

Active Learning via Regression Beyond Realizability

Atul Ganju
Cornell University
ag2222@cornell.edu

Shashaank Aiyer*
Cornell University
saa244@cornell.edu

Ved Sriraman*
Cornell University
vs346@cornell.edu

Karthik Sridharan
Cornell University
sridharan@cs.cornell.edu

May 24, 2025

We present a new active learning framework for multiclass classification based on surrogate risk minimization that operates beyond the standard realizability assumption. Existing surrogate-based active learning algorithms crucially rely on realizability—the assumption that the optimal surrogate predictor lies within the model class—limiting their applicability in practical, misspecified settings. In this work we show that under conditions significantly weaker than realizability, as long as the class of models considered is convex, one can still obtain a label and sample complexity comparable to prior work. Despite achieving similar rates, the algorithmic approaches from prior works can be shown to fail in non-realizable settings where our assumption is satisfied. Our epoch-based active learning algorithm departs from prior methods by fitting a model from the full class to the queried data in each epoch and returning an improper classifier obtained by aggregating these models.

1 Introduction

We study active learning for multi-class classification in the statistical learning framework, where the learner has access to a large pool of unlabeled data and can query labels from an oracle. The objective is to learn an accurate classifier while minimizing the number of label queries—that is, to simultaneously achieve low excess classification risk and label complexity.

A promising recent approach to active learning is to reduce the problem to regression [KAH⁺21, SSSW23, ZN22, HY19]. Rather than directly optimizing the classification loss, these methods relax the problem to one of minimizing a convex, differentiable surrogate loss over a class of real-valued functions, thereby shifting the objective to approximating a real-valued target function whose associated classifier is expected to perform well under the original classification loss. This formulation enables the hypothesis class to encode additional structure—such as smoothness, margin, or regularization—that reflects assumptions about the target function, ultimately enabling the learning of richer, more expressive model classes. Furthermore, the differentiability of the surrogate facilitates the use of efficient gradient-based optimization techniques, which are essential for training complex models in modern machine learning systems.

While surrogate-based methods have become a cornerstone of modern learning systems due to their flexibility and compatibility with gradient-based optimization, their theoretical guarantees—both in the active setting [KAH⁺21, SSSW23, ZN22, HY19] and even in the passive setting [BJM06]—have critically relied on the *realizability* assumption: that the minimizer of the surrogate risk lies within the function class accessible to the learner. However, assuming that a setting is realizable is often too strong to hold in practice. To address this limitation, we develop a regression-based active learning algorithm that matches the accuracy and label complexity of prior realizability-based methods, while operating under conditions strictly milder than realizability.

To this end, we first establish a passive learning result that extends the realizability-based theory of [BJM06]. Specifically, we show that for convex function classes, there exists a broad family of classification-calibrated surrogate losses for which the excess classification risk can be upper bounded by the excess surrogate risk, provided a structural condition holds: namely, that the bias of the best-in-class function is lower bounded by a non-decreasing function of the bias of the surrogate risk minimizer. To demonstrate the generality of this condition, we construct an example in which realizability, and even approximate realizability, fails, yet

*Equal contribution.

our condition remains satisfied. Under additional distributional assumptions—such as those introduced by Massart and Tsybakov—our analysis yields sharper convergence rates, analogous to those obtained in the realizable setting by [BJM06].

While this extension is conceptually simple, it serves as the foundation for our main contribution. Specifically, we design a novel, improper active learning algorithm for multi-class classification. Our algorithm achieves guarantees comparable to existing realizability-dependent, surrogate-based methods yet operates under a natural extension of our passive learning assumption that is strictly weaker than realizability. In contrast, existing algorithms such as those of [HY19, ZN22] fundamentally rely on realizability, and their analyses break down in the misspecified setting. Our algorithm is the first to offer surrogate-based active learning guarantees in such generality.

Related Work: The study of active learning initially focused on the noise-free setting, where the label distribution is assumed to be perfectly consistent with some classifier in a known hypothesis class [FSST97, Das04, Das05, HY15]. Foundational contributions in this regime include the general label complexity bounds for arbitrary hypothesis classes established by [Das05], and the minimax rates derived by [HY15].

As the noise-free setting was increasingly recognized as too restrictive, attention shifted to the agnostic setting, where no assumptions are made on the label distribution. The work of [BBL06] introduced the disagreement-based framework and proposed the first algorithm for agnostic active learning with a nontrivial label complexity bound, as was later analyzed in [Han07]. However, subsequent work showed that in the fully agnostic regime, active learning cannot, in general, significantly outperform passive learning [Kä06, Hsu10, Han14].

To circumvent this limitation, a line of research focused on characterizing label complexity as a function of distributional noise. This led to significant improvements in label complexity under structured noise models [CN08, Han09, Kol10, Han11, LCK17]. A particularly notable result by [HY15] established minimax label complexity bounds for a broad class of noise models, demonstrating that under Tsybakov noise, active learning can provably outperform passive learning for any VC class.

Although these results establish strong theoretical foundations, they rely on idealized algorithmic primitives such as classification oracles and are limited to relatively simple hypothesis classes. To address these shortcomings, a series of works explored surrogate loss minimization as an alternative framework for active learning, with several promising results under the realizability assumption [AT07, Min12, HY19, KAH⁺21, ZN22].

Among these, the work of [HY19] is most directly related to ours. Their key insight is that once the sign of the optimal predictor is known at a given point, its exact value becomes irrelevant for classification. This allows their algorithm to focus exclusively on the region where the sign remains uncertain. In particular, it operates by maintaining a version space that is progressively refined by optimizing the empirical surrogate risk over this region of uncertainty. As such, their algorithm obtains a good approximation of the optimal predictor and therefore is able to infer the sign at additional inputs with increasing confidence. Although this insight still holds in the misspecified setting, their algorithmic approach fundamentally relies on the realizability assumption. In fact, the authors remark that identifying a concise, *a priori* condition under which their guarantees would extend to the non-realizable case would require a substantially different algorithmic approach. In this respect, our algorithm departs significantly from theirs: rather than refining a single version space, we employ an improper learning strategy that constructs a piecewise aggregation of classifiers, each operating on a region of the input space where it maintains high confidence.

Also relevant are the works of [BJM06], [LRS15], and [ZN22], each of which contributes essential mathematical ingredients to our analysis. The results of [BJM06] provide the foundation for bounding classification excess risk in terms of surrogate excess risk. The work of [LRS15] offers sharp bounds on the excess surrogate risk via localized Rademacher complexity. Finally, [ZN22] introduce a framework for bounding label complexity in terms of the disagreement coefficient. As both the results of [HY19] and [ZN22] are for surrogate-minimization based active binary classification, we provide a more detailed comparison with their results later in the paper.

Further related is the recent work of [SSSW23], which develops surrogate-based regression algorithms in the setting of selective sampling—an online variant of active learning in which the data generation process

may be arbitrary, even adversarial. Separately, a number of works have investigated nonparametric active learning under smoothness and regularity assumptions [LCK17, KYZ22, Min12, Han17].

2 Preliminaries

Regression-Based Classification: We consider the problem of surrogate-based multi-class classification in the statistical learning framework, where input instances belong to a set \mathcal{X} and their labels belong to the finite set $\mathcal{Y} = [K]$ for $K \geq 2$. The learner is given access to a sample $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ drawn i.i.d. from an unknown distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. The learner selects a function f from a class of real-valued functions $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R}^K\}$. A classifier $h_f : \mathcal{X} \rightarrow \mathcal{Y}$ is obtained by applying a fixed decision rule to functions in \mathcal{F} (e.g., argmax), inducing the hypothesis class $\mathcal{H}_{\mathcal{F}} := \{h_f : f \in \mathcal{F}\}$. The learner's objective is to minimize excess classification risk, defined as

$$\mathcal{E}_{0-1}(f, \mathcal{F}) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_f(x) \neq y\}] - \inf_{f' \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_{f'}(x) \neq y\}].$$

We denote $\mathcal{E}_{0-1}(f) := \mathcal{E}_{0-1}(f, \mathcal{F})$ when \mathcal{F} is the class of all measurable f . Since the 0–1 loss is neither convex nor continuous, we instead optimize a convex, differentiable surrogate loss $\ell : \mathbb{R}^K \times \mathcal{Y} \rightarrow \mathbb{R}$ over \mathcal{F} minimizing the excess surrogate risk, defined as

$$\mathcal{E}_{\ell}(f, \mathcal{F}) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f(x), y)] - \inf_{f' \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f'(x), y)],$$

in hopes it serves as a tractable proxy for the excess classification risk. We also denote $h^* = h_{f^*}$, for $f^* := \text{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f(x), y)]$, to be the classifier induced by the function in \mathcal{F} that achieves the minimum surrogate risk.

In this work, we consider surrogate loss functions $\ell_{\Phi} : \mathbb{R}^K \times \mathcal{Y} \rightarrow \mathbb{R}$ of the form

$$\ell_{\Phi}(v, y) = \Phi(v) - v[y],$$

where $\Phi : \mathbb{R}^K \rightarrow \mathbb{R}$ is a β_{Φ} -strongly convex, L_{Φ} smooth function in its first argument over the set of realizable score vectors (see [Aga13] for more details). That is, for all $x, x' \in \mathcal{X}$ and all $f \in \mathcal{F}$,

$$\frac{\beta_{\Phi}}{2} \|f(x) - f(x')\|_2^2 \leq \Phi(f(x)) - \Phi(f(x')) - \langle \nabla \Phi(f(x')), f(x) - f(x') \rangle \leq \frac{L_{\Phi}}{2} \|f(x) - f(x')\|_2^2.$$

Each such loss function admits a link function $\phi : \mathbb{R}^K \rightarrow \Delta^K$, given by $\phi(v) = \nabla \Phi(v)$, which maps the minimizer of the surrogate loss over \mathcal{D} , denoted by f_{η} , to the conditional probability vector $\eta(x) = (\mathbb{P}[Y = c \mid X = x])_{c \in [K]}$. Moreover, the β_{Φ} -strong convexity of Φ ensures that ℓ_{Φ} is strongly convex in its first argument, and the L_{Φ} smoothness of Φ ensures the link function ϕ is L_{Φ} -Lipschitz on the set of realizable score vectors. We then define the regression-based classifier h_f , for any f , as:

$$h_f(x) := \mathbf{e}_{c_f(x)} \quad \text{where} \quad c_f(x) = \underset{c \in K}{\text{argmax}} \phi(f(x))[c],$$

ensuring that the Bayes optimal predictor under the surrogate loss induces an optimal classifier, i.e. that $f_{\eta} \in \text{argmin}_f \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_f(x) \neq y\}]$. Below we instantiate our framework for a couple of common surrogate losses.

- *Squared-Loss:* If we select $\Phi(v) = \frac{1}{2} \|v\|_2^2$, the surrogate loss is equivalent to the squared loss, i.e., $\ell_{sq}(\hat{y}, y) = \|\hat{y} - \mathbf{e}_y\|_2^2$. In this case, the link function ϕ is the identity function, i.e., $\phi(v) = v$. In this case, the Bayes optimal predictor is simply the conditional probability function (i.e., $f_{\eta} = \eta$).
- *Logistic-Loss:* If we select $\Phi(v) = \log(\sum_{j=1}^K e^{v[j]})$, the surrogate loss is equivalent to the logistic loss, i.e., $\ell_{log}(\hat{y}, y) = -\log(\hat{y}[y])$. In this case, the link function is the Boltzmann mapping, i.e., $\phi(v)[i] = e^{v[i]} / \sum_{j=1}^K e^{v[j]}$.

We assume access to an offline regression oracle for the surrogate loss, defined as follows:

Definition 2.1 (Offline Regression Oracle). *Given a class of functions \mathcal{F} an offline regression oracle is specified by mapping $\text{Alg}_{\ell_\Phi}^{\text{OR}} : \cup_{t=0}^\infty (\mathcal{X} \times \mathcal{Y})^t \mapsto \mathcal{F}$ and is such that for any distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ and any n , given sample $S = \{(x_i, y_i)\}_{i \in [n]}$ of n data drawn i.i.d. from this distribution \mathcal{D} , the output of the oracle $\hat{f} = \text{Alg}_{\ell_\Phi}^{\text{OR}}(S)$ is such that with probability at least $1 - \delta$ over draw of samples,*

$$\mathcal{E}_{\ell_\Phi}(\hat{f}, \mathcal{F}) \leq \frac{\text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K)}{n}.$$

Learning Protocols: We consider two standard statistical learning protocols for multi-class classification: the passive learning protocol and the active learning protocol.

Passive Learning: In the passive learning setting, the learner receives the labeled sample S drawn i.i.d. from an unknown distribution \mathcal{D} . Given a function class \mathcal{F} and a surrogate loss ℓ_Φ adhering to the specifications above, the learner's objective is to output a function \hat{f} that achieves low excess classification risk $\mathcal{E}_{0-1}(\hat{f})$ as defined earlier.

Active Learning: In the active learning setting, the learner has access to an unlabeled sample $U = \{x_1, \dots, x_n\}$ drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, the marginal distribution over \mathcal{X} . The learner may adaptively query the label y_i of any example $x_i \in U$. Then, given access to a function class \mathcal{F} and a surrogate loss ℓ_Φ , the learner's objective is to output a function \hat{f} that achieves low classification excess risk $\mathcal{E}_{0-1}(\hat{f})$ while simultaneously minimizing the number of queries N it makes to the labeling oracle.

Additional Notation and Definitions: For any probability vector $v \in \Delta^K$ and decision $c \in [K]$, we define:

$$\begin{aligned} \text{margin}(v) &:= v[c_v] - \max_{c' \neq c_v} v[c'] \quad \text{where } c_v = \underset{c'' \in [K]}{\text{argmax}} v[c''] \\ \text{gap}(v, c) &:= \max_{c'} v[c'] - v[c]. \end{aligned}$$

Here, $\text{margin}(v)$ quantifies the strength of the decision given by the probability vector v by measuring how much its largest value exceeds the next best alternative. Meanwhile, $\text{gap}(v, c)$ expresses how much the probability vector favors its top choice over a specific class c , capturing the bias toward the decision relative to any given alternative.

Definition 2.2 (Massart Noise Condition, [MN06]). *The marginal distribution $\mathcal{D}_\mathcal{X}$ satisfies the Massart noise condition with parameter $\gamma \in [0, \frac{1}{2}]$ if $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}[\text{margin}(\eta(x)) < \gamma] = 0$.*

Definition 2.3 (Tsybakov Noise Condition, [Tsy04]). *The marginal distribution $\mathcal{D}_\mathcal{X}$ satisfies the Tsybakov noise condition with parameter $\beta \geq 0$ and a universal constant $c > 0$ if $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}[\text{margin}(\eta(x)) < \gamma] \leq c\gamma^\beta$ for any $\gamma > 0$.*

Definition 2.4 (Pseudo Dimension, [Pol84]; [Hau92, Hau95]). *Consider a set of real-valued functions $\mathcal{F} : \mathcal{X} \rightarrow [0, 1]$. The pseudo-dimension $\text{Pdim}(\mathcal{F})$ of \mathcal{F} is defined as the VC dimension of the set of threshold functions $\{(x, \zeta) \mapsto \mathbb{1}(f(x) > \zeta) : f \in \mathcal{F}\}$.*

3 Prelude: Passive Binary Classification

In this section, we provide bounds on excess classification risk in terms of excess surrogate risk, directly using this result to guarantee the performance of our active learning algorithm. Although we focus on the case of binary classification via squared-error regression, our analysis easily extends to the general learning framework provided in [Section 2](#) and we provide the proof in [Appendix A.1](#).

We can equivalently express the problem of binary classification using a single-coordinate formulation by considering a function class where each function models the probability that the label of x is class 1. Specifically, we redefine the label space as $\mathcal{Y} = \{0, 1\}$ and have that \mathcal{F} consists of functions $f : \mathcal{X} \rightarrow [0, 1]$ which, rather than predicting the full conditional probability vector, estimate $\eta(x) = \mathbb{P}[Y = 1|X = x]$.¹ Under this

¹We overload notation for $\eta(x)$. In this special case, it refers to the conditional probability of the label of x being class 1, but for our more general framework it refers to the conditional probability vector from [Section 2](#)

formulation, the regression-based classifier induced by any $f \in \mathcal{F}$ is given by $h_f(x) = \mathbb{1}\{f(x) > \frac{1}{2}\}$ and the squared loss simplifies to $\ell_{sq}(\hat{y}, y) = (\hat{y} - y)^2$.

To obtain meaningful bounds on classification excess risk, we assume that the classifier induced by the best-in-class function agrees with that of the Bayes optimal predictor under squared loss, and that the confidence of the best-in-class function bears a nondecreasing relationship to the confidence of the Bayes optimal predictor.

Assumption 1. *For conditional probability function $\eta(x)$ and function class \mathcal{F} :*

1. $\mathbb{P}_{x \sim \mathcal{D}_X}[h_{f^*}(x) = h_{f_\eta}(x)] = 1$,
2. *There exists a non-decreasing function $\psi : [0, \frac{1}{2}] \rightarrow [0, \frac{1}{2}]$ such that:*

$$\mathbb{P}_{x \sim \mathcal{D}_X} \left[\left| f^*(x) - \frac{1}{2} \right| \geq \psi \left(\left| \eta(x) - \frac{1}{2} \right| \right) \right] = 1.$$

Under this assumption, we are able to prove the following bound on the excess classification risk of a function in terms of its excess surrogate risk.

Proposition 1. *For any convex function class \mathcal{F} , if [Assumption 1](#) holds, then for any $f \in \mathcal{F}$,*

$$\mathcal{E}_{0-1}(f) \leq 2 \inf_{\gamma} \left\{ \mathcal{E}_{sq}(f, \mathcal{F}) \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_X} \left[\left| \eta(x) - \frac{1}{2} \right| \leq \gamma \right] \right\}$$

Proof. Starting with a standard analysis of excess risk under [Assumption 1.1](#), we have:

$$\mathcal{E}_{0-1}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_f(x) \neq y\}] - \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_{f_\eta}(x) \neq y\}] = \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h^*(x)\} |2\eta(x) - 1|]$$

Now, splitting on the event of a γ margin on $\eta(x)$ and upper bounding the small-margin case by the maximum margin size times the probability of a data point falling within the margin, we obtain:

$$\leq 2 \mathbb{E}_{x \sim \mathcal{D}_X} [\mathbb{1}\{h_f(x) \neq h^*(x)\} \mathbb{1}\{|\eta(x) - \frac{1}{2}| > \gamma\} |\eta(x) - \frac{1}{2}|] + 2\gamma \mathbb{P}_{x \sim \mathcal{D}_X} [|\eta(x) - \frac{1}{2}| \leq \gamma].$$

To bound the first expectation, notice that if $h_f(x) \neq h^*(x)$ then $|f^*(x) - f(x)| \geq |f^*(x) - \frac{1}{2}|$, and by [Assumption 1.2](#), we know $|f^*(x) - \frac{1}{2}| \geq \psi(|\eta(x) - \frac{1}{2}|)$ and thus:

$$\begin{aligned} & \mathbb{E}_{x \sim \mathcal{D}_X} [\mathbb{1}\{h_f(x) \neq h^*(x)\} \mathbb{1}\{|\eta(x) - \frac{1}{2}| > \gamma\} |\eta(x) - \frac{1}{2}|] \\ & \leq \mathbb{E}_{x \sim \mathcal{D}_X} [\mathbb{1}\{|f^*(x) - f(x)| \geq \psi(|\eta(x) - \frac{1}{2}|)\} \mathbb{1}\{|\eta(x) - \frac{1}{2}| > \gamma\} |\eta(x) - \frac{1}{2}|], \end{aligned}$$

where we can upper bound the first indicator by the ratio of the terms being compared to get:

$$\begin{aligned} & \leq \mathbb{E}_{x \sim \mathcal{D}_X} \left[\mathbb{1}\{|\eta(x) - \frac{1}{2}| > \gamma\} \left| \eta(x) - \frac{1}{2} \right| \left(\frac{f(x) - f^*(x)}{\psi(|\eta(x) - \frac{1}{2}|)} \right)^2 \right] \\ & \leq \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{E}_{(x,y) \sim \mathcal{D}} [(f(x) - f^*(x))^2] \\ & \leq \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{E}_{(x,y) \sim \mathcal{D}} [(f(x) - y)^2 - (f^*(x) - y)^2] \\ & = \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathcal{E}_{sq}(f, \mathcal{F}), \end{aligned}$$

where the final inequality is true by our assumption that \mathcal{F} is convex. Putting this together with the noise term and optimizing over γ gives us our desired result. \square

This bound expresses excess classification risk as a balance between the excess squared-error risk and the probability mass near the decision boundary. The infimum over γ allows for the tightest tradeoff between these two terms and enables the bound to yield explicit rates when instantiated under margin-based noise models.

4 Multi-Class Active Learning via Regression

We now show it is possible to use our passive learning bounds on classification excess risk to obtain active learning algorithms for classification even for problem instances that are far from realizable.

Stream-Based Active Learning Algorithms: We study an online variant of active learning, where the learner receives instances sequentially and must decide in real time whether to query each label. In order to minimize the number of label queries, queries are concentrated in regions of uncertainty and therefore the data distribution observed by the learner differs from the underlying distribution. Formally, the learner's querying strategy is modeled by a function $q : \mathcal{X} \rightarrow \{0, 1\}$, which selects a subset $Q = \{x \in \mathcal{X} : q(x) = 1\}$ of inputs to query. The resulting observed distribution \mathcal{D}_Q is the renormalization of \mathcal{D} restricted to Q :

$$\mathcal{D}_Q(x) = \begin{cases} \frac{\mathcal{D}_{\mathcal{X}}(x)}{\mathbb{P}_{x' \sim \mathcal{D}_{\mathcal{X}}}[x' \in Q]} & \text{if } x \in Q, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that a stream-based algorithm can be used for active learning by providing the data points in the unlabeled sample U to the algorithm sequentially.

Active Multi-Class Classification Assumption: Since a stream-based active learning algorithm interacts with data drawn from modified distributions \mathcal{D}_Q determined by its querying strategy, assumptions made solely about the original distribution no longer suffice. We therefore posit a stronger condition, analogous to [Assumption 1](#), but required to hold for all modified distributions that may arise during the algorithm's execution:

Assumption 2. *There exists a nondecreasing function $\psi : [0, 1] \rightarrow [0, 1]$ such that, for any $Q \subseteq \mathcal{X}$, if we define $f_Q^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}_Q} [\ell_{\Phi}(f(x), y)]$, then for all $k \in [K]$ we have:*

$$\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\operatorname{gap}(\phi(f_Q^*(x)), c) \geq \psi(\operatorname{gap}(\phi(f_{\eta}(x)), c))] = 1.$$

Active Multi-Class Classification Algorithm: We now present our regression-based active learning algorithm for multi-class classification.

Our algorithm is an epoch-based, improper learning algorithm. During each epoch $m \in [M]$, the algorithm employs a consistent query condition q_{m-1} ensuring that all observed data points during the epoch are drawn from the same modified data distribution \mathcal{D}_m . Then, after enough data is observed during the epoch, the learner uses an offline regression oracle, formally defined in [Definition 2.1](#), to obtain a good approximation $\hat{f}_m \in \mathcal{F}$ of the optimal function f_m^* in \mathcal{F} over the modified data distribution, where we denote $f_m^* = f_{\mathcal{X}_m}^*$ for notational convenience.

Under [Assumption 2](#), whenever the classifiers induced by \hat{f}_m and f_m^* agree on a point $x \in \mathcal{X}_m$, the learner can safely assign a label. To facilitate this, the learner constructs a subset $\mathcal{F}_m \subseteq \mathcal{F}$ centered around \hat{f}_m , consisting of all functions consistent with the observed data in epoch m , such that f_m^* is contained in \mathcal{F}_m with high probability. Then, the algorithm can confidently predict the label of any point $x \in \mathcal{X}_m$ on which all classifiers induced by functions in \mathcal{F}_m agree. Then, the learner updates its query condition such that it only continues to query points $x \in \mathcal{X}_m$ for which there exists a pair of functions $f, f' \in \mathcal{F}_m$ such that $h_f(x) \neq h_{f'}(x)$.

Finally, the algorithm outputs the classifier \hat{h} which, for any input x , considers the smallest epoch $i \in [M]$ for which there did not exist a pair of functions $f, f' \in \mathcal{F}_i$ such that $h_f(x) \neq h_{f'}(x)$. If such an i exists, it outputs the classification of the consensus; otherwise, it defaults to $h_{\hat{f}_M}(x)$.

As can be seen above, the excess risk and the label complexity of [Algorithm 1](#) are directly related to the probability the query function is triggered. This probability can be shown to be bounded via the *Value Function Disagreement Coefficient* defined as follows.

Definition 4.1 (Value Function Disagreement Coefficient, [\[FRSLX20\]](#)). *For any $f^* \in \mathcal{F}$ and $\gamma_0, \epsilon_0 > 0$, let:*

$$\theta_{\text{val}}(\mathcal{F}, \gamma_0, \epsilon_0, f^*) = \sup_{\mathcal{D}_{\mathcal{X}}} \sup_{\gamma > \gamma_0, \epsilon > \epsilon_0} \left\{ \frac{\gamma^2}{\epsilon^2} \mathbb{P}_{\mathcal{D}_{\mathcal{X}}} \left(\exists f \in \mathcal{F} : \|f(x) - f^*(x)\|_2 > \gamma, \|f - f^*\|_{\mathcal{D}_{\mathcal{X}}} \leq \epsilon \right) \right\} \vee 1,$$

Algorithm 1 Active Learning in Epochs

```

1: Parameters:  $\delta \in (0, 1)$ 
2: Define  $\tau_m = 2^m - 1$ ,  $\tau_0 = 0$ , and  $q_0(x) = 1$  and  $B := C \log^3(n) \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n)$ .
3: for  $m = 1, \dots, M$  do
4:   for  $t = \tau_{m-1} + 1, \dots, \tau_m$  do
5:     Receive  $x_t$  for  $(x_t, y_t) \sim \mathcal{D}$ 
6:     if  $q_{m-1}(x_t) = 1$  then
7:       Query the label  $y_t$  of  $x_t$ 
8:     end if
9:   end for
10:  Compute estimate  $\hat{f}_m \leftarrow \text{Alg}_{\ell_\Phi}^{\text{OR}}(S_m, \mathcal{F})$  for  $S_m = \{(x_t, y_t) : q_{m-1}(x_t) = 1, t \in [\tau_{m-1} + 1, \tau_m]\}$ :
11:  Implicitly Construct Set  $\mathcal{F}_m := \left\{ f \in \mathcal{F} : \sum_{t=\tau_{m-1}+1}^{\tau_m} q_{m-1}(x_t) \|f(x_t) - \hat{f}_m(x_t)\|_2^2 \leq B \right\}$ 
12:  Update condition  $q_m(x) \leftarrow q_{m-1}(x) \cdot \mathbb{1} \{ \exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x) \}$ 
13: end for
14: return  $\hat{f}$  defined by:

```

$$\hat{f}(x) := \begin{cases} \hat{f}_i(x) & \text{if } q_M(x) = 0 \text{ and } i \text{ is the smallest index s.t. } \nexists f, f' \in \mathcal{F}_i, h_f(x) \neq h_{f'}(x), \\ \hat{f}_M(x) & \text{otherwise.} \end{cases}$$

where $\|f - f^*\|_{\mathcal{D}_X} := \mathbb{E}_{x \sim \mathcal{D}_X} [\|f(x) - f^*(x)\|_2^2]$. Also, $\theta_{\text{val}}(\mathcal{F}, \gamma) = \sup_{f^* \in \mathcal{F}, \epsilon > 0} \theta_{\text{val}}(\mathcal{F}, \gamma, \epsilon, f^*)$.

The following theorem provides a bound on the excess risk and query complexity of [Algorithm 1](#).

Theorem 4.2. *For any convex function class \mathcal{F} , if [Assumption 2](#) holds, then for the predictor \hat{f} returned by [Algorithm 1](#) using offline regression oracle in [Definition 2.1](#) as a subroutine, we have that with probability at least $1 - \delta$,*

$$\mathcal{E}_{0-1}(\hat{f}) \leq \tilde{\mathcal{O}} \left(\inf_{\gamma > 0} \left\{ \frac{L_\Phi \beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) \log \delta^{-1}}{n} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}[\text{margin}(\eta(x)) \leq \gamma] \right\} \right),$$

and simultaneously, the number of label queries is bounded as

$$N \leq \tilde{\mathcal{O}} \left(\inf_{\gamma > 0} \left\{ \frac{L_\Phi^2 \beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) \log \delta^{-1}}{\psi^2(\gamma)} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)) + n \mathbb{P}[\text{margin}(\eta(x)) \leq \gamma] \right\} \right),$$

where $\text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K)$ is the rate achieved by the offline regression oracle. The $\tilde{\mathcal{O}}$ hides constants, \log factors in comp , θ_{val} , and n , and $\log \log$ factors in δ .

We instantiate [Theorem 4.2](#) for the Tsybakov noise model for settings where [Assumption 2](#) holds for $\psi(x) = x$, an assumption far weaker than realizability, and provide bounds on labeled and unlabeled sample complexity in terms of the error rate. We provide a proof in [Appendix B.1](#).

Corollary 1. *For any convex function class \mathcal{F} , if [Assumption 2](#) holds for $\psi(x) = x$ and Tsybakov's noise condition for parameter $\beta \geq 0$, then for the predictor \hat{f} returned by [Algorithm 1](#) using the offline regression oracle in [Definition 2.1](#) as a subroutine, we have that with probability at least $1 - \delta$,*

$$n \leq \tilde{\mathcal{O}} \left(\text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) \epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1} \right),$$

and,

$$N \leq \tilde{\mathcal{O}} \left(\text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) \theta_{\text{val}}^{\frac{\beta}{\beta+2}} \epsilon^{-\frac{2}{\beta+1}} \log \delta^{-1} \right).$$

The \tilde{O} hides constants, log factors in $\text{comp}, \theta_{\text{val}}$, and ϵ , and $\log \log$ factors in δ .

Our Algorithm vs. Prior Methods: Table 1 compares the rates achieved by Algorithm 1 to those of state-of-the-art surrogate minimization-based active learning algorithms under the widely studied Tsybakov noise model. Notably, our sample complexity, like that of [HY19], is independent of the disagreement coefficient, and our label complexity exhibits a strictly better dependence on it than both [HY19] and [ZN22].² Furthermore, the dependence on the target error rate ϵ in both the sample and label complexity of our algorithm matches that of [HY19] and [ZN22].³

Algorithm	Assumption	Sample Complexity (n)	Label Complexity (N)
[HY19]	Realizability	$\tilde{O}\left(d\epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1}\right)$	$\tilde{O}\left(d\theta_{\text{sgn}}\epsilon^{-\frac{2}{\beta+1}} \log \delta^{-1}\right)$
[ZN22]	Realizability	$\tilde{O}\left(d\theta_{\text{val}}\epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1}\right)$	$\tilde{O}\left(d\theta_{\text{val}}\epsilon^{-\frac{2}{\beta+1}} \log \delta^{-1}\right)$
Algorithm 1	Assumption 2 where $\psi(x) = x$, convex \mathcal{F}	$\tilde{O}\left(d\epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1}\right)$	$\tilde{O}\left(d\theta_{\text{val}}^{\frac{\beta}{\beta+2}}\epsilon^{-\frac{2}{\beta+1}} \log \delta^{-1}\right)$

Table 1: Comparison of sample and label complexities of squared-error regression based active classification algorithms. In this table $d = \text{PDim}(\mathcal{F})$ and the \tilde{O} hides constants, log factors in d, θ , and ϵ , and $\log \log$ factors in δ .

Our Assumptions vs. Realizability: While it is evident that our assumptions are weaker than realizability, one might ask whether they remains qualitatively weaker than approximate versions of the realizability condition. In the binary case, when paired with Massart noise, Assumption 1 is implied under $\mathcal{L}_\infty(\mathcal{D}_\mathcal{X})$ ϵ -realizability (i.e., when $|f^*(x) - \eta(x)| \leq \epsilon$ for all $x \in \mathcal{X}$). We formalize this claim below and provide a proof in Appendix C.

Claim 1. *For $\gamma \geq \epsilon > 0$, if a squared-error regression problem instance is both $\mathcal{L}_\infty(\mathcal{D}_\mathcal{X})$ ϵ -realizable and satisfies Definition 2.2 for parameter γ then, Assumption 1 holds for the given problem and data distribution with $\psi(x) = (1 - \frac{\epsilon}{\gamma})x$.*

In fact, even when the problem satisfies Tsybakov’s noise condition and is only $\mathcal{L}_2(\mathcal{D}_\mathcal{X})$ ϵ -approximately realizable—that is, when $\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} [(f^*(x) - \eta(x))^2] \leq \epsilon$ —one can still obtain a version of Assumption 1 which holds outside a region of small probability measure. While Proposition 1 is shown under the original condition which is stated to hold almost surely, our analysis naturally extends to the setting where Assumption 1 is violated on a subset of the input space with measure at most a function of ϵ and we would pay this measure additively in our bounds.

Furthermore, there exist instances where the problem is far from approximately realizable (i.e., where ϵ is a constant), yet Assumption 2—and consequently, our guarantees—still hold. The following example illustrates such a setting, see Appendix C for more details.

Example 1. *Consider $\mathcal{X} = \{\pm \vec{e}_i : i \in [d]\}$ and $\mathcal{F} = \{x \mapsto \frac{1+w \cdot x}{2} : \|w\|_2 \leq 1\}$. Let $\mathcal{D}_\mathcal{X} = \text{Unif}[\mathcal{X}]$, and $\eta(x) = \frac{1+\vec{1} \cdot x}{2}$. For this example, Assumption 2 holds with $\psi(x) = d^{-1/2}x$. That is, for all $Q \subseteq \mathcal{X}$ that could be induced by Algorithm 1, for all $x \in Q$ we have:*

$$|f_Q^*(x) - \frac{1}{2}| \geq \psi(|\eta(x) - \frac{1}{2}|).$$

However, we also have that:

$$\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} [(f^*(x) - \eta(x))^2] = \left(\frac{1}{2} - \frac{1}{2}d^{-1/2}\right)^2,$$

i.e., the instance is not $\mathcal{L}_2(\mathcal{D}_\mathcal{X})$ ϵ -realizable for any $\epsilon < \left(\frac{1}{2} - \frac{1}{2}d^{-1/2}\right)^2$.

²Note that the analysis in [HY19] is expressed in terms of the *sign-based disagreement coefficient*, denoted θ_{sgn} , which was first introduced in [Han07].

³Although our guarantees are not expressed in terms of the Pseudo-dimension, it upper bounds the rate achieved by the ERM in the binary classification setting with squared loss. See Appendix B.2 for more details.

The above claim and example should convince the reader that our assumption is provably weaker than approximate versions of realizability and can be far from them. In fact, our assumption intuitively says that the optimal regression function is only not allowed to be close to the decision boundary in regions where the true label is fairly decisive. It does not preclude cases when the regression function is very confident in places where the true label is close to the margin. This is a natural requirement, as the primary source of excess classification error is when the true label has a clear bias that the regression solution is unable to capture.

The Hurdle of Non-Realizability: Previous works in surrogate-based active learning algorithms [HY19, ZN22] have all made the realizability assumption. Under realizability, in any epoch m , the minimizer of the surrogate loss in \mathcal{F} on \mathcal{D}_m is always the Bayes optimal predictor, which would be contained in \mathcal{F} . This consistency allowed for the learner to progressively aggregate data across epochs and refine a version space that, with high probability, contains the Bayes optimal predictor. The final classifier would then be chosen from \mathcal{F} in direct alignment with this version space.

In contrast, our algorithm is fundamentally *improper*. Since we make no realizability assumption, the surrogate loss minimizer within \mathcal{F} may not be the Bayes optimal predictor and can vary arbitrarily across epochs. This invalidates the version space construction central to prior approaches. Instead, to ensure robust performance under our new assumption, our algorithm treats each epoch independently and learns an approximation $\hat{f}_m \in \mathcal{F}$ to the local surrogate-optimal function f_m^* using only the labeled data observed during that epoch. The final classifier stitches together the epoch-wise approximations, assigning to each point the prediction of the earliest approximation that can be deemed correct with high confidence.

The Importance of Convexity of \mathcal{F} : [Assumption 2](#) alone is not sufficient to guarantee the success of our passive or active learning results, as demonstrated by the following simple example:

Example 2. Let $\gamma > 0$, $\delta > 0$, and define $\mathcal{X} = \{0\}$ with $\eta(0) = \frac{1}{2} + \gamma$. Let $\mathcal{F} = \{f, f^*\}$ where $f^*(0) = \frac{1}{2} + 2\gamma$ and $f(0) = \frac{1}{2} - \delta$. Then $\mathcal{E}_{0-1}(f, \mathcal{F}) = 2\gamma$, while $\mathcal{E}_{sq}(f, \mathcal{F}) \rightarrow 0$ as $\delta \rightarrow 0$.

This example illustrates that without additional structural assumptions on \mathcal{F} , excess classification risk may not be controlled by excess surrogate risk. We can reconcile this example by additionally requiring that there exists a constant $C > 0$ for which:

$$\mathbb{E}_{x \sim \mathcal{D}_X} [(f(x) - f^*(x))^2] \leq C \mathcal{E}_{sq}(f).$$

This is exactly the role convexity plays in the proof of [Proposition 1](#). We specifically posit convexity of \mathcal{F} as it is both the most tangible structural assumption under which this relationship provably holds and a standard one in the literature on surrogate loss minimization (e.g., [BJM06, HY19]). Furthermore, convex function classes include many widely used, expressive model families—such as linear predictors, generalized linear models, and kernel-based methods (e.g., RKHSs)—and are well-suited for optimization-based learning algorithms commonly employed in modern practice.

While we focus on convex \mathcal{F} , we note that our analysis lends itself to the weaker condition of star convexity around each optimal predictor f_m^* ; however, such conditions are significantly more difficult to verify in practice.

5 Discussion

Our results show that in the batch setting, for both the active and passive learning, even under relaxations of the stringent realizability assumption commonly made in regression-based classification literature, one can obtain effectively the same guarantees as proven under realizability. We note that a compelling part of our analysis of the performance of [Algorithm 1](#) directly incorporates our passive learning result as a black-box component. This shows that in general, bounds on excess classification risk in terms of excess surrogate risk under relaxations of realizability appear to be a gateway to relaxing realizability in other learning paradigms as well. We conclude with a discussion of avenues for future work:

Algorithms for Interactive Learning: Regression-based algorithms for the contextual bandit problem under realizability are provided in [FR20] for the worst case, and for instance dependent bounds, in

[FRSLX20]. One can ask the question of whether this realizability assumption can be relaxed with milder assumptions like the one we mention above for the multi-class setting. We do note however that since, in these reductions, one picks distributions over actions in an epoch using the current estimate of $\hat{f}_m(x)$, the modified distribution over context-action pairs is not only a function of the benchmark class, but also this estimate. One can still change the assumption to having modified data distributions over \mathcal{X} space but any distribution over actions. However, we are yet to carefully analyze the implications of such an assumption.

Exploring the Additional Power of Being Improper: We note that our active learning algorithm is improper and technically can work better than proper passive learning algorithms in certain scenarios. While we have toy examples that illustrate this, further principled investigation into when this happens could be interesting.

References

- [Aga13] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1220–1228, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- [AT07] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2), April 2007.
- [BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 65–72, 2006.
- [BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- [Bou02] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik, 2002.
- [CN08] Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.
- [Das04] Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.
- [Das05] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.
- [FR20] Dylan J. Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3199–3210, 2020.
- [FR23] Dylan J. Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making, 2023.
- [FRSLX20] Dylan J. Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *ArXiv*, abs/2010.03104:6, 2020.
- [FSST97] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
- [Han07] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 353–360, 2007.
- [Han09] Steve Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, pages 249–264, 2009.
- [Han11] Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.
- [Han14] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.
- [Han17] Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions, 2017.
- [Hau92] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.
- [Hau95] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

- [Hsu10] Daniel J. Hsu. *Algorithms for Active Learning*. PhD thesis, University of California, San Diego, 2010.
- [HY15] Steve Hanneke and Liu Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(109):3487–3602, 2015.
- [HY19] Steve Hanneke and Liu Yang. Surrogate losses in passive and active learning. *Electronic Journal of Statistics*, 13(2), January 2019.
- [KAH⁺21] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daume III, and John Langford. Active learning for cost-sensitive classification, 2021.
- [Kol10] Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.
- [KYZ22] Samory Kpotufe, Gan Yuan, and Yunfan Zhao. Nuances in margin conditions determine gains in active learning, 2022.
- [Kä06] Matti Kääriäinen. Active learning in the non-realizable case. In *Algorithmic Learning Theory, 17th International Conference, ALT 2006*, volume 4264 of *Lecture Notes in Computer Science*, pages 63–77. Springer, 2006.
- [LBW98] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.
- [LCK17] Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning, 2017.
- [LRS15] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015.
- [Min12] Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13:67–90, 2012.
- [MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5), October 2006.
- [Pol84] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.
- [RST17] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2), May 2017.
- [SSSW23] Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression. In *Advances in Neural Information Processing Systems*, 2023.
- [Tsy04] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
- [ZN22] Yinglun Zhu and Robert Nowak. Efficient active learning with abstention, 2022.

A Proof of Main Result (**Theorem 4.2**)

We restate **Theorem 4.2** for the reader's convenience. In the proof of this theorem, we denote the number of points whose label was queried during epoch m as k_m .

Theorem A.1 (Theorem 4.2 Restated). *For any convex function class \mathcal{F} , if **Assumption 2** holds, then for the predictor \hat{f} returned by **Algorithm 1** using offline regression oracle in **Definition 2.1** as subroutine, we have that with probability at least $1 - \delta$,*

$$\mathcal{E}_{0-1}(\hat{f}) \leq \tilde{\mathcal{O}} \left(\inf_{\gamma > 0} \left\{ \frac{L_{\Phi} \beta_{\Phi}^{-1} \text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \log \delta^{-1}}{n} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}[\text{margin}(\eta(x)) \leq \gamma] \right\} \right),$$

and simultaneously, the number of label queries is bounded as

$$N \leq \tilde{\mathcal{O}} \left(\inf_{\gamma > 0} \left\{ \frac{L_{\Phi}^2 \beta_{\Phi}^{-1} \text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \log \delta^{-1}}{\psi^2(\gamma)} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)) + n \mathbb{P}[\text{margin}(\eta(x)) \leq \gamma] \right\} \right),$$

where $\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K)$ is the rate achieved by the offline regression oracle. The $\tilde{\mathcal{O}}$ hides constants, log factors in comp , θ_{val} , and n , and $\log \log$ factors in δ .

Proof. We start by bounding the excess risk of the classifier outputted by **Algorithm 1** and then bound the number of queries it makes. Starting with the reformulation of excess risk from **Eq. (1)**, we have,

$$\mathcal{E}_{0-1}(h_{\hat{f}}) = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathbb{1}\{h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x))],$$

where by decomposing on the query condition, we get:

$$\begin{aligned} &= \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathbb{1}\{q_M(x) = 0, h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x))] \\ &\quad + \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathbb{1}\{q_M(x) = 1, h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x))]. \end{aligned}$$

We will now separately bound the excess risk incurred when the classifier would have chosen not to query, i.e. when $q_M(x) = 0$, and when it would have chosen to query, i.e. when $q_M(x) = 1$, under the intersection of the high probability events of **Lemma 2** and **Lemma 1**.

We begin by bounding the excess risk incurred when the classifier would have chosen not to query. For any $x \in \mathcal{X}$ such that $q_M(x) = 0$, from the definition of **Algorithm 1**, we have that $h_{\hat{f}}(x) = h_{\hat{f}_i}(x)$, where i is the earliest epoch such that every function $f \in \mathcal{F}_i$ is in consensus on x . By **Lemma 6**, we know that $f_i^* \in \mathcal{F}_i$ and therefore that $h_{f_i^*}(x) = h_{\hat{f}_i}(x) = h_{\hat{f}}(x)$. Finally, by **Assumption 2**, we know that $h_{f_i^*}(x) = h_{f_{\eta}}(x)$, implying that the classifier does not incur risk when it would not have queried.

We now bound the excess risk incurred when the classifier would have chosen to query. First note that this can be bounded by the probability that the algorithm queries a data point it encounters in the final epoch:

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathbb{1}\{q_M(x) = 1, h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x))] \leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1] \leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{M-1}(x) = 1].$$

Accordingly, in the case this probability is small, less than $\frac{4}{n_M} \log \left(\frac{M}{\delta} \right)$, then so is the algorithm's excess risk. Therefore, we are left to consider the case when this probability is at least $\frac{4}{n_M} \log \left(\frac{M}{\delta} \right)$. With this in mind, excess risk can be rewritten as:

$$\begin{aligned} &\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{q_M(x) = 1, h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x)) \right] \\ &= \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x)) \middle| q_M(x) = 1 \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1] \\ &= \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} \left[\mathbb{1}\{h_{\hat{f}}(x) \neq h_{f_{\eta}}(x)\} \text{gap}(\eta(x), c_{\hat{f}}(x)) \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1], \end{aligned}$$

where, by [Assumption 2](#), we have:

$$= \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_M}} \left[\mathbb{1}\{h_{\hat{f}}(x) \neq h_{f_m^*}(x)\} \text{gap}(\eta(x), c_f(x)) \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1].$$

Now, notice, the expectation term is simply the excess risk of the function \hat{f} as seen in [Eq. \(1\)](#). Therefore, we can invoke [Proposition 2](#) giving us that, for any $\gamma > 0$:

$$\begin{aligned} & \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_M}} \left[\mathbb{1}\{h_{\hat{f}}(x) \neq h_{f_m^*}(x)\} \text{gap}(\eta(x), c_f(x)) \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1] \\ & \leq \left(8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{k_m} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \right) \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1] \\ & \leq 8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{k_m} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_M(x) = 1] + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \end{aligned}$$

where, since $q_m(x) \leq q_{m-1}(x)$ by construction, we have:

$$\begin{aligned} & \leq 8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{k_m} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{M-1}(x) = 1] + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \\ & \leq 8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{k_m} \frac{\mathbb{E}[k_m]}{n_M} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \end{aligned}$$

Then, by the lower bound from [Lemma 1](#), we have that:

$$\leq 8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{\frac{1}{2} \mathbb{E}[k_m] - \log(\frac{M}{\delta})} \frac{\mathbb{E}[k_m]}{n_M} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma],$$

where since, $\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{M-1}(x) = 1] \geq \frac{4}{n_M} \log(\frac{M}{\delta}) \implies \mathbb{E}[k_m] \geq 4 \log(\frac{M}{\delta})$, and therefore:

$$\leq 8 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, k_m)}{\frac{1}{4} \mathbb{E}[k_m]} \frac{\mathbb{E}[k_m]}{n_M} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma]$$

Now, because comp is increasing in its third argument, we obtain that:

$$\begin{aligned} & \leq 32 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, n, K)}{n_M} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \\ & = 64 L_{\Phi} \beta_{\Phi}^{-1} \frac{\text{comp}(\mathcal{F}, \delta, n, K)}{n} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma], \end{aligned}$$

where in the last line we plug in $M = \log n$, and by extension $n_M = \frac{n}{2}$. Doing the same substitution for the other case and taking the maximum of the two recovers our claimed bound on excess risk.

We now bound the number of queries made by [Algorithm 1](#). We know by [Lemma 1](#) that,

$$\begin{aligned} N &= \sum_{m=1}^M k_m \\ &\leq \sum_{m=1}^M \left(\frac{3}{2} \mathbb{E}[k_m] + \log \frac{M}{\delta} \right) \\ &= \sum_{m=1}^M \frac{3}{2} n_m \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{m-1}(x) = 1] + \log \frac{M}{\delta}. \end{aligned}$$

Where, by plugging in the upper bound for $\mathbb{P}_{x \sim \mathcal{D}_X}[q_M(x) = 1]$ we get by [Lemma 5](#), we get:

$$\begin{aligned}
&\leq \sum_{m=1}^M \frac{3}{2} n_m \frac{4}{n_{m-1}} \left(\max \left\{ \frac{L_\Phi^2 (9C(\mathcal{F}, \delta, n, K) + 32\beta_\Phi^{-1} \text{comp}(\mathcal{F}, \delta, n, K))}{\psi^2(\gamma)} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)), \log \frac{M}{\delta} \right\} \right) \\
&\quad + \frac{3}{2} n_m \mathbb{P}_{x \sim \mathcal{D}_X}[\text{margin}(\eta(x)) \leq \gamma] + \log \frac{M}{\delta} \\
&\leq 12 \log n \left(\max \left\{ \frac{L_\Phi^2 (9C(\mathcal{F}, \delta, n, K) + 32\beta_\Phi^{-1} \text{comp}(\mathcal{F}, \delta, n, K))}{\psi^2(\gamma)} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)), \log \frac{\log n}{\delta} \right\} \right) \\
&\quad + \frac{3}{2} n \mathbb{P}_{x \sim \mathcal{D}_X}[\text{margin}(\eta(x)) \leq \gamma] + \log \frac{\log n}{\delta},
\end{aligned}$$

where in the last line we plug in $M = \log n$ and recognize that for any parameter setting $C(\mathcal{F}, \delta, n, K) = \tilde{\mathcal{O}}(\text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K))$ to recover our claimed bound on label complexity. \square

A.1 Bound on Multi-Class Classification Excess Risk

We generalize [Proposition 1](#) to the regression-based multi-class classification framework presented in [Section 2](#). This result is used as a blackbox component of our analysis of the labeled and unlabeled complexity of [Algorithm 1](#). Under the following assumption, we are able to prove the result below.

Assumption 3. *There exists a non-decreasing function $\psi : (\gamma, 1] \rightarrow [0, 1]$ such that, for all $c \in [K]$, we have:*

$$\mathbb{P}_{x \sim \mathcal{D}_X}[\text{gap}(\phi(f^*(x)), c) \geq \psi(\text{gap}(\phi(f_\eta(x)), c))] = 1.$$

Proposition 2. *For any convex function class \mathcal{F} , if [Assumption 3](#) holds, then for any $f \in \mathcal{F}$,*

$$\mathcal{E}_{0-1}(f) \leq \inf_{\gamma} \left\{ 4L_\Phi \beta_\Phi^{-1} \mathcal{E}_{\ell_\Phi}(f, \mathcal{F}) \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \mathbb{P}_{x \sim \mathcal{D}_X}[\text{margin}(\eta(x)) \leq \gamma] \right\}.$$

Proof. Starting with the definition of classification excess risk, we have:

$$\begin{aligned}
\mathcal{E}_{0-1}(f) &= \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_f(x) \neq y\}] - \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_{f_\eta}(x) \neq y\}] \\
&= \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\}(\mathbb{1}\{h_{f_\eta}(x) = y\} - \mathbb{1}\{h_f(x) = y\})] \\
&= \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x))].
\end{aligned} \tag{1}$$

Then, splitting on the event that $\eta(x)$ has a γ -gap with respect to the label $c_f(x)$, we have:

$$\begin{aligned}
&= \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\}] \\
&\quad + \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) \leq \gamma\}], \\
&= \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\}] \\
&\quad + \gamma \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) \leq \gamma\}],
\end{aligned}$$

where, since $h_f(x) \neq h_{f_\eta}(x)$, we know $\text{gap}(\eta(x), c_f(x)) \geq \text{margin}(\eta(x))$, and therefore:

$$\begin{aligned}
&= \mathbb{E}_{x \sim \mathcal{D}_X}[\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\}] \\
&\quad + \gamma \mathbb{P}_{x \sim \mathcal{D}_X}[\text{margin}(\eta(x)) \leq \gamma].
\end{aligned}$$

It remains to bound the first term. To do this, recall from [Assumption 3](#) that $\psi(\text{gap}(\eta(x), c_f(x))) \leq \text{gap}(\phi(f^*(x)), c_f(x))$. Then, since:

$$\begin{aligned}
\text{gap}(\phi(f^*(x)), c_f(x)) &\leq \text{gap}(\phi(f^*(x)), c_f(x)) + \text{gap}(\phi(f(x)), c_{f^*}(x)) \\
&\leq 2 \|\phi(f^*(x)) - \phi(f(x))\|_\infty
\end{aligned}$$

$$\leq 2 \|\phi(f^*(x)) - \phi(f(x))\|_2,$$

we have that:

$$\begin{aligned} & \mathbb{E}_{x \sim \mathcal{D}_X} [\mathbb{1}\{h_f(x) \neq h_{f_\eta}(x)\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\}] \\ & \leq \mathbb{E}_{x \sim \mathcal{D}_X} [\mathbb{1}\{2 \|\phi(f^*(x)) - \phi(f(x))\|_2 \geq \psi(\text{gap}(\eta(x), c_f(x)))\} \text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\}] \end{aligned}$$

where, since the ratio of the terms being compared in the first indicator is an upper bound on the indicator, we have:

$$\begin{aligned} & \leq \mathbb{E}_{x \sim \mathcal{D}_X} \left[\text{gap}(\eta(x), c_f(x)) \mathbb{1}\{\text{gap}(\eta(x), c_f(x)) > \gamma\} \left(\frac{2 \|\phi(f^*(x)) - \phi(f(x))\|_2}{\psi(\text{gap}(\eta(x), c_f(x)))} \right)^2 \right] \\ & \leq 4 \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{E}_{x \sim \mathcal{D}_X} [\|\phi(f^*(x)) - \phi(f(x))\|_2^2] \end{aligned}$$

where, since ϕ is L_Φ -Lipschitz over the set of realizable score vectors, we have:

$$\begin{aligned} & \leq 4 L_\Phi \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathbb{E}_{x \sim \mathcal{D}_X} [\|f^*(x) - f(x)\|_2^2] \\ & \leq 8 L_\Phi \beta_\Phi^{-1} \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \mathcal{E}_{\ell_\Phi}(f, \mathcal{F}) \end{aligned}$$

where the final inequality is because \mathcal{F} is convex. Putting this together with the noise term and optimizing over γ gives us our desired result. \square

A.2 Concentration Lemmas and Supporting Results

In the proof of our main result we bound the unlabeled and labeled sample complexities of [Algorithm 1](#) under a good event. In this section, we show that this good event happens with high probability. This good event is the event that for ever epoch $m \in [M]$:

1. the number of queries k_m made within the epoch is concentrated around its expectation, and
2. the empirical distance from any function $f \in \mathcal{F}$ to f_m^* on labeled data observed in epoch m is concentrated around its $L_2(\mathcal{D}_{\mathcal{X}_m})$ distance to f_m^* .

We independently show that each of these events happen with high probability – in [Lemma 1](#) and [Lemma 2](#) respectively – and therefore the intersection of these events also happens with high probability.

Lemma 1. *For all $m \in [M]$, with probability $1 - 2\delta$,*

$$\frac{1}{2} \mathbb{E}[k_m] - \log \frac{M}{\delta} \leq k_m \leq \frac{3}{2} \mathbb{E}[k_m] + \log \frac{M}{\delta}.$$

Proof. We provide a proof of the lower bound using the lower tail Chernoff bound. First, note that k_m is a Binomial random variable, as it is a sum of i.i.d. Bernoulli random variables each representing whether the learner queried on a round of epoch m . By the lower tail Chernoff bound for a sum of independent Bernoulli random variables, we have that for any $\epsilon \in (0, 1)$,

$$\Pr[k_m < (1 - \epsilon) \mathbb{E}[k_m]] \leq \exp\left(-\frac{\epsilon^2 \mathbb{E}[k_m]}{2}\right).$$

Setting the probability of this bad event to be δ/M , and union bounding over $m \in [M]$ gives us that with probability at least $1 - \delta$, for all $m \in [M]$, $k_m < (1 - \epsilon) \mathbb{E}[k_m]$ for $\epsilon = \sqrt{\frac{-2 \log(\delta/M)}{\mathbb{E}[k_m]}}$. Now, by the AM-GM inequality, we have,

$$\epsilon = \sqrt{\frac{-2 \log(\delta/M)}{\mathbb{E}[k_m]}} \leq \frac{1}{2} + \frac{\log(M/\delta)}{\mathbb{E}[k_m]},$$

where plugging this upper bound in for ϵ gives us our desired lower bound. The upper bound follows identically from using the upper tail Chernoff bound, also with probability at least $1 - \delta$, implying they happen simultaneously with probability at least $1 - 2\delta$. \square

Lemma 2. Fix any $\delta \in (0, 1)$. Then, for all $m \in [M]$ and any $f \in \mathcal{F}$, with probability $1 - \delta$, we have:

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f(x) - f_m^*(x)\|_2^2] \leq 2 \left(\frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t) = 1\} \|f(x_t) - f_m^*(x_t)\|_2^2 \right) + \frac{C(\mathcal{F}, \delta, n, K)}{k_m},$$

and

$$\frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t) = 1\} \|f(x_t) - f_m^*(x_t)\|_2^2 \leq 2 \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f(x) - f_m^*(x)\|_2^2] + \frac{C(\mathcal{F}, \delta, n, K)}{k_m},$$

where $C(\mathcal{F}, \delta, n, K) = C' \left(Kn \log^3(n) \text{Rad}_n^2(\mathcal{F}) + \log \left(\frac{K \log n}{\delta} \right) \right)$ for some absolute constant C' .

Proof. Take $m \in [M]$. Recall the definition of f_m^* to be the best-in-class function on the subdistribution induced by our query condition in epoch m . Then, for any $f \in \mathcal{F}$, consider the average squared distance between f and f_m^* over labeled data observed during epoch m . Since, each data point (x_t, y_t) of this epoch is sampled from \mathcal{D} , and its label y_t is observed exactly when $q_{m-1}(x_t) = 1$, we can imagine each data point whose label was observed as being sampled from \mathcal{D}_m — the original data distribution \mathcal{D} normalized after being restricted to the set \mathcal{X}_m . Therefore, if we denote the rounds for which [Algorithm 1](#) did query in epoch m as $t_1^m, \dots, t_{k_m}^m$ we have:

$$\begin{aligned} \frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t) = 1\} \|f(x_t) - f_m^*(x_t)\|_2^2 &= \frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{x_t \in \mathcal{X}_m\} \|f(x_t) - f_m^*(x_t)\|_2^2 \\ &= \frac{1}{k_m} \sum_{i=1}^{k_m} \|f(x_{t_i^m}) - f_m^*(x_{t_i^m})\|_2^2. \end{aligned}$$

Then, by applying the upper and lower bounds on this quantity from [Lemma 3](#), union bounding over all $m \in [M]$, and re-normalizing δ , we get our desired lemma except with the additive term $C(\mathcal{F}, \delta, k_m, K)$. Finally, we remark that since $C(\mathcal{F}, \delta, k_m, K)$ is an increasing function in its third argument we can replace k_m with n . \square

Lemma 3. Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow [0, 1]^K\}$, $n \geq 2$, and $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over samples $S = \{(x_i, y_i)\}_{i \in [n]}$ drawn i.i.d. from \mathcal{D} , the following inequalities hold for all $f, f' \in \mathcal{F}$:

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\|f(x) - f'(x)\|_2^2] \leq 2 \left(\frac{1}{n} \sum_{i=1}^n [\|f(x_i) - f'(x_i)\|_2^2] \right) + \frac{C(\mathcal{F}, \delta, n, K)}{n},$$

and,

$$\frac{1}{n} \sum_{i=1}^n \|f(x_i) - f'(x_i)\|_2^2 \leq 2 \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\|f(x) - f'(x)\|_2^2] + \frac{C(\mathcal{F}, \delta, n, K)}{n},$$

for $C(\mathcal{F}, \delta, n, K) = C' \left(Kn \log^3(n) \text{Rad}_n^2(\mathcal{F}) + \log \left(\frac{K \log n}{\delta} \right) \right)$ some absolute constant C' .

Proof. We directly apply the result from [Lemma 4](#) and extend to the multi-class setting. In particular, we consider $\mathcal{F}^k := \{f(\cdot)[k] : f \in \mathcal{F}\}$ to be the class of k -th coordinate functions corresponds to the functions in \mathcal{F} . Now, the guarantees from [Lemma 3](#) hold for all \mathcal{F}^k for each $k \in [K]$. Taking a union bound over all K coordinates gives our result. \square

Lemma 4 ([Bou02, RST17]). Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \rightarrow [0, 1]\}$, $n \geq 2$, and $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over samples $S = \{(x_i, y_i)\}_{i \in [n]}$ drawn i.i.d. from \mathcal{D} , the following inequalities hold for all $f, f' \in \mathcal{F}$:

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[(f(x) - f'(x))^2] \leq 2 \left(\frac{1}{n} \sum_{i=1}^n (f(x_i) - f'(x_i))^2 \right) + \frac{C(\mathcal{F}, \delta, n)}{n},$$

and,

$$\frac{1}{n} \sum_{i=1}^n (f(x_i) - f'(x_i))^2 \leq 2 \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[(f(x) - f'(x))^2] + \frac{C(\mathcal{F}, \delta, n)}{n},$$

for $C(\mathcal{F}, \delta, n) = C' \left(n \log^3(n) \text{Rad}_n^2(\mathcal{F}) + \log \left(\frac{\log n}{\delta} \right) \right)$ some absolute constant C' .

A.3 Bounding the Probability of Querying Via the Disagreement Coefficient

We show in our main result that the labeled sample complexity can be bounded given a bound on the probability that our query condition is triggered. The following lemma bounds this probability for any epoch in terms of the value-based disagreement coefficient.

Lemma 5. Under the high probability event of [Lemma 1](#), for all $m \in [M]$, we have:

$$\begin{aligned} \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_m(x) = 1] &\leq \frac{4}{n_m} \left(\max \left\{ \frac{L_{\Phi}^2(9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K))}{\psi^2(\gamma)} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)), \log \frac{M}{\delta} \right\} \right) \\ &\quad + \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\text{margin}(\eta(x)) \leq \gamma]. \end{aligned}$$

Proof. We consider two cases and independently prove a bound on our desired quantity for each case. Then, the final bound is the max over the two cases.

Case 1: If $\mathbb{E}[k_m] \leq 4 \log \left(\frac{M}{\delta} \right)$, then we have,

$$\mathbb{E}[k_m] = n_m \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1] \leq 4 \log \left(\frac{M}{\delta} \right)$$

which implies that,

$$\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_m(x) = 1] \leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1] \leq \frac{4 \log \left(\frac{M}{\delta} \right)}{n_m}.$$

Case 2: Otherwise, if $\mathbb{E}[k_m] > 4 \log \left(\frac{M}{\delta} \right)$, then by the construction of [Algorithm 1](#), we can decompose the query condition on epoch $m + 1$ as,

$$\begin{aligned} \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_m(x) = 1] &= \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1] \\ &= \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1, \text{margin}(\eta(x)) > \gamma] \\ &\quad + \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1, \text{margin}(\eta(x)) \leq \gamma], \end{aligned}$$

where we can upper bound the second term by the probability $x \sim \mathcal{D}$ falls inside the margin to get:

$$\begin{aligned} &\leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1, \text{margin}(\eta(x)) > \gamma] \\ &\quad + \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\text{margin}(\eta(x)) \leq \gamma]. \end{aligned}$$

Now, to bound the first term in this sum, we rewrite it as the following conditional probability:

$$\begin{aligned} &\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1, \text{margin}(\eta(x)) > \gamma] \\ &= \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), \text{margin}(\eta(x)) > \gamma \mid q_{m-1}(x) = 1 \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), \text{margin}(\eta(x)) > \gamma \mid x \in \mathcal{X}_m \right] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{m-1}(x) = 1] \\
&= \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), \text{margin}(\eta(x)) > \gamma] \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{m-1}(x) = 1].
\end{aligned}$$

To bound the first term in this product, we start by recalling that from [Lemma 6](#), we know $f_m^* \in \mathcal{F}_m$ for all $m \in [M]$. Then, for any $x \in \mathcal{X}_m$ for which $\exists f, f' \in \mathcal{F}_m$ such that $h_f(x) \neq h_{f'}(x)$, there must exist a function $f \in \mathcal{F}_m$ for which $\|f_m^*(x) - f(x)\|_2 \geq \frac{1}{L_{\Phi}} \|\phi(f_m^*(x)) - \phi(f(x))\|_2 \geq \frac{1}{L_{\Phi}} \text{margin}(\phi(f_m^*(x)))$. Furthermore, we know by [Assumption 2](#) that $\text{margin}(\phi(f_m^*(x))) > \psi(\text{margin}(\eta(x)))$. However, since f is in \mathcal{F}_m , by [Lemma 7](#) we also know an upper bound on $\|f - f_m^*\|_{\mathcal{D}_{\mathcal{X}_m}}^2$. Therefore, we can bound by the probability these two events happen simultaneously, to get:

$$\begin{aligned}
&\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), \text{margin}(\eta(x)) > \gamma] \\
&\leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}_m}} \left[\text{margin}(\eta(x)) > \gamma, \exists f \in \mathcal{F}_m : \|f(x) - f_m^*(x)\|_2 > \frac{\psi(\text{margin}(\eta(x)))}{L_{\Phi}}, \right. \\
&\quad \left. \|f - f_m^*\|_{\mathcal{D}_{\mathcal{X}_m}}^2 \leq \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m} \right].
\end{aligned}$$

Then, since ψ is a non-decreasing function, we have:

$$\begin{aligned}
&\leq \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}_m}} \left[\exists f \in \mathcal{F}_m : \|f(x) - f_m^*(x)\|_2 > \frac{\psi(\gamma)}{L_{\Phi}}, \right. \\
&\quad \left. \|f - f_m^*\|_{\mathcal{D}_{\mathcal{X}_m}}^2 \leq \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m} \right]
\end{aligned}$$

where by the definition of the disagreement coefficient, we have:

$$\begin{aligned}
&\leq \frac{L_{\Phi}^2}{\psi^2(\gamma)} \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m} \\
&\quad \cdot \theta_{\text{val}} \left(\mathcal{F}_m, \psi(\gamma), \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m}, f_m^* \right) \\
&\leq \frac{L_{\Phi}^2}{\psi^2(\gamma)} \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)).
\end{aligned}$$

Now, from [Lemma 1](#), we know that, with probability at least $1 - \delta$, the following inequality holds:

$$k_m \geq \frac{1}{2} \mathbb{E}[k_m] - \log\left(\frac{M}{\delta}\right) = \frac{1}{2} n_m \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{m-1}(x) = 1] - \log\left(\frac{M}{\delta}\right).$$

From this lower bound and our assumption, we have that,

$$\begin{aligned}
&\frac{L_{\Phi}^2}{\psi^2(\gamma)} \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{k_m} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)) \\
&\leq \frac{4L_{\Phi}^2}{\psi^2(\gamma)} \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{n_m \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [q_{m-1}(x) = 1]} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)).
\end{aligned}$$

Plugging this upper bound back in and simplifying gives us,

$$\begin{aligned}
&\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\exists f, f' \in \mathcal{F}_m \text{ s.t. } h_f(x) \neq h_{f'}(x), q_{m-1}(x) = 1, \text{margin}(\eta(x)) > \gamma] + \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [\text{margin}(\eta(x)) \leq \gamma] \\
&\leq \frac{4L_{\Phi}^2}{\psi^2(\gamma)} \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K)}{n_m} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)) + \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} [|\eta(x) - \frac{1}{2}| \leq \gamma].
\end{aligned}$$

Finally, by taking the max of the two bounds from both cases gives us our desired result. \square

A.4 Supporting Lemmas

In our main result, we show that bounding excess risk is contingent on showing that the version space \mathcal{F}_m contains the surrogate minimizer of the modified distribution f_m^* . Similarly, our bound on labeled sample complexity is also contingent on this being true. We now prove this holds true under our good event.

Lemma 6. *Under the high probability event of Lemma 2, with probability $1 - \delta$, it is true that for any $m \in [M]$, $f_m^* \in \mathcal{F}_m$.*

Proof. First recall that we denote t_i^m to be the i -th queried point in epoch m . Then, we can bound the empirical distance between f_m^* and \hat{f}_m on queried points in the m -th epoch by Lemma 2, to get:

$$\begin{aligned} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t) = 1\} \|\hat{f}_m(x_t) - f_m^*(x_t)\|_2^2 &= \sum_{i=1}^{k_m} \|\hat{f}_m(x_{t_i^m}) - f_m^*(x_{t_i^m})\|_2^2 \\ &\leq 2k_m \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f_m^*(x) - \hat{f}_m(x)\|_2^2] + C(\mathcal{F}, \delta, n, K). \end{aligned}$$

Then, since \mathcal{F} is convex, we have:

$$2k_m \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f_m^*(x) - \hat{f}_m(x)\|_2^2] + C(\mathcal{F}, \delta, n, K) \leq 4k_m \beta_\Phi^{-1} \mathcal{E}_{\ell_\Phi}(\hat{f}_m, \mathcal{F}) + C(\mathcal{F}, \delta, n, K),$$

where since this is just the excess risk of \hat{f}_m , we can apply the bound on the excess risk of the multi-regression oracle to get:

$$\begin{aligned} &\leq 4\beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, k_m, K) + C(\mathcal{F}, \delta, n, K) \\ &\leq 4\beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) + C(\mathcal{F}, \delta, n, K), \end{aligned}$$

where the final inequality is possible since comp_{ℓ_Φ} is an increasing function in its third argument. Plugging this back in gives us our desired result. \square

Our bound on labeled sample complexity is also contingent on the expected $L_2(\mathcal{D}_{\mathcal{X}_m})$ distance between \hat{f}_m and f_m^* being small. We prove this also holds true under our good event.

Lemma 7. *Under the high probability event of Lemma 2, for any $m \in [M]$ and $f \in \mathcal{F}_m$, we have,*

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f(x) - f_m^*(x)\|_2^2] \leq \frac{9C(\mathcal{F}, \delta, n, K) + 32\beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K)}{k_m}.$$

Proof. By Lemma 2, we have,

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} [\|f(x) - f_m^*(x)\|_2^2] \leq 2 \left(\frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_m(x_t) = 1\} \|f(x_t) - f_m^*(x_t)\|_2^2 \right) + \frac{C(\mathcal{F}, \delta, n, K)}{k_m}.$$

To bound the summation term, we apply a basic triangle inequality, $\|a - b\|_2^2 \leq 2\|a - c\|_2^2 + 2\|b - c\|_2^2$,

$$\begin{aligned} &\sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_m(x_t) = 1\} \|f(x_t) - f_m^*(x_t)\|_2^2 \\ &\leq \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_m(x_t) = 1\} \left(2\|f(x_t) - \hat{f}_m(x)\|_2^2 + 2\|\hat{f}_m(x) - f_m^*(x_t)\|_2^2 \right) \\ &\leq 4 \sup_{f' \in \mathcal{F}_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_m(x_t) = 1\} \|f'(x_t) - \hat{f}_m(x)\|_2^2, \end{aligned}$$

where, by the construction of Algorithm 1, we have a bound on the distance from any function in \mathcal{F}_m to \hat{f}_m ,

$$\leq 32\beta_\Phi^{-1} \text{comp}_{\ell_\Phi}(\mathcal{F}, \delta, n, K) + 4C(\mathcal{F}, \delta, n, K).$$

Finally, by plugging this bound back in, we achieve our desired result. \square

B Instantiating for Tsybakov's Noise Condition

In this section we prove [Corollary 1](#) and briefly discuss how we arrive at our results in Table 1.

B.1 Proof of [Corollary 1](#)

We restate [Corollary 1](#) for the reader's convenience. Note this is for a generic loss fitting the specifications from [Section 2](#).

Corollary 2 (Corollary 1 Restated). *For any convex function class \mathcal{F} , if [Assumption 2](#) holds for $\psi(x) = x$ and Tsybakov's noise condition for parameter $\beta \geq 0$, then for the predictor \hat{f} returned by [Algorithm 1](#) using the offline regression oracle in [Definition 2.1](#) as a subroutine, we have that with probability at least $1 - \delta$,*

$$n \leq \tilde{\mathcal{O}} \left(\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1} \right),$$

and,

$$N \leq \tilde{\mathcal{O}} \left(\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \theta_{\text{val}}^{\frac{\beta}{\beta+2}} \epsilon^{-\frac{2}{\beta+1}} \log \delta^{-1} \right).$$

The $\tilde{\mathcal{O}}$ hides constants, log factors in comp , θ_{val} , and ϵ , and $\log \log$ factors in δ .

Proof. Instantiating the bound from [Theorem 4.2](#) with $\psi(x) = x$ and Tsybakov's Noise Condition gives us:

$$\mathcal{E}_{0-1}(\hat{f}) = \frac{32 \beta_{\Phi}^{-1} L_{\Phi} \text{comp}(\mathcal{F}, \delta, n, K)}{\gamma n} + c \gamma^{\beta+1} + \frac{2}{n} \log \frac{\log n}{\delta},$$

where we purposefully take the tighter bound that appears in the proof of [Theorem 4.2](#).

Now, optimizing with respect to γ and plugging the optimal value back into our excess risk bound gives

$$\begin{aligned} \mathcal{E}_{0-1}(\hat{f}) &\leq \tilde{\mathcal{O}} \left(\left(\frac{\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K)}{n} \right)^{\frac{\beta+1}{\beta+2}} + \frac{1}{n} \log \delta^{-1} \right) \\ &\leq \tilde{\mathcal{O}} \left(\left(\frac{\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \log \delta^{-1}}{n} \right)^{\frac{\beta+1}{\beta+2}} \right). \end{aligned}$$

Now, upper bounding by the error rate ϵ and isolating n gives the first part of our result,

$$n \leq \tilde{\mathcal{O}} \left(\text{comp}_{\ell_{\Phi}}(\mathcal{F}, \delta, n, K) \epsilon^{-\frac{\beta+2}{\beta+1}} \log \delta^{-1} \right).$$

To compute the bound on labeled sample complexity, we again begin by plugging $\psi(x) = x$ and Tsybakov's Noise Condition into the label complexity bound from [Theorem A.1](#). We again use the tighter bound that appears in the analysis.

$$\begin{aligned} N &\leq 12 \log n \left(\max \left\{ \frac{L_{\Phi}^2 (9 C(\mathcal{F}, \delta, n, K) + 32 \beta_{\Phi}^{-1} \text{comp}(\mathcal{F}, \delta, n, K))}{\gamma^2} \theta_{\text{val}}(\mathcal{F}, \psi(\gamma)), \log \frac{\log n}{\delta} \right\} \right) \\ &\quad + \frac{3c}{2} n \gamma^{\beta} + \log \frac{\log n}{\delta} \\ &\leq \tilde{\mathcal{O}} \left(\frac{\text{comp}(\mathcal{F}, \delta, n, K) \theta_{\text{val}}(\mathcal{F})}{\gamma^2} + \log \delta^{-1} + n \gamma^{\beta} \right) \end{aligned}$$

Optimizing over γ and plugging the optimal value back into our label complexity bound gives us:

$$N \leq \tilde{\mathcal{O}} \left((\text{comp}(\mathcal{F}, \delta, n, K) \theta_{\text{val}}(\mathcal{F}))^{\frac{\beta}{\beta+2}} n^{\frac{2}{\beta+2}} + \log \delta^{-1} \right),$$

where plugging the n from above back in gives us our desired result. \square

B.2 Discussion of Results in Table 1

We arrive at our results in Table 1 for binary classification with squared error by instantiating Corollary 1 for squared-error and bounding the excess risk of the offline regression oracle using the following bound for the ERM. Prior work shows that this quantity can be bounded in terms of the Covering Number defined below.

Definition B.1 (Covering Number). *V is an ℓ_2 cover of \mathcal{F} on x_1, \dots, x_n at scale β if for all $f \in \mathcal{F}$, there exists a collection of its elements such that the union of the β -balls with centers at the elements contains \mathcal{F} . That is, there exists $\mathbf{v}_f \in V$ such that*

$$\left(\frac{1}{n} \sum_{i=1}^n |f(x_i) - \mathbf{v}_f[i]| \right)^{\frac{1}{2}} \leq \beta.$$

The empirical covering number $\mathcal{N}_2(\mathcal{F}, \beta; x_1, \dots, x_n)$ is the size of the minimal set of such a V , and we define the covering number as

$$\mathcal{N}_2(\mathcal{F}, \beta, n) = \sup_{x_1, \dots, x_n} \mathcal{N}_2(\mathcal{F}, \beta; x_1, \dots, x_n)$$

In particular, we have the following bound.

Lemma 8. *[[LRS15]] For any convex function class \mathcal{F} with probability $1 - \delta$, if \hat{f} is the ERM, then*

$$\text{comp}_{sq}(\mathcal{F}, \delta, n) = n \log^3 n \log \frac{1}{\delta} \left(\inf_{\kappa > 0, \nu \in [0, \kappa]} \left(4\nu + \frac{12}{\sqrt{n}} \int_{\nu}^{\kappa} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \beta)} d\beta \right) + \frac{\log \mathcal{N}_2(\mathcal{F}, \kappa) + \log \frac{1}{\delta}}{n} \right)$$

for some absolute constant $C_1 > 0$.

Proof. This is a direct consequence of applying Lemma 7 to the upper bound of Theorem 4 of [LRS15], then upper bounding the resulting complexity by multiplying by an additional $\log^3 n$. \square

In key cases, the complexity measure from Lemma 8 simplifies:

- *Finite pseudo-dimension:* If \mathcal{F} is convex with finite pseudo-dimension (i.e. $\text{Pdim}(\mathcal{F}) < \infty$), then [LBW98] shows that $\text{comp}_{sq}(\mathcal{F}, \delta, n) = \mathcal{O}(\text{Pdim}(\mathcal{F}) \log n)$.
- *Bounded Covering Number:* If \mathcal{F} is convex and the covering number at scale κ satisfies the following $\log \mathcal{N}_2(\mathcal{F}, \kappa) \leq \kappa^{-p}$ for some $p > 0$, then [LRS15] shows that: $\text{comp}_{sq}(\mathcal{F}, \delta, n) = \mathcal{O}(n^{p/(2+p)})$ when $p \in (0, 2)$, and $\text{comp}_{sq}(\mathcal{F}, \delta, n) = n^{1-1/p}$ when $p \geq 2$ with an extra logarithmic factor when $p = 2$.

C Comparing Our Assumptions to Approximate Realizability

In this section, we prove our Claim 1 stating that Assumption 1 is implied under approximate realizability and Massart's Noise condition and provide a proof that our Example 1 is a problem instance satisfying our assumption that is in fact far from being approximately realizable.

C.1 Proof of Claim 1

We restate the Claim 1 for the reader's convenience.

Claim 2 (Claim 1 Restated). *For $\gamma \geq \epsilon > 0$, if a squared-error regression problem instance is both $\mathcal{L}_{\infty}(\mathcal{D}_{\mathcal{X}})$ ϵ -realizable and satisfies Definition 2.2 for parameter γ then, Assumption 1 holds for the given problem and data distribution with $\psi(x) = (1 - \frac{\epsilon}{\gamma})x$.*

Proof. To show [Assumption 1.1](#) holds, assume for contradiction that there exists x such that $h_{f^*}(x) \neq h_{f_\eta}(x)$. Then, it must be that $|f^*(x) - \eta(x)| \geq \gamma \geq \epsilon$, implying the problem is not $L_\infty(\mathcal{D}_\mathcal{X})$ ϵ -realizable.

Now for [Assumption 1.2](#), starting with the margin on f^* , we have by the triangle inequality that:

$$|f^*(x) - \tfrac{1}{2}| = |f^*(x) - \eta(x) + \eta(x) - \tfrac{1}{2}| \geq |\eta(x) - \tfrac{1}{2}| - |f^*(x) - \eta(x)| \geq |\eta(x) - \tfrac{1}{2}| - \epsilon,$$

where the final inequality is true since the problem is point-wise ϵ -realizable. Now, consider the event where $|\eta(x) - \tfrac{1}{2}| \geq \gamma$. If we factor out the bias of η , then under this event, which takes place with probability 1, we have:

$$|\eta(x) - \tfrac{1}{2}| - \epsilon = |\eta(x) - \tfrac{1}{2}| \left(1 - \epsilon |\eta(x) - \tfrac{1}{2}|^{-1}\right) \geq |\eta(x) - \tfrac{1}{2}| \left(1 - \frac{\epsilon}{\gamma}\right),$$

and therefore, [Assumption 1.2](#) holds with $\psi(x) = (1 - \frac{\epsilon}{\gamma})x$. \square

C.2 Proof of Correctness for [Example 1](#)

We restate [Example 1](#) for the reader's convenience.

Example 3 (Example 1 Restated). Consider $\mathcal{X} = \{\pm \vec{e}_i : i \in [d]\}$ and $\mathcal{F} = \{x \mapsto \frac{1+w \cdot x}{2} : \|w\|_2 \leq 1\}$. Let $\mathcal{D}_\mathcal{X} = \text{Unif}[\mathcal{X}]$, and $\eta(x) = \frac{1+\vec{1} \cdot x}{2}$. For this example, [Assumption 2](#) holds with $\psi(x) = d^{-1/2}x$. That is, for all $Q \subseteq \mathcal{X}$ that could be induced by [Algorithm 1](#), for all $x \in Q$ we have:

$$|f_Q^*(x) - \tfrac{1}{2}| \geq \psi(|\eta(x) - \tfrac{1}{2}|).$$

However, we also have that:

$$\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} [(f^*(x) - \eta(x))^2] = \left(\tfrac{1}{2} - \tfrac{1}{2}d^{-1/2}\right)^2,$$

i.e., the instance is not $\mathcal{L}_2(\mathcal{D}_\mathcal{X})$ ϵ -realizable for any $\epsilon < \left(\tfrac{1}{2} - \tfrac{1}{2}d^{-1/2}\right)^2$.

Proof. First, we claim that the only possible regions of uncertainty that can be induced by the algorithm are of the form $\{\pm e_i : i \in S\}$ for some $S \subseteq [d]$. To see this, note that for any $f \in \mathcal{F}$, $f(x) = -f(-x)$ and thus $h_f(x) \neq h_f(-x)$ for all $x \in \mathcal{X}$. Therefore, for any $f, f' \in \mathcal{F}$, $h_f(x) \neq h_{f'}(x)$ if and only if $h_f(-x) \neq h_{f'}(-x)$, directly implying our claim.

Now, if we consider a subset $Q \subseteq \mathcal{X}$ of the form given above, we have that the optimal regression function on this subset is given by:

$$w_Q^*[i] = \begin{cases} \left(\frac{|Q|}{2}\right)^{-\frac{1}{2}} & \text{if } e_i \in Q \\ 0 & \text{otherwise.} \end{cases}$$

and therefore, $(w_Q^*)^T x = (|Q|/2)^{-1/2}(\vec{1} \cdot x)$ for any $x \in Q$. As such, we have that:

$$|f_Q^*(x) - \tfrac{1}{2}| = \tfrac{1}{2} \left(\frac{|Q|}{2}\right)^{-\frac{1}{2}} \geq \tfrac{1}{2}d^{-1/2} = \psi(|\eta(x) - \tfrac{1}{2}|),$$

where the inequality holds true because $|Q| \leq 2d$ by construction and the final equality holds true because $|\eta(x) - \tfrac{1}{2}| = \tfrac{1}{2}$ by construction as well.

Alternatively, since $\eta(x) = 1$ exactly when $f^*(x) = \frac{1}{2} + \frac{1}{2}d^{-1/2}$ and $\eta(x) = 0$ exactly when $f^*(x) = \frac{1}{2} - \frac{1}{2}d^{-1/2}$, we have that $|\eta(x) - f^*(x)| = \frac{1}{2} - \frac{1}{2}d^{-1/2}$ for all x implying our claim on its approximate realizability. \square

D The Importance of Convexity of the Benchmark Class

In this section, we provide a proof that for any strongly convex surrogate loss, convexity of \mathcal{F} ensures the excess surrogate risk of a function is bounded by its $L_2(\mathcal{D}_X)$ distance to the surrogate risk minimizer in \mathcal{F} . Although the strong convexity constant of the squared loss is 1, for this case the factor of two can be shaved off (see Lemma 1 of [FR23] for example).

Lemma 9. *For a convex class \mathcal{F} and surrogate loss $\ell_\Phi : \mathbb{R}^K \times \mathcal{Y} \rightarrow \mathbb{R}$ that is β_Φ -strongly convex in its first argument, we have, for all $f \in \mathcal{F}$:*

$$\|f - f^*\|_{\mathcal{D}_X}^2 \leq \frac{2}{\beta_\Phi} \mathcal{E}_{\ell_\Phi}(f, \mathcal{F}).$$

Proof. By β_Φ -strong convexity of $\ell_\Phi(\cdot, y)$ over all realizable inputs, we have:

$$\ell_\Phi(f(x), y) \geq \ell_\Phi(f^*(x), y) + \langle \nabla \ell_\Phi(f^*(x), y), f(x) - f^*(x) \rangle + \frac{\beta_\Phi}{2} \|f(x) - f^*(x)\|_2^2.$$

Taking an expectation over $(x, y) \sim \mathcal{D}$ gives us:

$$\mathcal{E}_{\ell_\Phi}(f, \mathcal{F}) \geq \mathbb{E}_{(x, y) \sim \mathcal{D}} [\langle \nabla \ell_\Phi(f^*(x), y), f(x) - f^*(x) \rangle] + \frac{\beta_\Phi}{2} \|f - f^*\|_{\mathcal{D}_X}^2.$$

Since f^* minimizes surrogate risk over the convex set \mathcal{F} , the expectation is nonnegative implying our desired result. \square