(x) + \zeta(x) \equiv 1, \gamma(x), \xi(x), \zeta(x) \geq 0\}$. Define $\phi(r, \eta) := \mathbf{\Phi}_D(B_D(h^*, r), \eta)$. The label complexity of our active learning algorithm can be stated as follows.

Theorem 4 (Label Complexity)**.** *Suppose we run Algorithm 1 with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P of Algorithm 3, target excess error ϵ and target confidence δ. Then there exist constants $c_3, c_4 > 0$ such that with probability $1 - \delta$:*
(1) In the realizable case, the total number of labels queried by Algorithm 1 is at most:

$$c_3 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{\lceil \log(1/\epsilon) \rceil - k + 1}{\delta})) \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k}$$

(2) In the agnostic case, the total number of labels queried by Algorithm 1 is at most:

$$c_4 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{\lceil \log(1/\epsilon) \rceil - k + 1}{\delta})) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} (1 + \frac{\nu^*(D)}{\epsilon_k})$$

Comparison. The label complexity of disagreement-based active learning is characterized in terms of the *disagreement coefficient*. Given a radius r, the disagreement coefficent $\theta(r)$ is defined as:

$$\theta(r) = \sup_{r' \geq r} \frac{\mathbb{P}(\text{DIS}(B_D(h^*, r')))}{r'},$$

where for any $V \subseteq \mathcal{H}$, $\text{DIS}(V)$ is the disagreement region of V. As $\mathbb{P}(\text{DIS}(B_D(h^*, r))) \geq \phi(r, 0)$ [13], in our notation, $\theta(r) \geq \sup_{r' \geq r} \frac{\phi(r', 0)}{r'}$.

In the realizable case, the label complexity of disagreement-based active learning is $\tilde{O}(\theta(\epsilon) \cdot \ln(1/\epsilon) \cdot (d \ln \theta(\epsilon) + \ln\ln(1/\epsilon)))$ [20][4]. Our label complexity bound may be simplified to:

$$\tilde{O}\left(\ln\frac{1}{\epsilon} \cdot \sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k} \cdot \left(d \ln\left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k}\right) + \ln\ln\frac{1}{\epsilon}\right)\right),$$

which is essentially the bound of [20] with $\theta(\epsilon)$ replaced by $\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k}$. As enforcing a lower error guarantee requires more abstention, $\phi(r, \eta)$ is a decreasing function of η; as a result,

$$\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k} \leq \theta(\epsilon),$$

and our label complexity is better.

In the agnostic case, [12] provides a label complexity bound of $\tilde{O}(\theta(2\nu^*(D) + \epsilon) \cdot (d\frac{\nu^*(D)^2}{\epsilon^2} \ln(1/\epsilon) + d \ln^2(1/\epsilon)))$ for disagreement-based active-learning. In contrast, by Proposition 1 our label complexity is at most:

$$\tilde{O}\left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu^*(D) + \epsilon_k} \cdot \left(d\frac{\nu^*(D)^2}{\epsilon^2} \ln(1/\epsilon) + d \ln^2(1/\epsilon)\right)\right)$$

Again, this is essentially the bound of [12] with $\theta(2\nu^*(D) + \epsilon)$ replaced by the smaller quantity

$$\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu^*(D) + \epsilon_k},$$

[20] has provided a more refined analysis of disagreement-based active learning that gives a label complexity of $\tilde{O}(\theta(\nu^*(D) + \epsilon)(\frac{\nu^*(D)^2}{\epsilon^2} + \ln\frac{1}{\epsilon})(d \ln \theta(\nu^*(D) + \epsilon) + \ln\ln\frac{1}{\epsilon}))$; observe that their dependence is still on $\theta(\nu^*(D) + \epsilon)$. We leave a more refined label complexity analysis of our algorithm for future work.

An important sub-case of learning from noisy data is learning under the Tsybakov noise conditions [30]. We defer the discussion into the Appendix.

3.1 Case Study: Linear Classification under the Log-concave Distribution

We now consider learning linear classifiers with respect to log-concave data distribution on R^d. In this case, for any r, the disagreement coefficient $\theta(r) \leq O(\sqrt{d} \ln(1/r))$ [4]; however, for any $\eta > 0$, $\frac{\phi(r,\eta)}{r} \leq O(\ln(r/\eta))$ (see Lemma 14 in the Appendix), which is much smaller so long as η/r is not too small. This leads to the following label complexity bounds.

Corollary 1. *Suppose $D_{\mathcal{X}}$ is isotropic and log-concave on R^d, and $\mathcal{H}$ is the set of homogeneous linear classifiers on R^d. Then Algorithm 1 with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P of Algorithm 3, target excess error ϵ and target confidence δ satisfies the following properties. With probability $1 - \delta$:*
(1) In the realizable case, there exists some absolute constant $c_8 > 0$ such that the total number of labels queried is at most $c_8 \ln\frac{1}{\epsilon}(d + \ln\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})$.
(2) In the agnostic case, there exists some absolute constant $c_9 > 0$ such that the total number of labels queried is at most $c_9(\frac{\nu^*(D)^2}{\epsilon^2} + \ln\frac{1}{\epsilon}) \ln\frac{\epsilon + \nu^*(D)}{\epsilon}(d \ln\frac{\epsilon + \nu^*(D)}{\epsilon} + \ln\frac{1}{\delta}) + \ln\frac{1}{\epsilon} \ln\frac{\epsilon + \nu^*(D)}{\epsilon} \ln\ln\frac{1}{\epsilon}$.

[4] Here the $\tilde{O}()$ notation hides factors logarithmic in $1/\delta$

(3) If (C_0, κ)-Tsybakov Noise condition holds for D with respect to $\mathcal{H}$, then there exists some constant $c_{10} > 0$ (that depends on C_0, κ) such that the total number of labels queried is at most $c_{10}\epsilon^{\frac{2}{\kappa}-2} \ln \frac{1}{\epsilon}(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})$.

In the realizable case, our bound matches [4]. For disagreement-based algorithms, the bound is $\tilde{O}(d^{\frac{3}{2}} \ln^2 \frac{1}{\epsilon}(\ln d + \ln \ln \frac{1}{\epsilon}))$, which is worse by a factor of $O(\sqrt{d} \ln(1/\epsilon))$. [4] does not address the fully agnostic case directly; however, if $\nu^*(D)$ is known a-priori, then their algorithm can achieve roughly the same label complexity as ours.

For the Tsybakov Noise Condition with $\kappa > 1$, [3, 4] provides a label complexity bound for $\tilde{O}(\epsilon^{\frac{2}{\kappa}-2} \ln^2 \frac{1}{\epsilon}(d + \ln \ln \frac{1}{\epsilon}))$ with an algorithm that has a-priori knowledge of C_0 and κ. We get a slightly better bound. On the other hand, a disagreement based algorithm [20] gives a label complexity of $\tilde{O}(d^{\frac{3}{2}} \ln^2 \frac{1}{\epsilon}\epsilon^{\frac{2}{\kappa}-2}(\ln d + \ln \ln \frac{1}{\epsilon}))$. Again our bound is better by factor of $\Omega(\sqrt{d})$ over disagreement-based algorithms. For $\kappa = 1$, we can tighten our label complexity to get a $\tilde{O}(\ln \frac{1}{\epsilon}(d + \ln \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}))$ bound, which again matches [4], and is better than the ones provided by disagreement-based algorithm – $\tilde{O}(d^{\frac{3}{2}} \ln^2 \frac{1}{\epsilon}(\ln d + \ln \ln \frac{1}{\epsilon}))$ [20].

4 Related Work

Active learning has seen a lot of progress over the past two decades, motivated by vast amounts of unlabelled data and the high cost of annotation [28, 10, 20]. According to [10], the two main threads of research are exploitation of cluster structure [31, 11], and efficient search in hypothesis space, which is the setting of our work. We are given a hypothesis class $\mathcal{H}$, and the goal is to find an $h \in \mathcal{H}$ that achieves a target excess generalization error, while minimizing the number of label queries.

Three main approaches have been studied in this setting. The first and most natural one is generalized binary search [17, 8, 9, 27], which was analyzed in the realizable case by [9] and in various limited noise settings by [23, 27, 26]. While this approach has the advantage of low label complexity, it is generally inconsistent in the fully agnostic setting [11]. The second approach, disagreement-based active learning, is consistent in the agnostic PAC model. [7] provides the first disagreement-based algorithm for the realizable case. [2] provides an agnostic disagreement-based algorithm, which is analyzed in [18] using the notion of disagreement coefficient. [12] reduces disagreement-based active learning to passive learning; [5] and [6] further extend this work to provide practical and efficient implementations. [19, 24] give algorithms that are adaptive to the Tsybakov Noise condition. The third line of work [3, 4, 1], achieves a better label complexity than disagreement-based active learning for linear classifiers on the uniform distribution over unit sphere and logconcave distributions. However, a limitation is that their algorithm applies only to these specific settings, and it is not apparent how to apply it generally.

Research on confidence-rated prediction has been mostly focused on empirical work, with relatively less theoretical development. Theoretical work on this topic includes KWIK learning [25], conformal prediction [29] and the weighted majority algorithm of [16]. The closest to our work is the recent learning-theoretic treatment by [13, 14]. [13] addresses confidence-rated prediction with guaranteed error in the realizable case, and provides a predictor that abstains in the disagreement region of the version space. This predictor achieves zero error, and coverage equal to the measure of the agreement region. [14] shows how to extend this algorithm to the non-realizable case and obtain zero error with respect to the best hypothesis in $\mathcal{H}$. Note that the predictors in [13, 14] generally achieve less coverage than ours for the same error guarantee; in fact, if we plug them into our Algorithm 1, then we recover the label complexity bounds of disagreement-based algorithms [12, 19, 24].

A formal connection between disagreement-based active learning in realizable case and perfect confidence-rated prediction (with a zero error guarantee) was established by [15]. Our work can be seen as a step towards bridging these two areas, by demonstrating that active learning can be further reduced to imperfect confidence-rated prediction, with potentially higher label savings.

Acknowledgements. We thank NSF under IIS-1162581 for research support. We thank Sanjoy Dasgupta and Yoav Freund for helpful discussions. We thank Steve Hanneke for pointing out a mistake on an initial version of the paper. CZ would like to thank Liwei Wang for introducing the problem of selective classification to him.

References

[1] P. Awasthi, M-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *STOC*, 2014.

[2] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.

[3] M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.

[4] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[5] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

[6] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

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

[8] S. Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, 2004.

[9] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[10] S. Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19), 2011.

[11] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *ICML*, 2008.

[12] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

[13] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *JMLR*, 2010.

[14] R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *NIPS*, 2011.

[15] R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *JMLR*, 2012.

[16] Y. Freund, Y. Mansour, and R. E. Schapire. Generalization bounds for averaged classifiers. *The Ann. of Stat.*, 32, 2004.

[17] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[18] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[19] S. Hanneke. Adaptive rates of convergence in active learning. In *COLT*, 2009.

[20] S. Hanneke. A statistical theory of active learning. Manuscript, 2013.

[21] S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *CoRR*, abs/1207.3772, 2012.

[22] D. Hsu. *Algorithms for Active Learning*. PhD thesis, UC San Diego, 2010.

[23] M. Kääriäinen. Active learning in the non-realizable case. In *ALT*, 2006.

[24] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *JMLR*, 2010.

[25] L. Li, M. L. Littman, and T. J. Walsh. Knows what it knows: a framework for self-aware learning. In *ICML*, 2008.

[26] M. Naghshvar, T. Javidi, and K. Chaudhuri. Noisy bayesian active learning. In *Allerton*, 2013.

[27] R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

[28] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

[29] G. Shafer and V. Vovk. A tutorial on conformal prediction. *JMLR*, 2008.

[30] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

[31] R. Urner, S. Wulff, and S. Ben-David. Plal: Cluster-based active learning. In *COLT*, 2013.

A Tsybakov Noise Conditions

An important sub-case of learning from noisy data is learning under the Tsybakov noise conditions [30].

Definition 1. *(Tsybakov Noise Condition) Let $\kappa \geq 1$. A labelled data distribution D over $\mathcal{X} \times \mathcal{Y}$ satisfies (C_0, κ)-Tsybakov Noise Condition with respect to a hypothesis class $\mathcal{H}$ for some constant $C_0 > 0$, if for all $h \in \mathcal{H}$, $\rho_D(h, h^*(D)) \leq C_0(err_D(h) - err_D(h^*(D)))^{\frac{1}{\kappa}}$.*

The following theorem shows the performance guarantees achieved by Algorithm 1 under the Tsybakov noise conditions.

Theorem 5. *Suppose (C_0, κ)-Tsybakov Noise Condition holds for D with respect to $\mathcal{H}$. Then Algorithm 1 with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class $\mathcal{H}$, confidence-rated predictor P of Algorithm 3, target excess error ϵ and target confidence δ satisfies the following properties. There exists a constant $c_5 > 0$ such that with probability $1 - \delta$, the total number of labels queried by Algorithm 1 is at most:*

$$c_5 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \epsilon_k/256)\epsilon_k^{\frac{1}{\kappa}-2}) + \ln(\frac{\lceil \log \frac{1}{\epsilon} \rceil - k + 1}{\delta}))\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \epsilon_k/256)\epsilon_k^{\frac{1}{\kappa}-2}$$

Comparison. [20] provides a label complexity bound of $\tilde{O}(\theta(C_0\epsilon^{\frac{1}{\kappa}})\epsilon^{\frac{2}{\kappa}-2} \ln \frac{1}{\epsilon}(d \ln \theta(C_0 \epsilon^{\frac{1}{\kappa}}) + \ln \ln \frac{1}{\epsilon}))$ for disagreement-based active learning. For $\kappa > 1$, by Proposition 2, our label complexity is at most:

$$\tilde{O}\left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k^{1/\kappa}, \epsilon_k/256)}{\epsilon_k^{1/\kappa}} \cdot \epsilon_k^{2/\kappa - 2} \cdot d \ln(1/\epsilon) \right),$$

For $\kappa = 1$, our label complexity is at most

$$\tilde{O}\left(\ln \frac{1}{\epsilon} \cdot \sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k, \epsilon_k/256)}{\epsilon_k} \cdot \left(d \ln(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k, \epsilon_k/256)}{\epsilon_k}) + \ln \ln \frac{1}{\epsilon} \right) \right).$$

In both cases, our bounds are better, as $\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \cdot \frac{\phi(C_0 \epsilon_k^{1/\kappa}, \epsilon_k/256)}{C_0 \epsilon_k^{1/\kappa}} \leq \theta(C_0 \epsilon^{1/\kappa})$. In further work, [21] provides a refined analysis with a bound of $\tilde{O}(\theta(C_0 \epsilon^{\frac{1}{\kappa}})\epsilon^{\frac{2}{\kappa}-2} \; d \ln \theta(C_0 \epsilon^{\frac{1}{\kappa}}))$; however, this work is not directly comparable to ours, as they need prior knowledge of C_0 and κ.

B Additional Notation and Concentration Lemmas

We begin with some additional notation that will be used in the subsequent proofs. Recall that we define:

$$\sigma(n, \delta) = \frac{16}{n}(2d \ln \frac{2en}{d} + \ln \frac{24}{\delta}), \tag{3}$$

where d is the VC dimension of the hypothesis class $\mathcal{H}$.

The following lemma is an immediate corollary of the multiplicative VC bound; we pick the version of the multiplicative VC bound due to [22].

Lemma 7. *Pick any $n \geq 1$, $\delta \in (0, 1)$. Let S_n be a set of n iid copies of (X, Y) drawn from a distribution D over labelled examples. Then, the following hold with probability at least $1 - \delta$ over the choice of S_n:*
(1) For all $h \in \mathcal{H}$,

$$|err_D(h) - err_{S_n}(h)| \leq \min(\sigma(n, \delta) + \sqrt{\sigma(n, \delta) err_D(h)}, \sigma(n, \delta) + \sqrt{\sigma(n, \delta) err_{S_n}(h)}) \tag{4}$$

In particular, all classifiers h in $\mathcal{H}$ consistent with S_n satisfies

$$err_D(h) \leq \sigma(n, \delta) \tag{5}$$

(2) For all h, h' in $\mathcal{H}$,

$$|(err_D(h) - err_D(h')) - (err_{S_n}(h) - err_{S_n}(h'))| \leq \sigma(n,\delta) + \min(\sqrt{\sigma(n,\delta)\rho_D(h,h')}, \sqrt{\sigma(n,\delta)\rho_{S_n}(h,h')}) \tag{6}$$

$$|\rho_D(h,h') - \rho_{S_n}(h,h')| \leq \sigma(n,\delta) + \min(\sqrt{\sigma(n,\delta)\rho_D(h,h')}, \sqrt{\sigma(n,\delta)\rho_{S_n}(h,h')}) \tag{7}$$

Where $\sigma(n,\delta)$ is defined in Equation (3).

We occasionally use the following (weaker) version of Lemma 7.

Lemma 8. *Pick any $n \geq 1$, $\delta \in (0,1)$. Let S_n be a set of n iid copies of (X,Y). The following holds with probability at least $1-\delta$: (1) For all $h \in \mathcal{H}$,*

$$|err_D(h) - err_{S_n}(h)| \leq \sqrt{4\sigma(n,\delta)} \tag{8}$$

(2) For all h, h' in $\mathcal{H}$,

$$|(err_D(h) - err_D(h')) - (err_{S_n}(h) - err_{S_n}(h'))| \leq \sqrt{4\sigma(n,\delta)} \tag{9}$$

$$|\rho_D(h,h') - \rho_{S_n}(h,h')| \leq \sqrt{4\sigma(n,\delta)} \tag{10}$$

Where $\sigma(n,\delta)$ is defined in Equation (3).

For an unlabelled sample U_k, we use $\tilde{U}_k$ to denote the joint distribution over $\mathcal{X} \times \mathcal{Y}$ induced by uniform distribution over U_k and $D_{Y|X}$. We have:

Lemma 9. *If the size of n_k of the unlabelled dataset U_k is at least $192(\frac{512}{\epsilon_k})^2(d \ln 192(\frac{512}{\epsilon_k})^2 + \ln \frac{288}{\delta_k})$, then with probability $1 - \delta_k/4$, the following conditions hold for all $h, h' \in V_k$:*

$$|err_D(h) - err_{\tilde{U}_k}(h)| \leq \frac{\epsilon_k}{64} \tag{11}$$

$$|(err_D(h) - err_D(h')) - (err_{\tilde{U}_k}(h) - err_{\tilde{U}_k}(h'))| \leq \frac{\epsilon_k}{32} \tag{12}$$

$$|\rho_D(h,h') - \rho_{\tilde{U}_k}(h,h')| \leq \frac{\epsilon_k}{64} \tag{13}$$

Lemma 10. *If the size of n_k of the unlabelled dataset U_k is at least $192(\frac{512}{\epsilon_k})^2(d \ln 192(\frac{512}{\epsilon_k})^2 + \ln \frac{288}{\delta_k})$, then with probability $1 - \delta_k/4$, the following hold:*
(1) The outputs $\{(\xi_{k,i}, \zeta_{k,i}, \gamma_{k,i})\}_{i=1}^{n_k}$ of any confidence-rated predictor with inputs hypothesis set V_k, unlabelled data U_k, and error bound $\epsilon_k/64$ satisfy:

$$\frac{1}{n_k}\sum_{i=1}^{n_k}[I(h(x_i) \neq h'(x_i))(1-\gamma_{k,i})] \leq \frac{\epsilon_k}{32}; \tag{14}$$

(2) The outputs $\{(\xi_{k,i}, \zeta_{k,i}, \gamma_{k,i})\}_{i=1}^{n_k}$ of the confidence-rated predictor of Algortihm 3 with inputs hypothesis set V_k, unlabelled data U_k, and error bound $\epsilon_k/64$ satisfy:

$$\phi_k \leq \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \tag{15}$$

We use $\tilde{\Gamma}_k$ to denote the joint distribution over $\mathcal{X} \times \mathcal{Y}$ induced by Γ_k and $D_{Y|X}$. Denote $\gamma_k(x) : \mathcal{X} \to [0,1]$, where $\gamma_k(x_i) = \gamma_{k,i}$, and 0 elsewhere. Clearly, $\Gamma_k(\{x\}) = \frac{\gamma_k(x)}{n_k\phi_k}$ and $\tilde{\Gamma}_k(\{(x,y)\}) = \frac{\tilde{U}_k(\{(x,y)\})\gamma_k(x)}{\phi_k}$. Also, Equations (14) and (15) of Lemma 10 can be restated as

$$\forall h, h' \in V_k, \mathbb{E}_{\tilde{U}_k}[(1-\gamma_k(x))I(h(x) \neq h'(x))] \leq \frac{\epsilon_k}{32}$$

$$\mathbb{E}_{\tilde{U}_k}[\gamma_k(x)] = \phi_k \leq \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256}$$

In the realizable case, define event

$$E_r = \{\text{For all } k = 1, 2, \ldots, k_0\text{: Equations (11), (12), (13), (14), (15) hold for } \tilde{U}_k$$
$$\text{and all classifiers consistent with } S_k \text{ have error at most } \frac{\epsilon_k}{8\phi_k} \text{ with respect to } \tilde{\Gamma}_k\ \}.$$

Fact 1. $\mathbb{P}(E_r) \geq 1 - \delta$.

Proof. By Equation (5) of Lemma 7, with probability $1 - \delta_k/2$, if $h \in V_k$ is consistent with S_k, then

$$\mathrm{err}_{\tilde{\Gamma}_k}(h) \leq \sigma(m_k, \delta_k/2)$$

Because $m_k = \frac{1536\phi_k}{\epsilon_k}(d \ln \frac{1536\phi_k}{\epsilon_k} + \ln \frac{48}{\delta_k})$, we have $\mathrm{err}_{\tilde{\Gamma}_k}(h) \leq \epsilon_k/8\phi_k$. The fact follows from combining the fact above with Lemma 9 and Lemma 10, and the union bound. □

In the non-realizable case, define event

$E_a = \{$For all $k = 1, 2, \ldots, k_0$: Equations (11), (12), (13), (14), (15) hold for $\tilde{U}_k$,
and Algorithm 2 succeeds with inputs hypothesis set $V = V_k$, example distribution $\Delta = \Gamma_k$,
labelling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon} = \dfrac{\epsilon_k}{8\phi_k}$ and target confidence $\tilde{\delta} = \dfrac{\delta_k}{2}\}$.

Fact 2. $\mathbb{P}(E_a) \geq 1 - \delta$.

Proof. This is an immediate consequence of Lemma 9, Lemma 10, Lemma 4 and union bound. □

Recall that we assume the hypothesis space is "dense", in the sense that $\forall r > 0$, $\sup_{h \in B_D(h^*(D), r)} \rho(h, h^*(D)) = r$. We will call this the "denseness assumption".

C Proofs related to the properties of Algorithm 2

We first establish some properties of Algorithm 2. The inputs to Algorithm 2 are a set V of hypotheses of VC dimension d, an example distribution Δ, a labeling oracle $\mathcal{O}$, a target excess error $\tilde{\epsilon}$ and a target confidence $\tilde{\delta}$.

We define the event

$\tilde{E} = \{$For all $j = 1, 2, \ldots$: Equations (4)-(7) hold for sample S_j with $n = n_j$ and $\delta = \tilde{\delta}_j$ $\}$

By union bound, $\mathbb{P}(\tilde{E}) \geq 1 - \sum_j \tilde{\delta}_j \geq 1 - \tilde{\delta}$.

Proof. (of Lemma 4) Assume $\tilde{E}$ happens. For the proof of (1), define j_{max} as the smallest integer j such that $\sigma(n_j, \tilde{\delta}_j) \leq \tilde{\epsilon}^2/144$. Since $n_{j_{max}}$ is a power of 2,

$$n_{j_{max}} \leq 2 \min\{n = 1, 2, \ldots : \frac{16(2d \ln \frac{2en}{d} + \ln \frac{24 \log n (\log n + 1)}{\delta})}{n} \leq \frac{\tilde{\epsilon}^2}{144}\}$$

Thus, $n_{j_{max}} \leq 384 \frac{144}{\tilde{\epsilon}^2}(d \ln 192 \frac{144}{\tilde{\epsilon}^2} + \ln \frac{24}{\tilde{\delta}})$. Then in round j_{max}, the stopping criterion (6) of Algorithm 2 is satisified; thus, Algorithm 2 halts with $j_0 \leq j_{max}$.

To prove (2.1), we observe that as $h^*(\tilde{\Delta})$ is the risk minimizer in V, if h satisfies $\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \frac{\tilde{\epsilon}}{2}$, then $\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(\hat{h}_{j_0}) \leq \frac{\tilde{\epsilon}}{2}$. By Equation (6) of Lemma 7,

$$\begin{aligned}(\mathrm{err}_{S_{j_0}}(h) - \mathrm{err}_{S_{j_0}}(\hat{h}_{j_0})) &\leq (\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(\hat{h}_{j_0})) + \sigma(n_{j_0}, \tilde{\delta}_{j_0}) + \sqrt{\sigma(n_{j_0}, \tilde{\delta}_{j_0}) \rho_{S_{j_0}}(h, \hat{h}_{j_0})} \\ &\leq \frac{\tilde{\epsilon}}{2} + \sigma(n_{j_0}, \tilde{\delta}_{j_0}) + \sqrt{\sigma(n_{j_0}, \tilde{\delta}_{j_0}) \rho_{S_{j_0}}(h, \hat{h}_{j_0})}\end{aligned}$$

Hence $h \in V_{j_0}$.

For the proof of (2.2), note first that by (2.1), in particular, $h^*(\tilde{\Delta}) \in V_{j_0}$. Hence by Equation (6) of Lemma 7, and the stopping criterion Equation (6),

$$(\mathrm{err}_{\tilde{\Delta}}(\hat{h}_{j_0}) - \mathrm{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta}))) - (\mathrm{err}_{S_{j_0}}(\hat{h}_{j_0}) - \mathrm{err}_{S_{j_0}}(h^*(\tilde{\Delta}))) \leq \sigma(n_{j_0}, \tilde{\delta}_{j_0}) + \sqrt{\sigma(n_{j_0}, \tilde{\delta}_{j_0}) \rho_{S_{j_0}}(\hat{h}_{j_0}, h^*(\tilde{\Delta}))} \leq \frac{\tilde{\epsilon}}{6}$$

Thus,

$$\text{err}_{\tilde{\Delta}}(\hat{h}_{j_0}) - \text{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \frac{\tilde{\epsilon}}{6} \tag{16}$$

On the other hand, if $h \in V_{j_0}$, then

$$(\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\hat{h}_{j_0})) - (\text{err}_{S_{j_0}}(h) - \text{err}_{S_{j_0}}(\hat{h}_{j_0})) \leq \sigma(n_{j_0}, \tilde{\delta}_{j_0}) + \sqrt{\sigma(n_{j_0}, \tilde{\delta}_{j_0})\rho_{S_{j_0}}(h, \hat{h}_{j_0})} \leq \frac{\tilde{\epsilon}}{6}$$

By definition of V_{j_0},

$$(\text{err}_{S_{j_0}}(h) - \text{err}_{S_{j_0}}(\hat{h}_{j_0})) \leq \sigma(n_{j_0}, \tilde{\delta}_{j_0}) + \sqrt{\sigma(n_{j_0}, \tilde{\delta}_{j_0})\rho_{S_{j_0}}(h, \hat{h}_{j_0})} + \frac{\tilde{\epsilon}}{2} \leq \frac{2\tilde{\epsilon}}{3}$$

Hence,

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\hat{h}_{j_0}) \leq \frac{5\tilde{\epsilon}}{6} \tag{17}$$

Combining Equations (16) and (17), we have

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \tilde{\epsilon}$$

□

Proof. (of Lemma 5) Assume $\tilde{E}$ happens. For each j, by triangle inequality, we have that $\rho_{S_j}(\hat{h}_j, h) \leq \text{err}_{S_j}(\hat{h}_j) + \text{err}_{S_j}(h)$. If $h \in V_j$, then, by defintion of V_j,

$$\text{err}_{S_j}(h) - \text{err}_{S_j}(\hat{h}_j) \leq \frac{\tilde{\epsilon}}{2} + \sigma(n_j, \tilde{\delta}_j) + \sqrt{\sigma(n_j, \tilde{\delta}_j)\text{err}_{S_j}(\hat{h}_j)} + \sqrt{\sigma(n_j, \tilde{\delta}_j)\text{err}_{S_j}(h)}$$

Using the fact that $A \leq B + C\sqrt{A} \Rightarrow A \leq 2B + C^2$,

$$\text{err}_{S_j}(h) \leq \tilde{\epsilon} + 2\text{err}_{S_j}(\hat{h}_j) + 2\sqrt{\sigma(n_j, \tilde{\delta}_j)\text{err}_{S_j}(\hat{h}_j)} + 3\sigma(n_j, \tilde{\delta}_j) \leq 3\text{err}_{S_j}(\hat{h}_j) + 4\sigma(n_j, \tilde{\delta}_j) + \tilde{\epsilon}$$

Since

$$\text{err}_{S_j}(\hat{h}_j) \leq \text{err}_{S_j}(h^*(\tilde{\Delta})) \leq \nu^*(\tilde{\Delta}) + \sqrt{\sigma(n_j, \tilde{\delta}_j)\nu^*(\tilde{\Delta})} + \sigma(n_j, \tilde{\delta}_j) \leq 2\nu^*(\tilde{\Delta}) + 2\sigma(n_j, \tilde{\delta}_j),$$

by the triangle inequality, we get that for all $h \in V_j$,

$$\rho_{S_j}(h, \hat{h}_j) \leq \text{err}_{S_j}(h) + \text{err}_{S_j}(\hat{h}_j) \leq 8\nu^*(\tilde{\Delta}) + 12\sigma(n_j, \tilde{\delta}_j) + \tilde{\epsilon} \tag{18}$$

Now observe that for any j,

$$\begin{aligned}
& \sup_{h \in V_j} \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)} + \sigma(n_j, \tilde{\delta}_j) \\
\leq \quad & \sup_{h \in V_j} \max(2\sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)}, 2\sigma(n_j, \tilde{\delta}_j)) \\
\leq \quad & \max(2\sqrt{(8\nu^*(\tilde{\Delta}) + 12\sigma(n_j, \tilde{\delta}_j) + \tilde{\epsilon})\sigma(n_j, \tilde{\delta}_j)}, 2\sigma(n_j, \tilde{\delta}_j)) \\
\leq \quad & \max(12\sqrt{2\nu^*(\tilde{\Delta})\sigma(n_j, \tilde{\delta}_j)}, \tilde{\epsilon}/6, 216\sigma(n_j, \tilde{\delta}_j)),
\end{aligned}$$

Where the first inequality follows from $A + B \leq 2\max(A, B)$, the second inequality follows from Equation (18), the third inequality follows from $\sqrt{A + B} \leq \sqrt{A} + \sqrt{B}$, $A + B + C \leq 3\max(A, B, C)$ and $\sqrt{AB} \leq \max(A, B)$.

It can be easily seen that there exists some constant $c_1 > 0$, such that taking $j_1 = \lceil \log\left(\frac{c_1}{2}(d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})(\frac{\nu^*(\tilde{\Delta}) + \tilde{\epsilon}}{\tilde{\epsilon}^2})\right) \rceil$ ensures that $n_{j_1} \geq \frac{c_1}{2}(d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})(\frac{\nu^*(\tilde{\Delta}) + \tilde{\epsilon}}{\tilde{\epsilon}^2})$; this, in turn, suffices to make

$$\max(12\sqrt{2\nu^*(\tilde{\Delta})\sigma(n_j, \tilde{\delta}_j)}, 216\sigma(n_j, \tilde{\delta}_j)) \leq \tilde{\epsilon}/6$$

Hence the stopping criterion $\sup_{h \in V_j} \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)} + \sigma(n_j, \tilde{\delta}_j) \leq \tilde{\epsilon}/6$ is satisfied in iteration j_1, and Algorithm 2 exits at iteration $j_0 \leq j_1$, which ensures that $n_{j_0} \leq n_{j_1} \leq c_1(d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})(\frac{\nu^*(\tilde{\Delta}) + \tilde{\epsilon}}{\tilde{\epsilon}^2})$. □

The following lemma examines the behavior of Algorithm 2 under the Tsybakov Noise Condition and is crucial in the proof of Theorem 5. We observe that even if the (C_0, κ)-Tsybakov Noise Conditions hold with respect to D, they do not necessarily hold with respect to Γ_k. In particular, it is not necessarily true that:

$$\rho_{\tilde{\Gamma}_k}(h, h^*(D)) \leq C_0(\text{err}_{\tilde{\Gamma}_k}(h) - \text{err}_{\tilde{\Gamma}_k}(h^*(D)))^{\frac{1}{\kappa}}, \forall h \in V_k$$

However, we show that an "approximate" Tsybakov Noise Condition with a significantly larger "C_0", namely Condition (19) is met by $\tilde{\Gamma}_k$ and V_k, with $C = \max(8C_0, 4)\phi_k^{\frac{1}{\kappa}-1}$ and $\tilde{h} = h^*(D)$. In the Lemma below, we carefully track the dependence of the number of our label queries on C, since $C = \max(8C_0, 4)\phi_k^{\frac{1}{\kappa}-1}$ can be $\omega(1)$ in our particular application.

Lemma 11. *Suppose we run Algorithm 2 with inputs hypothesis set V, example distribution $\tilde{\Delta}$, labelling oracle $\mathcal{O}$, excess generalization error $\tilde{\epsilon}$ and confidence $\tilde{\delta}$. Then there exists some absolute constant $c_2 > 0$ (independent of C) such that the following holds. Suppose there exist $C > 0$ and a classifier $\tilde{h} \in V$, such that*

$$\forall h \in V, \rho_{\tilde{\Delta}}(h, \tilde{h}) \leq C \max(\tilde{\epsilon}, err_{\tilde{\Delta}}(h) - err_{\tilde{\Delta}}(\tilde{h}))^{\frac{1}{\kappa}}, \tag{19}$$

where $\tilde{\epsilon}$ is the target exccess error parameter in Algorithm 2. Then, on the event that Algorithm 2 succeeds,

$$n_{j_0} \leq c_2 \max((d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})\tilde{\epsilon}^{-1}, (d \ln(C\tilde{\epsilon}^{\frac{1}{\kappa}-2}) + \ln \frac{1}{\tilde{\delta}})C\tilde{\epsilon}^{\frac{1}{\kappa}-2})$$

Observe that Condition (19), the approximate Tsybakov Noise Condition in the statement of Lemma 11, is with respect to $\tilde{h}$, which is not necessarily the true risk minimizer in V with respect to $\tilde{\Delta}$. We therefore prove Lemma 11 in three steps; first, in Lemma 12, we analyze the difference $\text{err}_{\tilde{\Delta}}(\hat{h}) - \text{err}_{\tilde{\Delta}}(\tilde{h})$, where $\hat{h}$ is the empirical risk minimizer. Then, in Lemma 13, we bound the difference $\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h})$ for any $h \in V_j$ for some j. Finally, we combine these two lemmas to provide sample complexity bounds for the V_{j_0} output by Algorithm 2.

Proof. (of Lemma 11) Assume the event $\tilde{E}$ happens. Then,

Consider iteration j, by Lemma 13, if $h \in V_j$, then

$$\rho_{\tilde{\Delta}}(h, \hat{h}_j) \leq \rho_{\tilde{\Delta}}(h, \tilde{h}) + \rho_{\tilde{\Delta}}(\hat{h}_j, \tilde{h}) \leq \max(2C(36\tilde{\epsilon})^{\frac{1}{\kappa}}, 2C(52\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{\kappa}}, 2C(6400C\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2\kappa-1}}). \tag{20}$$

We can write:

$$\begin{aligned}
\sup_{h \in V_j} \sigma(n_j, \tilde{\delta}_j) + \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)} &\leq \sup_{h \in V_j} 3\sigma(n_j, \tilde{\delta}_j) + \sqrt{2\sigma(n_j, \tilde{\delta}_j)\rho_{\tilde{\Delta}}(h, \hat{h}_j)} \\
&\leq \sup_{h \in V_j} \max(6\sigma(n_j, \tilde{\delta}_j), 2\sqrt{2\sigma(n_j, \tilde{\delta}_j)\rho_{\tilde{\Delta}}(h, \hat{h}_j)}),
\end{aligned}$$

where the first inequality follows from Equation (23) and the second inequality follows $A + B \leq 2\max(A, B)$. We can further use Equation (20) to show that this is at most:

$$\begin{aligned}
&\leq \max(6\sigma(n_j, \tilde{\delta}_j), (16C\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2}}(36\tilde{\epsilon})^{\frac{1}{2\kappa}}, (16C\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2}}(52\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2\kappa}}, (6400C\sigma(n_j, \tilde{\delta}_j))^{\frac{\kappa}{2\kappa-1}}) \\
&\leq \max(6\sigma(n_j, \tilde{\delta}_j), \tilde{\epsilon}/6, (6400C\sigma(n_j, \tilde{\delta}_j))^{\frac{\kappa}{2\kappa-1}})
\end{aligned}$$

Here the last inequality follows from the fact that $(16C\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2}}(36\tilde{\epsilon})^{\frac{1}{2\kappa}} \leq \max((3456C\sigma(n_j, \tilde{\delta}_j))^{\frac{\kappa}{2\kappa-1}}, \tilde{\epsilon}/6)$ and $(16C\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2}}(52\sigma(n_j, \tilde{\delta}_j))^{\frac{1}{2\kappa}} \leq \max((144C\sigma(n_j, \tilde{\delta}_j))^{\frac{\kappa}{2\kappa-1}}, 6\sigma(n_j, \tilde{\delta}_j))$, since $A^{\frac{2\kappa-1}{2\kappa}}B^{\frac{1}{2\kappa}} \leq \max(A, B)$.

It can be easily seen that there exists $c_2 > 0$, such that taking $j_1 = \lceil \log \frac{c_2}{2}(d \ln \frac{\max(C,1)}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})(C\tilde{\epsilon}^{\frac{1}{\kappa}-2} + \tilde{\epsilon}^{-1}) \rceil$, so that $n_j \geq \frac{c_2}{2}(d \ln \frac{\max(C,1)}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})(C\tilde{\epsilon}^{\frac{1}{\kappa}-2} + \tilde{\epsilon}^{-1})$ suffices to make

$$\max(6\sigma(n_j, \tilde{\delta}_j), (6400C\sigma(n_j, \tilde{\delta}_j))^{\frac{\kappa}{2\kappa-1}}) \leq \tilde{\epsilon}/6$$

Hence the stopping criterion $\sup_{h \in V_j} \sqrt{\sigma(n_j, \tilde{\delta}_j)\rho_{S_j}(h, \hat{h}_j)} + \sigma(n_j, \tilde{\delta}_j) \leq \tilde{\epsilon}/6$ is satisfied in iteration j_1. Thus the number of the exit iteration j_0 satisfies $j_0 \leq j_1$, and $n_{j_0} \leq n_{j_1} \leq c_2 \max((d \ln \frac{1}{\tilde{\epsilon}} + \ln \frac{1}{\tilde{\delta}})\tilde{\epsilon}^{-1}, (d \ln(C\tilde{\epsilon}^{\frac{1}{\kappa}-2}) + \ln \frac{1}{\tilde{\delta}})C\tilde{\epsilon}^{\frac{1}{\kappa}-2})$.

□

Lemma 12. *Suppose there exist $C > 0$ and a classifier $\tilde{h} \in V$, such that Equation* (19) *holds. Suppose we draw a set S of n examples, denote the empirical risk minimizer over S as $\hat{h}$, then with probability $1 - \delta$:*

$$\mathit{err}_{\tilde{\Delta}}(\hat{h}) - \mathit{err}_{\tilde{\Delta}}(\tilde{h}) \leq \max(2\sigma(n,\delta), (4C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}}, 2\tilde{\epsilon})$$

$$\rho_{\tilde{\Delta}}(\hat{h}, \tilde{h}) \leq \max(C(2\sigma(n,\delta))^{\frac{1}{\kappa}}, C(4C\sigma(n,\delta))^{\frac{1}{2\kappa-1}}, C(2\tilde{\epsilon})^{\frac{1}{\kappa}})$$

Proof. By Lemma 7, with probability $1 - \delta$, Equation (6) holds. Assume this happens.

$$\begin{aligned}
& \mathrm{err}_{\tilde{\Delta}}(\hat{h}) - \mathrm{err}_{\tilde{\Delta}}(\tilde{h}) \\
\leq \quad & \sigma(n,\delta) + \sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(\hat{h}, \tilde{h})} \\
\leq \quad & 2\max(\sigma(n,\delta), \sqrt{\sigma(n,\delta)C(\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(\tilde{h})^{\frac{1}{\kappa}})}, \sqrt{\sigma(n,\delta)C\tilde{\epsilon}^{\frac{1}{\kappa}}}) \\
\leq \quad & \max(2\sigma(n,\delta), (4C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}}, 2\tilde{\epsilon})
\end{aligned}$$

Where the first inequality is by Equation (6) of Lemma 7; the second inequality follow from Equation (19) and $A + B \leq 2\max(A, B)$. The third inequality follows from $2\sqrt{\sigma(n,\delta)C\tilde{\epsilon}^{\frac{1}{\kappa}}} \leq \max(2(C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}}, 2\tilde{\epsilon})$, since $A^{\frac{2\kappa-1}{2\kappa}}B^{\frac{1}{2\kappa}} \leq \max(A, B)$. As a consequence, by Equation (19),

$$\rho_{\tilde{\Delta}}(\hat{h}, \tilde{h}) \leq \max(C(2\sigma(n,\delta))^{\frac{1}{\kappa}}, C(4C\sigma(n,\delta))^{\frac{1}{2\kappa-1}}, C(2\tilde{\epsilon})^{\frac{1}{\kappa}})$$

□

Lemma 13. *Suppose there exist a $C > 0$ and a classifier $\tilde{h} \in V$ such that Equation* (19) *holds. Suppose we draw a set S of n iid examples, and let $\hat{h}$ denote the empirical risk minimizer over S. Moreover, we define:*

$$\tilde{V} = \left\{h \in V : \mathit{err}_S(h) \leq \mathit{err}_S(\hat{h}) + \frac{\tilde{\epsilon}}{2} + \sigma(n,\delta) + \sqrt{\sigma(n,\delta)\rho_S(h, \hat{h})}\right\}$$

then with probability $1 - \delta$, for all $h \in \tilde{V}$,

$$\mathit{err}_{\tilde{\Delta}}(h) - \mathit{err}_{\tilde{\Delta}}(\tilde{h}) \leq \max(52\sigma(n,\delta), 36\tilde{\epsilon}, (6400C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}})$$

$$\rho_{\tilde{\Delta}}(h, \tilde{h}) \leq \max(C(36\tilde{\epsilon})^{\frac{1}{\kappa}}, C(52\sigma(n,\delta))^{\frac{1}{\kappa}}, C(6400C\sigma(n,\delta))^{\frac{1}{2\kappa-1}})$$

Proof. First, by Lemma 12,

$$\mathrm{err}_{\tilde{\Delta}}(\hat{h}) - \mathrm{err}_{\tilde{\Delta}}(\tilde{h}) \leq \max(2\sigma(n,\delta), (4C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}}, 2\tilde{\epsilon}) \tag{21}$$

$$\rho_{\tilde{\Delta}}(\hat{h}, \tilde{h}) \leq \max(C(2\sigma(n,\delta))^{\frac{1}{\kappa}}, C(4C\sigma(n,\delta))^{\frac{1}{2\kappa-1}}, C(2\tilde{\epsilon})^{\frac{1}{\kappa}}) \tag{22}$$

Next, if $h \in \tilde{V}$, then

$$\mathrm{err}_S(h) - \mathrm{err}_S(\hat{h}) \leq \sigma(n,\delta) + \sqrt{\sigma(n,\delta)\rho_S(h, \hat{h})} + \frac{\tilde{\epsilon}}{2}$$

Combining it with Equation (6) of Lemma 7: $\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(\hat{h}) \leq \mathrm{err}_S(h) - \mathrm{err}_S(\hat{h}) + \sqrt{\sigma(n,\delta)\rho_S(h, \hat{h})} + \sigma(n,\delta)$, we get

$$\mathrm{err}_{\tilde{\Delta}}(h) - \mathrm{err}_{\tilde{\Delta}}(\hat{h}) \leq 2\sigma(n,\delta) + 2\sqrt{\sigma(n,\delta)\rho_S(h, \hat{h})} + \frac{\tilde{\epsilon}}{2}$$

By Equation (7) of Lemma 7,

$$\rho_S(h,\hat{h}) \leq \rho_{\tilde{\Delta}}(h,\hat{h}) + \sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\hat{h})} + \sigma(n,\delta) \leq 2\rho_{\tilde{\Delta}}(h,\hat{h}) + 2\sigma(n,\delta) \tag{23}$$

Therefore,

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\hat{h}) \leq 5\sigma(n,\delta) + 3\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\hat{h})} + \frac{\tilde{\epsilon}}{2} \tag{24}$$

Hence

$$\begin{aligned}
& \text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}) \\
= \;& (\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\hat{h})) + (\text{err}_{\tilde{\Delta}}(\hat{h}) - \text{err}_{\tilde{\Delta}}(\tilde{h})) \\
\leq \;& (4C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}} + 7\sigma(n,\delta) + 3\tilde{\epsilon} + 3\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\hat{h})} \\
\leq \;& (4C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}} + 7\sigma(n,\delta) + 3\tilde{\epsilon} + 3\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\tilde{h})} + 3\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(\tilde{h},\hat{h})}
\end{aligned}$$

Here the first inequality follows from Equations (21) and (24) and $\max(A,B,C) \leq A+B+C$, and the second inequality follows from triangle inequality and $\sqrt{A+B} \leq \sqrt{A}+\sqrt{B}$.

From Equation (22), $\sigma(n,\delta)\rho_{\tilde{\Delta}}(\hat{h},\tilde{h})$ is at most:

$$\begin{aligned}
\leq \;& C\sigma(n,\delta) \cdot ((2\tilde{\epsilon})^{1/\kappa} + (2\sigma(n,\delta))^{1/\kappa} + (4C\sigma(n,\delta))^{1/(2\kappa-1)}) \\
\leq \;& (4C\sigma(n,\delta))^{2\kappa/(2\kappa-1)} + C\sigma(n,\delta)((2\tilde{\epsilon})^{1/\kappa} + (2\sigma(n,\delta))^{1/\kappa}) \\
\leq \;& (4C\sigma(n,\delta))^{2\kappa/(2\kappa-1)} + \max(4\tilde{\epsilon}^2, (C\sigma(n,\delta))^{2\kappa/(2\kappa-1)}) + \max(4\sigma(n,\delta)^2, (C\sigma(n,\delta))^{2\kappa/(2\kappa-1)}),
\end{aligned}$$

where the first step follows from Equation (22), the second step from algebra, and the third step from using the fact that $A^{\frac{2\kappa-1}{\kappa}}B^{\frac{1}{\kappa}} \leq \max(A^2, B^2)$. Plugging this in to the previous equation, and using $\max(A,B) \leq A+B$ and $\sqrt{A+B} \leq \sqrt{A}+\sqrt{B}$, we get that:

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}) \quad \leq \quad 10(4C\sigma(n,\delta))^{\kappa/(2\kappa-1)} + 9\tilde{\epsilon} + 13\sigma(n,\delta) + 3\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\tilde{h})}$$

Combining this with the fact that $A+B+C+D \leq 4\max(A,B,C,D)$, we get that this is at most:

$$\leq \quad \max(40(4C\sigma(n,\delta))^{\kappa/(2\kappa-1)}, 36\tilde{\epsilon}, 52\sigma(n,\delta), 12\sqrt{\sigma(n,\delta)\rho_{\tilde{\Delta}}(h,\tilde{h})})$$

Combining this with Condition (19), we get that this is at most:

$$\max(40(4C\sigma(n,\delta))^{\kappa/(2\kappa-1)}, 36\tilde{\epsilon}, 52\sigma(n,\delta), 12\sqrt{C\sigma(n,\delta)\tilde{\epsilon}^{1/\kappa}}, 12\sqrt{C\sigma(n,\delta)(\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}))^{1/\kappa}})$$

Using $A^{(2\kappa-1)/2\kappa}B^{1/2\kappa} \leq \max(A,B)$, we get that $\sqrt{C\sigma(n,\delta)\tilde{\epsilon}^{1/\kappa}} \leq \max(\tilde{\epsilon}, (C\sigma(n,\delta))^{\kappa/(2\kappa-1)})$. Also note $\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}) \leq 12\sqrt{C\sigma(n,\delta)(\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}))^{1/\kappa}}$ implies $\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}) \leq (144C\sigma(n,\delta))^{\kappa/(2\kappa-1)}$. Thus we have

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\tilde{h}) \leq \max(36\tilde{\epsilon}, 52\sigma(n,\delta), (6400C\sigma(n,\delta))^{\frac{\kappa}{2\kappa-1}})$$

Invoking (19) again, we have that:

$$\rho_{\tilde{\Delta}}(h,\tilde{h}) \leq \max(C(36\tilde{\epsilon})^{\frac{1}{\kappa}}, C(52\sigma(n,\delta))^{\frac{1}{\kappa}}, C(6400C\sigma(n,\delta))^{\frac{1}{2\kappa-1}})$$

□

D Remaining Proofs from Section 2

Proof. (Of Lemma 1) Assuming E_r happens, we prove the lemma by induction.
Base Case: For $k = 1$, clearly $h^*(D) \in V_1 = \mathcal{H}$.
Inductive Case: Assume $h^*(D) \in V_k$. As we are in the realizable case, $h^*(D)$ is consistent with the examples S_k drawn in Step 8 of Algorithm 1; thus $h^*(D) \in V_{k+1}$. The lemma follows. □

Proof. (Of Lemma 2) We use $\tilde{h}_k = \text{argmin}_{h \in V_k} \text{err}_{\tilde{\Gamma}_k}(h)$ to denote the optimal classifier in V_k with respect to the distribution $\tilde{\Gamma}_k$. Assuming E_a happens, we prove the lemma by induction.
Base Case: For $k = 1$, clearly $h^*(D) \in V_1 = \mathcal{H}$.
Inductive Case: Assume $h^* \in V_k$. In order to show the inductive case, our goal is to show that:

$$\mathbb{P}_{\tilde{\Gamma}_k}(h^*(D)(x) \neq y) - \mathbb{P}_{\tilde{\Gamma}_k}(\tilde{h}_k(x) \neq y) \leq \frac{\epsilon_k}{16\phi_k} \tag{25}$$

If (25) holds, then, by (2.1) of Lemma 4, we know that if Algorithm 2 succeeds when called in iteration k of Algorithm 1, then, it is guaranteed that $h^* \in V_{k+1}$.

We therefore focus on showing (25). First, from Equation (12) of Lemma 9, we have:

$$(\text{err}_{\tilde{U}_k}(h^*(D)) - \text{err}_{\tilde{U}_k}(\tilde{h}_k)) - (\text{err}_D(h^*(D)) - \text{err}_D(\tilde{h}_k)) \leq \frac{\epsilon_k}{32}$$

As $\text{err}_D(h^*(D)) \leq \text{err}_D(\tilde{h}_k)$, we get:

$$\text{err}_{\tilde{U}_k}(h^*(D)) \leq \text{err}_{\tilde{U}_k}(\tilde{h}_k) + \frac{\epsilon_k}{32} \tag{26}$$

On the other hand, by Equation (14) of Lemma 10 and triangle inequality,

$$\begin{aligned} & \mathbb{E}_{\tilde{U}_k}[I(\tilde{h}_k(x) \neq y)(1 - \gamma_k(x))] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)(1 - \gamma_k(x))] && (27) \\ \leq \; & \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq \tilde{h}_k(x))(1 - \gamma_k(x))] \leq \frac{\epsilon_k}{32} && (28) \end{aligned}$$

Combining Equations (26) and (27), we get:

$$\begin{aligned} \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)] &= \text{err}_{\tilde{U}_k}(h^*(D)(x)) - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)(1 - \gamma_k(x))] \\ &\leq \text{err}_{\tilde{U}_k}(\tilde{h}_k(x)) + \epsilon_k/32 - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)(1 - \gamma_k(x))] \\ &\leq \mathbb{E}_{\tilde{U}_k}[I(\tilde{h}_k(x) \neq y)\gamma_k(x)] + \mathbb{E}_{\tilde{U}_k}[I(\tilde{h}(x) \neq y)(1 - \gamma_k(x))] + \epsilon_k/32 \\ &\quad - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)(1 - \gamma_k(x))] \\ &\leq \mathbb{E}_{\tilde{U}_k}[I(\tilde{h}_k(x) \neq y)\gamma_k(x)] + \epsilon_k/16 \end{aligned}$$

Dividing both sides by ϕ_k, we get:

$$\mathbb{P}_{\tilde{\Gamma}_k}(h^*(D)(x) \neq y) - \mathbb{P}_{\tilde{\Gamma}_k}(\tilde{h}_k(x) \neq y) \leq \frac{\epsilon_k}{16\phi_k},$$

from which the lemma follows. □

Proof. (of Lemma 3) Assuming E_r happens, we prove the lemma by induction.
Base Case: For $k = 1$, clearly $\text{err}_D(h) \leq 1 \leq \epsilon_1 = \epsilon 2^{k_0}, \forall h \in V_1 = \mathcal{H}$.
Inductive Case: Note that $\forall h, h' \in V_{k+1} \subseteq V_k$, by Equation (14) of Lemma 10, we have:

$$\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq h'(x))(1 - \gamma_k(x))] \leq \frac{\epsilon_k}{8}$$

By the proof of Lemma 1, $h^*(D) \in V_{k+1}$ on event E_r, thus $\forall h \in V_{k+1}$,

$$\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq h^*(D)(x))(1 - \gamma_k(x))] \leq \frac{\epsilon_k}{8} \tag{29}$$

Since any $h \in V_{k+1}$, h is consistent with S_k of size $m_k = \frac{1536\phi_k}{\epsilon_k}(d \ln \frac{1536\phi_k}{\epsilon_k} + \ln \frac{48}{\delta_k})$, we have that for all $h \in V_{k+1}$,

$$\mathbb{P}_{\tilde{\Gamma}_k}(h(x) \neq h^*(D)(x)) \leq \frac{\epsilon_k}{8\phi_k}$$

That is,

$$\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq h^*(D)(x))\gamma_k(x)] \leq \frac{\epsilon_k}{8}$$

Combining this with Equation (29) above,

$$\mathbb{P}_{\tilde{U}_k}(h(x) \neq h^*(D)(x)) \leq \frac{\epsilon_k}{4}$$

By Equation (11) of Lemma 9,

$$\mathbb{P}_D(h(x) \neq h^*(D)(x)) \leq \frac{\epsilon_k}{2} = \epsilon_{k+1}$$

The lemma follows. □

Proof. (of Lemma 6) Assuming E_a happens, we prove the lemma by induction.
Base Case: For $k = 1$, clearly $\mathrm{err}_D(h) - \mathrm{err}_D(h^*(D)) \le 1 \le \epsilon_1 = \epsilon 2^{k_0}, \forall h \in V_1 = \mathcal{H}$.
Inductive Case: Note that $\forall h, h' \in V_{k+1} \subseteq V_k$, by Equation (14) of Lemma 10,

$$\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)(1-\gamma_k(x))] - \mathbb{E}_{\tilde{U}_k}[I(h'(D)(x) \neq y)(1-\gamma_k(x))] \le \mathbb{E}_{\tilde{U}_k}[I(h(x) \neq h'(D)(x))(1-\gamma_k(x))] \le \frac{\epsilon_k}{8}$$

From Lemma 2, $h^*(D) \in V_k$ whenever the event E_a happens. Thus $\forall h \in V_{k+1}$,

$$\mathbb{E}_{\tilde{U}_k} I(h(x) \neq y)(1 - \gamma_k(x)) - \mathbb{E}_{\tilde{U}_k} I(h^*(D)(x) \neq y)(1 - \gamma_k(x)) \le \frac{\epsilon_k}{8} \tag{30}$$

On the other hand, if Algorithm 2 succeeds with target excess error $\frac{\epsilon_k}{8\phi_k}$, by item(2.2) of Lemma 4, for any $h \in V_{k+1}$,

$$\mathbb{P}_{\tilde{\Gamma}_k}(h(x) \neq y) - \min_{h \in V_k} \mathbb{P}_{\tilde{\Gamma}_k}(h(x) \neq y) \le \frac{\epsilon_k}{8\phi_k}$$

Moreover, as $h^*(D) \in V_k$ from Lemma 2,

$$\mathbb{P}_{\tilde{\Gamma}_k}(h(x) \neq y) - \mathbb{P}_{\tilde{\Gamma}_k}(h^*(D)(x) \neq y) \le \frac{\epsilon_k}{8\phi_k}$$

In other words,

$$\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)] \le \frac{\epsilon_k}{8}$$

Combining this with Equation (30), we get that for all $h \in V_{k+1}$,

$$\mathbb{P}_{\tilde{U}_k}(h(x) \neq y) - \mathbb{P}_{\tilde{U}_k}(h^*(D)(x) \neq y) \le \frac{\epsilon_k}{4}$$

Finally, combining this with Equation (12) of Lemma 9, we have that:

$$\mathbb{P}_D(h(x) \neq y) - \mathbb{P}_D(h^*(D)(x) \neq y) \le \frac{\epsilon_k}{2} = \epsilon_{k+1}$$

The lemma follows. □

Proof. (of Theorem 1) In the realizable case, We observe that for example z_i, $\zeta_i = \mathbb{P}(P(z_i) = -1)$, $\xi_i = \mathbb{P}(P(z_i) = 1)$, and $\gamma_i = \mathbb{P}(P(z_i) = 0)$. Suppose $h^* \in \mathcal{H}$ is the true hypothesis which has 0 error with respect to the data distribution. By the realizability assumption, $h^* \in V$. Moreover, $\mathbb{P}_U(P(x) \neq h^*(x), P(x) \neq 0) = \frac{1}{m}(\sum_{i:h^*(z_i)=+1} \zeta_i + \sum_{i:h^*(z_i)=-1} \xi_i) \le \eta$ by Algorithm 3.
In the non-realizable case, we still have $\mathbb{P}_{x \sim U}(h^*(x) \neq P(x), P(x) \neq 0) \le \eta$, hence by triangle inequality, $\mathbb{P}_{x \sim U}(P(x) \neq x, P(x) \neq 0) - \mathbb{P}_{x \sim U}(h^*(x) \neq y, P(x) \neq 0) \le \eta$. Thus

$$\mathbb{P}_{x \sim U}(P(x) \neq y, P(x) \neq 0) \le \mathbb{P}_{x \sim U}(h^*(x) \neq y) + \eta$$

□

Proof. (of Theorem 2) Suppose P' assigns probabilities $\{[\xi_i', \zeta_i', \gamma_i'], i = 1, \ldots, m\}$ to the unlabelled examples z_i, and suppose for the sake of contradiction that $\sum_{i=1}^m \xi_i' + \zeta_i' > \sum_{i=1}^m \xi_i + \zeta_i$. Then, $\{\xi_i', \zeta_i', \gamma_i'\}$'s cannot satisfy the LP in Algorithm 3, and thus there exists some $h' \in V$ for which constraint (2) is violated. The true hypothesis that generates the data could be any $h \in V$; if this true hypothesis is h', then $\mathbb{P}_{x \sim U}(P'(x) \neq h'(x), P'(x) \neq 0) > \delta$. □

E Proofs from Section 3

Proof. (of Theorem 4)
(1) In the realizable case, suppose that event E_r happens. Then from Equation (15) of Lemma 10, while running Algorithm 3, we have that:

$$\phi_k \le \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \le \mathbf{\Phi}_D(B_D(h^*, \epsilon_k), \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \le \mathbf{\Phi}_D(B_D(h^*, \epsilon_k), \frac{\epsilon_k}{256}) = \phi(\epsilon_k, \frac{\epsilon_k}{256})$$

where the second inequality follows from the fact that $V_k \subseteq B_D(h^*(D), \epsilon_k)$, and third inequality follows from Lemma 18 and denseness assuption.
Thus, there exists $c_3 > 0$ such that, in round k,

$$m_k = (d \ln \frac{1536\phi_k}{\epsilon_k} + \ln \frac{48}{\delta_k}) \frac{1536\phi_k}{\epsilon_k} \le c_3(d \ln \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{k_0 - k + 1}{\delta})) \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k}$$

Hence the total number of labels queried by Algorithm 1 is at most

$$\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} m_k \leq c_3 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{k_0 - k + 1}{\delta})) \frac{\phi(\epsilon_k, \epsilon_k/256)}{\epsilon_k}$$

(2) In the agnostic case, suppose the event E_a happens.
First, given E_a, from Equation (15) of Lemma 10 when running Algorithm 3,

$$\phi_k \leq \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \leq \mathbf{\Phi}_D(B_D(h^*, 2\nu^*(D) + \epsilon_k), \frac{\epsilon_k}{256}) = \phi(2\nu^*(D) + \epsilon_k, \frac{\epsilon_k}{256}) \quad (31)$$

where the second inequality follows from the fact that $V_k \subseteq B_D(h^*(D), 2\nu^*(D) + \epsilon_k)$ and the third inequality follows from Lemma 18 and denseness assumption.
Second, recall that $\tilde{h}_k = \text{argmin}_{h \in V_k} \text{err}_{\tilde{\Gamma}_k}(h)$,

$$\begin{aligned} \text{err}_{\tilde{\Gamma}_k}(\tilde{h}_k) &= \min_{h \in V_k} \text{err}_{\tilde{\Gamma}_k}(h) \\ &\leq \text{err}_{\tilde{\Gamma}_k}(h^*(D)) \\ &= \frac{\mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)]}{\phi_k} \\ &\leq \frac{\mathbb{P}_{\tilde{U}_k}(h^*(D)(x) \neq y)}{\phi_k} \\ &\leq \frac{\nu^*(D) + \epsilon_k/64}{\phi_k} \end{aligned}$$

Here the first inequality follows from the suboptimality of $h^*(D)$ under distribution $\tilde{\Gamma}_k$, the second inequality follows from $\gamma_k(x) \leq 1$, and the third inequality follows from Equation (11).
Thus, conditioned on E_a, in iteration k, Algorithm 2 succeeds by Lemma 5, and there exists a constant $c_4 > 0$ such that the number of labels queried is

$$\begin{aligned} & m_k \leq c_1 \frac{\frac{\epsilon_k}{8\phi_k} + \text{err}_{\tilde{\Gamma}_k}(\tilde{h}_k)}{(\frac{\epsilon_k}{8\phi_k})^2} (d \ln \frac{1}{\frac{\epsilon_k}{8\phi_k}} + \ln \frac{2}{\delta_k}) \\ \leq \; & c_4 (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{k_0 - k + 1}{\delta})) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} (1 + \frac{\nu^*(D)}{\epsilon_k}) \end{aligned}$$

Here the last line follows from Equation (31). Hence the total number of examples queried is at most:

$$\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} m_k \leq c_4 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{k_0 - k + 1}{\delta})) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} (1 + \frac{\nu^*(D)}{\epsilon_k})$$

□

Proof. (of Theorem 5) Assume E_a happens.
First, from Equation (15) of Lemma 10 when running Algorithm 3,

$$\phi_k \leq \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \leq \mathbf{\Phi}_D(B_D(h^*, C_0 \epsilon_k^{\frac{1}{\kappa}}), \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256} \leq \mathbf{\Phi}_D(B_D(h^*, C_0 \epsilon_k^{\frac{1}{\kappa}}), \frac{\epsilon_k}{256}) = \phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \quad (32)$$

where the second inequality follows from the fact that $V_k \subseteq B_D(h^*(D), C_0 \epsilon_k^{\frac{1}{\kappa}})$, and the third inequality follows from Lemma 18 and denseness assumption.

Second, for all $h \in V_k$,

$$\begin{aligned}
& \phi_k \rho_{\tilde{\Gamma}_k}(h, h^*(D)) \\
= \; & \mathbb{E}_{\tilde{U}_k} I(h(x) \neq h^*(D)(x)) \gamma_k(x) \\
\leq \; & \rho_{\tilde{U}_k}(h, h^*(D)) \\
\leq \; & \rho_D(h, h^*(D)) + \epsilon_k/32 \\
\leq \; & C_0(\text{err}_D(h) - \text{err}_D(h^*(D)))^{\frac{1}{\kappa}} + \epsilon_k/32 \\
\leq \; & C_0(\text{err}_{\tilde{U}_k}(h) - \text{err}_{\tilde{U}_k}(h^*(D)) + \epsilon_k/64)^{\frac{1}{\kappa}} + \epsilon_k/32 \\
= \; & C_0(\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)] \\
& + \mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)(1 - \gamma_k(x))] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)(1 - \gamma_k(x))] + \epsilon_k/16)^{\frac{1}{\kappa}} + \epsilon_k/32
\end{aligned}$$

Here the first inequality follows from $\gamma_k(x) \leq 1$, the second inequality follows from Equation (13) of Lemma 9, the third inequality follows from Definition 1 and the fourth inequality follows from Equation (12) of Lemma 9. The above can be upper bounded by:

$$\begin{aligned}
\leq \; & C_0(\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)] + \epsilon_k/16)^{\frac{1}{\kappa}} + \epsilon_k/32 \\
\leq \; & 2C_0(\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)])^{\frac{1}{\kappa}} + 2C_0(\epsilon_k/16)^{\frac{1}{\kappa}} + \epsilon_k/32 \\
\leq \; & \max(8C_0, 4) \max((\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)]), \frac{\epsilon_k}{16})^{\frac{1}{\kappa}} \\
= \; & \max(8C_0, 4)(\phi_k)^{\frac{1}{\kappa}} \max(\mathbb{P}_{\tilde{\Gamma}_k}(h(x) \neq y) - \mathbb{P}_{\tilde{\Gamma}_k}(h^*(D)(x) \neq y), \frac{\epsilon_k}{8\phi_k})^{\frac{1}{\kappa}}
\end{aligned}$$

Here the first inequality follows from Equation (14) of Lemma 10 and triangle inequality $\mathbb{E}_{\tilde{U}_k}[I(h(x) \neq y)\gamma_k(x)] - \mathbb{E}_{\tilde{U}_k}[I(h^*(D)(x) \neq y)\gamma_k(x)] \leq \mathbb{E}_{\tilde{U}_k}[I(h(x) \neq h^*(D)(x))\gamma_k(x)] \leq \epsilon_k/32$, and the last two inequalities follow from simple algebra.

Dividing both sides by ϕ_k, we get:

$$\rho_{\tilde{\Gamma}_k}(h, h^*(D)) \leq C_1(\phi_k)^{\frac{1}{\kappa}-1} \max(\text{err}_{\tilde{\Gamma}_k}(h) - \text{err}_{\tilde{\Gamma}_k}(h^*(D)), \frac{\epsilon_k}{8\phi_k})^{\frac{1}{\kappa}}$$

where $C_1 = \max(8C_0, 4)$. Thus in iteration k, Condition (19) in Lemma 11 holds with $C := C_1(\phi_k)^{\frac{1}{\kappa}-1}$ and $\tilde{h} := h^*(D)$. Thus, from Lemma 11, Algorithm 2 succeeds, and there exists a constant $c_5 > 0$, such that the number of labels queried is

$$\begin{aligned}
m_k \leq \; & c_2 \max((d \ln(C_1(\phi_k)^{\frac{1}{\kappa}-1}(\frac{\epsilon_k}{8\phi_k})^{\frac{1}{\kappa}-2}) + \ln \frac{2}{\delta_k})(C_1(\phi_k)^{\frac{1}{\kappa}-1}(\frac{\epsilon_k}{8\phi_k})^{\frac{1}{\kappa}-2}), \\
& (d \ln(\frac{\epsilon_k}{8\phi_k})^{-1} + \ln \frac{2}{\delta_k})(\frac{\epsilon_k}{8\phi_k})^{-1}) \\
\leq \; & c_5(d \ln(\phi_k \epsilon_k^{\frac{1}{\kappa}-2}) + \ln(\frac{k_0 - k + 1}{\delta}))\phi_k \epsilon_k^{\frac{1}{\kappa}-2} \\
\leq \; & c_5(d \ln(\phi(C_0\epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})\epsilon_k^{\frac{1}{\kappa}-2}) + \ln(\frac{k_0 - k + 1}{\delta}))\phi(C_0\epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})\epsilon_k^{\frac{1}{\kappa}-2}
\end{aligned}$$

Where the last line follows from Equation (31). Hence the total number of examples queried is at most

$$\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} m_k \leq c_5 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\phi(C_0\epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})\epsilon_k^{\frac{1}{\kappa}-2}) + \ln(\frac{k_0 - k + 1}{\delta}))\phi(C_0\epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})\epsilon_k^{\frac{1}{\kappa}-2}$$

□

The following lemma is an immediate corollary of Theorem 21, item (a) of Lemma 2 and Lemma 3 of [4]:

Lemma 14. *Suppose D is isotropic and log-concave on R^d, and $\mathcal{H}$ is the set of homogeneous linear classifiers on R^d, then there exist absolute constants $c_6, c_7 > 0$ such that $\phi(r, \eta) \le c_6 r \ln \frac{c_7 r}{\eta}$.*

Proof. (of Lemma 14) Denote w_h as the unit vector w such that $h(x) = \text{sign}(w \cdot x)$, and $\theta(w, w')$ to be the angle between vectors w and w'. If $h \in B_D(h^*, r)$, then by Lemma 3 of [4], there exists some constant $c_{11} > 0$ such that $\theta(w_h, w_{h^*}) \le \frac{r}{c_{11}}$. Also, by Lemma 21 of [4], there exists some constants $c_{12}, c_{13} > 0$, such that, if $\theta(w, w') = \alpha$ then

$$\mathbb{P}_D(\text{sign}(w \cdot x) \ne \text{sign}(w' \cdot x), |w \cdot x| \ge b) \le c_{12} \alpha \exp(-c_{13} \frac{b}{\alpha})$$

We define a special solution (ξ, ζ, γ) as follows:

$$\begin{aligned} \xi(x) &:= I(w_{h^*} \cdot x \ge \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}) \\ \zeta(x) &:= I(w_{h^*} \cdot x \le -\frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}) \\ \gamma(x) &:= I(|w_{h^*} \cdot x| \le \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}) \end{aligned}$$

Then it can be checked that for all $h \in B_D(h^*, r)$,

$$\mathbb{E}[I(h(x) = +1)\zeta(x) + I(h(x) = -1)\xi(x)] = \mathbb{P}_D(\text{sign}(w_{h^*} \cdot x) \ne \text{sign}(w_h \cdot x), |w_{h^*} \cdot x| \ge \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}) \le \eta$$

And by item (a) of Lemma 2 of [4], we have

$$\mathbb{E}\gamma(x) = \mathbb{P}_D(|w_{h^*} \cdot x| \le \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}) \le \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}$$

Hence,

$$\phi(r, \eta) \le \frac{r}{c_{11} c_{13}} \ln \frac{c_{12} r}{c_{11} \eta}$$

□

Proof. (of Corollary 1) This is an immediate consequence of Lemma 14 and Theorems 4 and 5 and algebra. □

F A Suboptimal Alternative to Algorithm 2

Algorithm 4 An Nonadaptive Algorithm for Label Query Given Target Excess Error

```
Inputs: Hypothesis set $V$ of VC dimension $d$, Example distribution $\Delta$, Labeling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon}$, target confidence $\tilde{\delta}$.
Draw $n = \frac{12288}{\tilde{\epsilon}^2}(d \ln \frac{12288}{\tilde{\epsilon}^2} + \ln \frac{24}{\tilde{\delta}})$ i.i.d examples from $\Delta$; query their labels from $\mathcal{O}$ to get a labelled dataset $S$.
Train an ERM classifier $\hat{h} \in V$ over $S$.
Define the set $V$ as follows:
    $$V_1 = \left\{ h \in V : \text{err}_S(h) \le \text{err}_S(\hat{h}) + \frac{3\tilde{\epsilon}}{4} \right\}$$
return $V_1$.
```

It is immediate that we have the following lemma.

Lemma 15. *Suppose we run Algorithm 4 with inputs hypothesis set V, example distribution Δ, labelling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon}$ and target confidence $\tilde{\delta}$. Then there exists an event $\tilde{E}$, $\mathbb{P}(\tilde{E}) \ge 1 - \tilde{\delta}$, such that on $\tilde{E}$, the set V_1 has the following property. (1) If for $h \in \mathcal{H}$, $err_{\tilde{\Delta}}(h) - err_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \le \tilde{\epsilon}/2$, then $h \in V_1$. (2) On the other hand, if $h \in V_1$, then $err_{\tilde{\Delta}}(h) - err_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \le \tilde{\epsilon}$.*

When $\tilde{E}$ happens, we say that Algorithm 4 succeeds.

Proof. By Equation (9) of Lemma 8 and because $n = \frac{12288}{\tilde{\epsilon}^2}(d \ln \frac{12288}{\tilde{\epsilon}^2} + \ln \frac{24}{\tilde{\delta}})$, we have for all $h, h' \in \mathcal{H}$,

$$(\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(h')) - (\text{err}_S(h) - \text{err}_S(h')) \leq \frac{\tilde{\epsilon}}{4}$$

For the proof of (1), for any $h \in V$, $\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \tilde{\epsilon}/2$, then

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(\hat{h}) \leq \tilde{\epsilon}/2$$

Thus

$$\text{err}_S(h) - \text{err}_S(\hat{h}) \leq \frac{3\tilde{\epsilon}}{4}$$

proving $h \in V_1$.
For the proof of (2), for any $h \in V_1$,

$$\text{err}_S(h) - \text{err}_S(h') \leq \frac{3\tilde{\epsilon}}{4}$$

Thus

$$\text{err}_S(h) - \text{err}_S(h^*(\tilde{\Delta})) \leq \frac{3\tilde{\epsilon}}{4}$$

Combining with the fact that $(\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta}))) - (\text{err}_S(h) - \text{err}_S(h^*(\tilde{\Delta}))) \leq \frac{\tilde{\epsilon}}{4}$ we have

$$\text{err}_{\tilde{\Delta}}(h) - \text{err}_{\tilde{\Delta}}(h^*(\tilde{\Delta})) \leq \tilde{\epsilon}$$

□

Corollary 2. *Suppose we replace the calls to Algorithm 2 with Algorithm 4 in Algorithm 1, then run it with inputs example oracle $\mathcal{U}$, labelling oracle $\mathcal{O}$, hypothesis class V, confidence-rated predictor P of Algorithm 3, target excess error ϵ and target confidence δ. Then the modified algorithm has a label complexity of*

$$\tilde{O}(\sum_{k=1}^{\lceil \log 1/\epsilon \rceil} (d(\frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k})^2)$$

in the agnostic case and

$$\tilde{O}(\sum_{k=1}^{\lceil \log 1/\epsilon \rceil} d(\frac{\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})}{\epsilon_k^{\frac{1}{\kappa}}})^2 \epsilon_k^{\frac{2}{\kappa}-2})$$

under (C_0, κ)-Tsybakov Noise Condition.

Under denseness assumption, by Lemma 17, we have $\phi(r, \eta) \geq r - 2\eta$, the label complexity bounds given by Corollary 2 is always no better than the ones given by Theorem 4 and 5.

Proof. (Sketch) Define event

> $E_a = \{$For all $k = 1, 2, \ldots, k_0$: Equations (11), (12), (13), (14), (15) hold for $\tilde{U}_k$ with confidence $\delta_k/2$, and Algorithm 4 succeeds with inputs hypothesis set $V = V_k$, example distribution $\Delta = \Gamma_k$, labelling oracle $\mathcal{O}$, target excess error $\tilde{\epsilon} = \dfrac{\epsilon_k}{8\phi_k}$ and target confidence $\tilde{\delta} = \dfrac{\delta_k}{2}\}$.

Clealy, $\mathbb{P}(E_a) \geq 1 - \delta$. On the event E_a, there exists an absolute constant $c_{13} > 0$, such that the number of examples queried in interation k is

$$m_k \leq c_{13}(d \ln \frac{8\phi_k}{\epsilon_k} + \ln \frac{2}{\delta})(\frac{8\phi_k}{\epsilon_k})^2$$

Combining it with Equation (15) of Lemma 10

$$\phi_k \leq \mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256}$$

we have

$$m_k \leq O((d \ln \frac{\mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256}}{\epsilon_k} + \ln \frac{2}{\delta_k})(\frac{\mathbf{\Phi}_D(V_k, \frac{\epsilon_k}{128}) + \frac{\epsilon_k}{256}}{\epsilon_k})^2)$$

The rest of the proof follows from Lemma 18 and denseness assumption, along with algebra. □

G Proofs of Concentration Lemmas

Proof. (of Lemma 9) We begin by observing that:

$$\mathrm{err}_{\tilde{U}_k}(h) = \frac{1}{n_k}\sum_{i=1}^{n_k}[\mathbb{P}_D(Y=+1|X=x_i)I(h(x_i)=-1) + \mathbb{P}_D(Y=-1|X=x_i)I(h(x_i)=+1)]$$

Moreover, $\max(\mathcal{S}(\{I(h(x)=1, h\in\mathcal{H})\}, n), \mathcal{S}(\{I(h(x)=-1, h\in\mathcal{H})\}, n)) \leq (\frac{en}{d})^d$. Combining this fact with Lemma 16, the following equations hold simultaneously with probability $1-\delta_k/6$:

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{P}_D(Y=+1|X=x_i)I(h(x_i)=-1) - \mathbb{P}_D(h(x)=-1, y=+1)\Big| \leq \sqrt{\frac{16(d\ln\frac{en_k}{d} + \ln\frac{24}{\delta_k})}{n_k}} \leq \frac{\epsilon_k}{128}$$

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{P}_D(Y=-1|X=x_i)I(h(x_i)=+1) - \mathbb{P}_D(h(x)=+1, y=-1)\Big| \leq \sqrt{\frac{16(d\ln\frac{en_k}{d} + \ln\frac{24}{\delta_k})}{n_k}} \leq \frac{\epsilon_k}{128}$$

Thus Equation (11) holds with probability $1-\delta_k/6$. Moreover, we observe that Equation (11) implies Equation (12). To show Equation (13), we observe that by Lemma 8, with probability $1-\delta_k/12$,

$$|\rho_D(h,h') - \rho_{\tilde{U}_k}(h,h')| = |\rho_D(h,h') - \rho_{S_k}(h,h')| \leq 2\sqrt{\sigma(n_k, \delta_k/12)} \leq \frac{\epsilon_k}{64}$$

Thus, Equation (13) holds with probability $\geq 1-\delta_k/12$. By union bound, with probability $1-\delta_k/4$, Equations (11), (12), and (13) hold simultaneously. □

Proof. (of Lemma 10) (1) Given a confidence-rated predictor with inputs hypothesis set V_k, unlabelled data U_k, and error bound $\epsilon_k/64$, the outputs $\{(\xi_{k,i}, \zeta_{k,i}, \gamma_{k,i})\}_{i=1}^{n_k}$ must satisfy that for all $h, h' \in V_k$,

$$\frac{1}{n_k}\sum_{i=1}^{n_k}[I(h(x_{k,i})=-1)\xi_{k,i} + I(h(x_{k,i})=+1)\zeta_{k,i}] \leq \frac{\epsilon_k}{64}$$

$$\frac{1}{n_k}\sum_{i=1}^{n_k}[I(h'(x_{k,i})=-1)\xi_{k,i} + I(h'(x_{k,i})=+1)\zeta_{k,i}] \leq \frac{\epsilon_k}{64}$$

Since $I(h(x)\neq h'(x)) \leq \min(I(h(x)=-1)+I(h'(x)=-1), I(h(x)=+1)+I(h'(x)=+1))$, adding up the two inequalities above, we get

$$\frac{1}{n_k}\sum_{i=1}^{n_k}[I(h(x_{k,i})\neq h'(x_{k,i}))(\xi_{k,i}+\zeta_{k,i})] \leq \frac{\epsilon_k}{32}$$

That is,

$$\frac{1}{n_k}\sum_{i=1}^{n_k}[I(h(x_{k,i})\neq h'(x_{k,i}))(1-\gamma_{k,i})] \leq \frac{\epsilon_k}{32}$$

(2) By definition of $\mathbf{\Phi}_D(V,\eta)$, there exist nonnegative functions ξ, ζ, γ such that $\xi(x)+\zeta(x)+\gamma(x)\equiv 1$, $\mathbb{E}_D[\gamma(x)] = \mathbf{\Phi}_D(V_k, \epsilon_k/128)$ and for all $h\in V_k$,

$$\mathbb{E}_D[\xi(x)I(h(x)=-1) + \zeta(x)I(h(x)=+1)] \leq \frac{\epsilon_k}{128}$$

Consider the linear progam in Algorithm 3 with inputs hypothesis set V_k, unlabelled data U_k, and error bound $\epsilon_k/64$. We consider the following special (but possibly non-optimal) solution for this LP: $\xi_{k,i} = \xi(z_{k,i}), \zeta_{k,i} = \zeta(z_{k,i}), \gamma_{k,i} = \gamma(z_{k,i})$. We will now show that this solution is feasible and has coverage $\mathbf{\Phi}_D(V_k, \epsilon_k/128)$ plus $O(\epsilon_k)$ with high probability.

Observe that $\max(\mathcal{S}(\{I(h(x)=1, h\in\mathcal{H})\}, n), \mathcal{S}(\{I(h(x)=-1, h\in\mathcal{H})\}, n)) \leq (\frac{en}{d})^d$. Therefore, from Lemma 16 and the union bound, with probability $1-\delta_k/4$, the following hold simultaneously for all $h\in\mathcal{H}$:

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}\gamma(z_{k,i}) - \mathbb{E}_D\gamma(x)\Big| \leq \sqrt{\frac{\ln\frac{2}{\delta_k}}{2n_k}} \leq \frac{\epsilon_k}{256} \tag{33}$$

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}\xi(z_{k,i})I(h(z_{k,i})=-1)-\mathbb{E}_D[\xi(x)I(h(x)=-1)]\Big| \leq \sqrt{\frac{8(d\ln\frac{en_k}{d}+\ln\frac{24}{\delta_k})}{n_k}} \leq \frac{\epsilon_k}{256} \quad (34)$$

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}\zeta(z_{k,i})I(h(z_{k,i})=+1)-\mathbb{E}_D[\zeta(x)I(h(x)=+1)]\Big| \leq \sqrt{\frac{8(d\ln\frac{en_k}{d}+\ln\frac{24}{\delta_k})}{n_k}} \leq \frac{\epsilon_k}{256} \quad (35)$$

Adding up Equations (34) and (35),

$$\Big|\frac{1}{n_k}\sum_{i=1}^{n_k}[\zeta(x_i)I(h(x_i)=+1)+\xi(x_i)I(h(x_i)=-1)]-\mathbb{E}_D[\xi(x)I(h(x)=-1)+\zeta(x)I(h(x)=+1))]\Big| \leq \frac{\epsilon_k}{128}$$

Thus $\{(\xi(z_{k,i}),\zeta(z_{k,i})\}_{i=1}^{n_k}$ is a feasible solution of the linear program of Algorithm 3. Also, by Equation (33), $\frac{1}{n_k}\sum_{i=1}^{n_k}\gamma(z_{k,i}) \leq \mathbf{\Phi}_D(V_k,\frac{\epsilon_k}{128})+\frac{\epsilon_k}{64}$. Thus, the outputs $\{(\xi_{k,i},\zeta_{k,i},\gamma_{k,i})\}_{i=1}^{n_k}$ of the linear program in Algorithm 3 satisfy

$$\phi_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\gamma_{k,i} \leq \frac{1}{n_k}\sum_{i=1}^{n_k}\gamma(z_{k,i}) \leq \mathbf{\Phi}_D(V_k,\frac{\epsilon_k}{128})+\frac{\epsilon_k}{256}$$

due to their optimality. □

Lemma 16. *Pick any $n \geq 1$, $\delta \in (0,1)$, a family $\mathcal{F}$ of functions $f:\mathcal{Z}\to\{0,1\}$, a fixed weighting function $w:\mathcal{Z}\to[0,1]$. Let S_n be a set of n iid copies of Z. The following holds with probability at least $1-\delta$:*

$$\Big|\frac{1}{n}\sum_{i=1}^{n}w(z_i)f(z_i)-\mathbb{E}[w(z)f(z)]\Big| \leq \sqrt{\frac{16(\ln\mathcal{S}(\mathcal{F},n)+\ln\frac{2}{\delta})}{n}}$$

where $\mathcal{S}(\mathcal{F},n)=\max_{z_1,\ldots,z_n\in\mathcal{Z}}|\{(f(z_1),\ldots,f(z_n)):f\in\mathcal{F}\}|$ is the growth function of $\mathcal{F}$.

Proof. The proof is fairly standard, and follows immediately from the proof of additive VC bounds. With probability $1-\delta$,

$$\begin{aligned}
& \sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}w(z_i)f(z_i)-\mathbb{E}w(z)f(z)\Big| \\
\leq\ & \mathbb{E}_{S\sim D^n}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}w(z_i)f(z_i)-\mathbb{E}w(z)f(z)\Big|+\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} \\
\leq\ & \mathbb{E}_{S\sim D^n,S'\sim D^n}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}(w(z_i)f(z_i)-w(z_i')f(z_i'))\Big|+\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} \\
\leq\ & \mathbb{E}_{S\sim D^n,S'\sim D^n,\sigma\sim U(\{-1,+1\}^n)}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(w(z_i)f(z_i)-w(z_i')f(z_i'))\Big|+\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} \\
\leq\ & 2\mathbb{E}_{S\sim D^n,\sigma\sim U(\{-1,+1\}^n)}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(z_i)f(z_i)\Big|+\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} \\
\leq\ & 2\sqrt{\frac{2\ln(2\mathcal{S}(\mathcal{F},n))}{n}}+\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} \leq \sqrt{\frac{16(\ln\mathcal{S}(\mathcal{F},n)+\ln\frac{2}{\delta})}{n}}
\end{aligned}$$

Where the first inequality is by McDiarmid's Lemma; the second inequality follows from Jensen's Inequality; the third inequality follows from symmetry; the fourth inequality follows from $|A+B| \leq |A|+|B|$; the fifth inequality follows from Massart's Finite Lemma. □

Lemma 17. *Let $0 < 2\eta \leq r \leq 1$. Given a hypothesis set V and data distribution D over $\mathcal{X}\times\mathcal{Y}$, if there exist $h_1,h_2\in V$ such that $\rho_D(h_1,h_2)\geq r$, then $\mathbf{\Phi}_D(V,\eta)\geq r-2\eta$.*

Proof. Let (ξ, ζ, γ) be a triple of functions from $\mathcal{X}$ to R^3 satisfying the following conditions: $\xi, \zeta, \gamma \geq 0$, $\xi + \zeta + \gamma \equiv 1$, and for all $h \in V$,

$$\mathbb{E}_D[\xi(x)I(h(x) = +1) + \zeta(x)I(h(x) = -1)] \leq \eta$$

Then, in particular, we have:

$$\mathbb{E}_D[\xi(x)I(h_1(x) = +1) + \zeta(x)I(h_1(x) = -1)] \leq \eta$$

$$\mathbb{E}_D[\xi(x)I(h_1(x) = +1) + \zeta(x)I(h_2(x) = -1)] \leq \eta$$

Thus, by $I(h_1(x) \neq h_2(x)) \leq \min(I(h_1(x) = -1) + I(h_1(x) = -1), I(h_2(x) = +1) + I(h_2(x) = +1))$, adding the two inequalities up,

$$\mathbb{E}_D[(\xi(x) + \zeta(x))I(h_1(x) \neq h_2(x))] \leq 2\eta$$

Since

$$\rho_D(h_1, h_2) = \mathbb{E}_D I(h_1(x) \neq h_2(x)) \geq r$$

We have

$$\mathbb{E}_D[\gamma(x)I(h_1(x) \neq h_2(x))] = \mathbb{E}_D[(1 - \xi(x) - \zeta(x))I(h_1(x) \neq h_2(x))] \geq r - 2\eta$$

Thus,

$$\mathbb{E}_D[\gamma(x)] \geq \mathbb{E}_D[\gamma(x)I(h_1(x) \neq h_2(x))] \geq r - 2\eta$$

Hence $\mathbf{\Phi}_D(V, \eta) \geq r - 2\eta$. □

Lemma 18. *Given hypothesis set V and data distribution D over $\mathcal{X} \times \mathcal{Y}$, $0 < \lambda < \eta < 1$, if there exist $h_1, h_2 \in V$ such that $\rho_D(h_1, h_2) \geq 2\eta - \lambda$, then $\mathbf{\Phi}_D(V, \eta) + \lambda \leq \mathbf{\Phi}_D(V, \eta - \lambda)$.*

Proof. Suppose $(\xi_1, \zeta_1, \gamma_1)$ are nonnegative functions satisfying $\xi_1 + \zeta_1 + \gamma_1 \equiv 1$, and for all $h \in V$, $\mathbb{E}_D[\zeta_1(x)I(h(x) = +1) + \xi_1(x)I(h(x) = -1)] \leq \eta - \lambda$, and $\mathbb{E}_D\gamma_1(x) = \mathbf{\Phi}_D(V, \eta - \lambda)$. Notice by Lemma 17,$\mathbf{\Phi}_D(V, \eta - \lambda) \geq 2\eta - \lambda - 2(\eta - \lambda) = \lambda$.

Then we pick nonnegative functions $(\xi_2, \zeta_2, \gamma_2)$ as follows. Let $\xi_2 = \xi_1$, $\gamma_2 = (1 - \frac{\lambda}{\mathbf{\Phi}_D(V, \eta - \lambda)})\gamma_1$, and $\zeta_2 = 1 - \xi_2 - \gamma_2$. It is immediate that $(\xi_2, \zeta_2, \gamma_2)$ is a valid confidence rated predictor and $\zeta_2 \geq \zeta_1$, $\gamma_2 \leq \gamma_1$, $\mathbb{E}_D\gamma_2(x) = \mathbf{\Phi}_D(V, \eta - \lambda) - \lambda$. It can be readily checked that the confidence rated predictor $(\xi_2, \zeta_2, \gamma_2)$ has error guarantee η, specifically:

$$\begin{aligned}
& \mathbb{E}_D[\zeta_2(x)I(h(x) = +1) + \xi_2(x)I(h(x) = -1)] \\
\leq\ & \mathbb{E}_D[(\zeta_2(x) - \zeta_1(x))I(h(x) = +1) + (\xi_2(x) - \xi_1(x))I(h(x) = -1)] + \eta - \lambda \\
\leq\ & \mathbb{E}_D[(\zeta_2(x) - \zeta_1(x)) + (\xi_2(x) - \xi_1(x))] + \eta - \lambda \\
\leq\ & \lambda + \eta - \lambda = \eta
\end{aligned}$$

Thus, $\mathbf{\Phi}_D(V, \eta)$, which is the minimum abstention probability of a confidence-rated predictor with error guarantee η with respect to hypothesis set V and data distribution D, is at most $\mathbf{\Phi}_D(V, \eta - \lambda) - \lambda$. □

H Detailed Derivation of Label Complexity Bounds

H.1 Agnostic

Proposition 1. *In agnostic case, the label complexity of Algorithm 1 is at most*

$$\tilde{O}(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu^*(D) + \epsilon_k}(d\frac{\nu^*(D)^2}{\epsilon^2}\ln\frac{1}{\epsilon} + d\ln^2\frac{1}{\epsilon})),$$

where the $\tilde{O}$ notation hides factors logarithmic in $1/\delta$.

Proof. Applying Theorem 5, the total number of labels queried is at most:

$$c_4 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{\lceil \log(1/\epsilon) \rceil - k + 1}{\delta})) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k}(1 + \frac{\nu^*(D)}{\epsilon_k})$$

Using the fact that $\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256) \leq 1$, this is

$$
\begin{aligned}
& c_4 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln(\frac{\lceil \log(1/\epsilon) \rceil - k + 1}{\delta})) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} (1 + \frac{\nu^*(D)}{\epsilon_k}) \\
= \; & \tilde{O} \left(\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{\epsilon_k} + \ln \log(1/\epsilon)) \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu + \epsilon_k} (1 + \frac{\nu^*(D)^2}{\epsilon_k^2}) \right) \\
\leq \; & \tilde{O} \left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu^*(D) + \epsilon_k} \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (1 + \frac{\nu^*(D)^2}{\epsilon_k^2})(d \ln \frac{1}{\epsilon} + \ln \ln \frac{1}{\epsilon}) \right) \\
\leq \; & \tilde{O} \left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256)}{2\nu^*(D) + \epsilon_k} (d \frac{\nu^*(D)^2}{\epsilon^2} \ln \frac{1}{\epsilon} + d \ln^2 \frac{1}{\epsilon}) \right),
\end{aligned}
$$

where the last line follows as ϵ_k is geometrically decreasing. □

H.2 Tsybakov Noise Condition with $\kappa > 1$

Proposition 2. *Suppose the hypothesis class $\mathcal{H}$ and the data distribution D satisfies (C_0, κ)-Tsybakov Noise Condition with $\kappa > 1$. Then the label complexity of Algorithm 1 is at most*

$$
\tilde{O}(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})}{\epsilon_k^{\frac{1}{\kappa}}} \epsilon^{\frac{2}{\kappa} - 2} d \ln \frac{1}{\epsilon}),
$$

where the $\tilde{O}$ notation hides factors logarithmic in $1/\delta$.

Proof. Applying Theorem 5, the total number of labels queried is at most:

$$
c_5 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2}) + \ln(\frac{k_0 - k + 1}{\delta})) \phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2}
$$

Using the fact that $\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \leq 1$, we get

$$
\begin{aligned}
& c_5 \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2}) + \ln(\frac{k_0 - k + 1}{\delta})) \phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2} \\
\leq \; & \tilde{O} \left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})}{\epsilon_k^{\frac{1}{\kappa}}} \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} \epsilon_k^{\frac{2}{\kappa} - 2} d \ln \frac{1}{\epsilon} \right) \\
\leq \; & \tilde{O} \left(\sup_{k \leq \lceil \log(1/\epsilon) \rceil} \frac{\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256})}{\epsilon_k^{\frac{1}{\kappa}}} \epsilon^{\frac{2}{\kappa} - 2} d \ln \frac{1}{\epsilon} \right)
\end{aligned}
$$

□

H.3 Agnostic, Linear Classification under Log-Concave Distribution

We show in this subsection that in agnostic case, if $\mathcal{H}$ is the class of homogeneous linear classifiers in R^d, $D_{\mathcal{X}}$ is isotropic log-concave in R^d, then, our label complexity bound is at most

$$
O(\ln \frac{\epsilon + \nu^*(D)}{\epsilon} (\ln \frac{1}{\epsilon} + \frac{\nu^*(D)^2}{\epsilon^2})(d \ln \frac{\epsilon + \nu^*(D)}{\epsilon} + \ln \frac{1}{\delta}) + \ln \frac{1}{\epsilon} \ln \frac{\epsilon + \nu^*(D)}{\epsilon} \ln \ln \frac{1}{\epsilon})
$$

Recall by Lemma 14, we have $\phi(2\nu^*(D) + \epsilon_k, \epsilon_k/256) \leq C(\nu^*(D) + \epsilon_k) \ln \frac{\nu^*(D) + \epsilon_k}{\epsilon_k}$ for some constant $C > 0$. Applying Theorem 4, the label complexity is

$$
O(\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\frac{2\nu^*(D) + \epsilon_k}{\epsilon_k} \ln \frac{2\nu^*(D) + \epsilon_k}{\epsilon_k}) + \ln(\frac{\log(1/\epsilon) - k + 1}{\delta})) \ln \frac{2\nu^*(D) + \epsilon_k}{\epsilon_k} (1 + \frac{\nu^*(D)^2}{\epsilon_k^2}))
$$

This can be simplified to (treating 1 and $\frac{\nu^*(D)^2}{\epsilon_k^2}$ separately)

$$
\begin{aligned}
& O(\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} \ln \frac{\nu^*(D) + \epsilon_k}{\epsilon_k} (d \ln \frac{\nu^*(D) + \epsilon_k}{\epsilon_k} + \ln \frac{k_0 - k + 1}{\delta}) \\
& + \sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} \frac{\nu^*(D)^2}{\epsilon_k^2} \ln \frac{\nu^*(D) + \epsilon_k}{\epsilon_k} (d \ln \frac{\nu^*(D) + \epsilon_k}{\epsilon_k} + \ln \frac{k_0 - k + 1}{\delta})) \\
\leq \; & O(\ln \frac{1}{\epsilon} \ln \frac{\epsilon + \nu^*(D)}{\epsilon} (d \ln \frac{\epsilon + \nu^*(D)}{\epsilon} + \ln \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}) + \frac{\nu^*(D)^2}{\epsilon^2} \ln \frac{\epsilon + \nu^*(D)}{\epsilon} (d \ln \frac{\epsilon + \nu^*(D)}{\epsilon} + \ln \frac{1}{\delta})) \\
\leq \; & O(\ln \frac{\epsilon + \nu^*(D)}{\epsilon} (\ln \frac{1}{\epsilon} + \frac{\nu^*(D)^2}{\epsilon^2}) (d \ln \frac{\epsilon + \nu^*(D)}{\epsilon} + \ln \frac{1}{\delta}) + \ln \frac{1}{\epsilon} \ln \frac{\epsilon + \nu^*(D)}{\epsilon} \ln \ln \frac{1}{\epsilon})
\end{aligned}
$$

H.4 Tsybakov Noise Conditon with $\kappa > 1$, Linear Classification under Log-Concave Distribution

We show in this subsection that under (C_0, κ)-Tsybakov Noise Condition, if $\mathcal{H}$ is the class of homogeneous linear classifiers in R^d, and $D_{\mathcal{X}}$ is isotropic log-concave in R^d, our label complexity bound is at most

$$
O(\epsilon^{\frac{2}{\kappa} - 2} \ln \frac{1}{\epsilon} (d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}))
$$

Recall by Lemma 14, we have$\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \leq C \epsilon_k^{\frac{1}{\kappa}} \ln \frac{1}{\epsilon_k}$ for some constant $C > 0$. Applying Theorem 5, the label complexity is:

$$
O(\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2}) + \ln(\frac{k_0 - k + 1}{\delta})) \phi(C_0 \epsilon_k^{\frac{1}{\kappa}}, \frac{\epsilon_k}{256}) \epsilon_k^{\frac{1}{\kappa} - 2})
$$

This can be simplified to :

$$
\begin{aligned}
& O(\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} (d \ln(\epsilon_k^{\frac{2}{\kappa} - 2} \ln \frac{1}{\epsilon_k}) + \ln(\frac{k_0 - k + 1}{\delta})) \epsilon_k^{\frac{2}{\kappa} - 2} \ln \frac{1}{\epsilon_k}) \\
\leq \; & O((\sum_{k=1}^{\lceil \log \frac{1}{\epsilon} \rceil} \epsilon_k^{\frac{2}{\kappa} - 2}) \ln \frac{1}{\epsilon} (d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})) \\
\leq \; & O(\epsilon^{\frac{2}{\kappa} - 2} \ln \frac{1}{\epsilon} (d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}))
\end{aligned}
$$