Active and Passive Learning Via Regression: Beyond Realizability

Shashaank Aiyer SAA244@CORNELL.EDU
Atul Ganju AG2222@CORNELL.EDU
Karthik Sridharan KS999@CORNELL.EDU
Ved Sriraman VS346@CORNELL.EDU
Cornell University

Abstract

In this work, we consider the problem of binary classification in the statistical learning framework under both passive and active learning settings. We consider a regression based approach to solving these problems. Previous analysis using regression based approaches for solving the problems (e.g. Bartlett et al. (2006) for passive learning and Zhu and Nowak (2022) for active learning) all rely on the so called realizability assumption (or an approximate version) that assume that the class or regression model one uses for the problem is rich enough to contain the Bayes optimal predictor w.r.t. the squared loss used in regression. In this paper, we relax this assumption for both the active and passive learning settings and show that one can effectively obtain the same effective rates under conditions that are far milder than realizability.

1. Introduction

We consider the problem of binary classification in the statistical setting where input instances belong to a set $\mathcal{X}$ and their labels belong to the set $\mathcal{Y} = \{0, 1\}$. In this learning paradigm, the learner is given access to a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from a fixed unknown distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. The learner's objective is to then use this sample to output a function $\hat{h} : \mathcal{X} \to \{0, 1\}$ that has low expected classification error. A classical result in statistical learning theory shows that if the learner only knows that the optimal classifier belongs to some fixed hypothesis class $\mathcal{H}$ and no additional assumptions on the distribution are made, then the function class is learnable if and only if the VC dimension of $\mathcal{H}$ is finite (Vapnik and Alexey (1971); Valiant (1984)).

However, since function classes with infinite VC dimension are often used in practice, to facilitate learning, a common approach is to first impose additional structure onto the problem. Then, select a rich class of real-valued functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ designed to capture this structure, where each $f \in \mathcal{F}$ naturally induces a classifier h_f. Reformulating the problem in this way allows for the performance of any $f \in \mathcal{F}$ to be measured via a smooth, differentiable surrogate loss function $\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ that can be minimized through gradient-based optimization methods. Ultimately, the goal is to find algorithms outputting a classifier $\hat{f}$ with low classification excess risk, defined as:

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

As a concrete example, it is typical to choose $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ to model the conditional probability $\eta(x) = \mathbb{P}[Y = 1 \mid X = x]$. Then, each function $f \in \mathcal{F}$ induces the classifier $h_f(x) = \mathbb{1}\{f(x) > 1/2\}$ and its performance over $\mathcal{F}$ is measured using the squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$. In this case, the conditional probability η induces an optimal classifier h_η.

Under the rather strong assumption that the conditional probability η belongs to the class of regression functions $\mathcal{F}$, which is often referred to as realizability, one can show that classification excess risk can be upper bounded by a function of excess risk with respect to a surrogate loss (i.e., squared error excess risk) (Bartlett et al. (2006)). Remarkably, under this realizability assumption, one can also obtain results for a binary hypothesis class $\mathcal{H}$, induced by a class of regression functions $\mathcal{F}$, even when it has infinite VC dimension. As a result, regression based algorithms for classification are used extensively in practice. Moreover, with additional margin-based noise assumptions, such as the ones introduced by Massart and Tsybakov, it is possible to attain fast rates of convergence. In fact, the power of these assumptions extends to the active learning regime, where one obtains unlabeled samples drawn i.i.d. and the learner chooses points to be labeled. A recent result by Zhu and Nowak (2022) shows that with additional assumption that the class $\mathcal{F}$ has a small disagreement coefficient, it is possible to obtain an oracle efficient active learning algorithm that maintains performance guarantees comparable to passive learning while attaining optimal label complexity.

In this paper, we focus on relaxing this realizability assumption while still obtaining strong guarantees on classification excess risk when using regression based algorithms. Specifically, we introduce an assumption that ensures that (1) the classifier induced by the minimizer of expected square loss in the class $\mathcal{F}$ matches the Bayes Optimal Classifier and (2) the bias of the minimizer of expected square loss on the data distribution is lower bounded by a non-decreasing function of the bias of the Bayes Optimal Classifier. Then, we show that under this assumption, we can bound the classification excess risk in terms of squared error excess risk with respect to any convex class of regression functions $\mathcal{F}$. This yields results comparable to the ones in Bartlett et al. (2006) but without the realizability assumption imposed on the class $\mathcal{F}$. We further illustrate with an example that realizability, or even approximate realizability, can fail to hold in instances where our assumption above is satisfied. In the other direction, we also show that approximate realizability, along with a margin-based noise condition, implies our assumption, thus showing our assumption is in fact weaker than the realizability assumption. We instantiate our result to provide concrete bounds on classification excess risk for several families of function classes.

We then consider the problem of classification in the active learning setting (see, e.g., Hanneke (2014); Settles (2009)), which, as mentioned earlier, is a setting where the learner is given access to unlabeled data and can request labels for selected points. The goal of the learner, as before, is to output a classifier with low classification excess risk. However, the learner also aims to learn such a classifier while making as few label queries as possible. Most prior work in active learning adopts a *disagreement-based approach*, where the learning algorithm maintains a version space, containing all plausible models consistent with the data, and a region of uncertainty, where the models in the version space disagree. Predictions are made by evaluating a function in the version space, limiting errors to the region of uncertainty. The version space is designed to include the optimal classifier with high probability. Thus, when all models in the version space agree, the learner can confidently predict, knowing the optimal classifier aligns with the consensus. As more labeled data is acquired, the version space shrinks, reducing the region of uncertainty and refining predictions.

In Balcan et al. (2006); Zhang and Chaudhuri (2014), the version space is explicitly enumerated, whereas in Dasgupta and Freund (2023); Beygelzimer et al. (2009, 2010); Huang et al. (2024), it is constructed using a classification oracle. These methods, however, often face significant computational challenges. More recent works Krishnamurthy et al. (2021); Zhu and Nowak (2022); Sekhari et al. (2023) adopt a regression-based approach to define the version space, making the algorithms more practical, but they rely on the realizability assumption.

In this paper, we relax the realizability assumption when learning convex function classes and provide minimax classification excess risk bounds that match those of passive learning, while achieving label complexity comparable to Zhu and Nowak (2022). Specifically, we impose a relaxed assumption that extends the milder condition used in passive learning to all distributions induced by the query condition of our algorithm.

Our approach is also epoch-based and uses a regression oracle at the end of each epoch. However, unlike most prior disagreement-based analyses, the absence of the realizability assumption means we cannot maintain a version space guaranteed to contain the true underlying model across epochs. Instead, we perform regression separately on the region of uncertainty of each epoch and aggregate the induced classifiers. Our new approach, combined with our relaxed condition, allows us to overcome this limitation.

1.1. Preliminary Definitions

In this section, we introduce the key definitions that underpin our main results. First, we provide a useful characterization the complexity of the class $\mathcal{F}$.

Definition 1 (Covering Number) *V is an ℓ_2 cover of $\mathcal{F}$ on $x_1, \ldots, x_n$ 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 $\mathcal{N}_2(\mathcal{F}, \beta; x_1, \ldots, x_n)$ is the size of the minimal set of such a V, and we define the covering number as

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

In the regression setting, the "size" of $\mathcal{F}$ can be measured by the following well-known complexity measure, which we make use of in some of our examples:

Definition 2 (Pseudo Dimension, Pollard (1984); Haussler (1992, 1995)) *Consider a set of real-valued functions $\mathcal{F} : \mathcal{X} \to [0, 1]$. The pseudo-dimension $\mathrm{Pdim}(\mathcal{F})$ of $\mathcal{F}$ is defined as the VC dimension of the set of threshold functions $\{(x, \zeta) \mapsto \mathbb{1}(f(x) > \zeta) : f \in \mathcal{F}\}$.*

The notion of disagreement coefficient plays a role in characterizing the query complexity of active learning (see Zhu and Nowak (2022)). The value based disagreement coefficient is defined below:

Definition 3 (Value Function Disagreement Coefficient, Foster et al. (2020)) *For any $f^* \in \mathcal{F}$ and $\gamma_0, \epsilon_0 > 0$, let:*

$$\theta_{f^*}(\mathcal{F}, \gamma_0, \epsilon_0) = \sup_{\mathcal{D}_\mathcal{X}} \sup_{\gamma > \gamma_0, \varepsilon > \varepsilon_0} \left\{ \frac{\gamma^2}{\varepsilon^2} \cdot \mathbb{P}_{\mathcal{D}_\mathcal{X}}\Big(\exists f \in \mathcal{F} : |f(x) - f^*(x)| > \gamma,\ \|f - f^*\|_{\mathcal{D}_\mathcal{X}} \leq \varepsilon\Big)\right\} \vee 1,$$

where $\|f - f^*\|_{\mathcal{D}_\mathcal{X}} := \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[(f(x) - f^*(x))^2]$. We also define $\theta(\mathcal{F}, \gamma) = \sup_{f^* \in \mathcal{F}, \epsilon > 0} \theta_{f^*}(\mathcal{F}, \gamma, \epsilon)$.

2. Passive Learning via Regression

As mentioned in the introduction, past work has found smooth, optimizable surrogate loss functions (e.g. squared loss) to be a powerful tool when learning good binary classifiers. However, proving bounds on classification excess risk of such learned classifiers in terms of the excess risk under the square loss with respect to some class of models $\mathcal{F}$ has been mostly limited to the case when either the Bayes optimal predictor — or a good approximation of it — under the surrogate loss lies in $\mathcal{F}$. An alternative paradigm involves proving bounds on classification error in terms of margin error (Koltchinskii and Panchenko (2004)), but such bounds are typically too pessimistic.

A seminal work that conducted a thorough study of how classification excess risk relates to excess risk under square loss under the Bayes optimal function was conducted in Bartlett et al. (2006). They showed in Theorem 3 in the paper that

$$\mathcal{E}_{0\text{-}1}(f) \leq \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] - \inf_{f'} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f'(x)-y)^2]}.$$

They further improved this bound in Theorem 10 by leveraging a noise-based margin condition. However, their results crucially utilize excess risk under squared loss. To this end, even if one assumes that the class of binary hypothesis induced by a class of regression $\mathcal{F}$ contains the Bayes optimal predictor with respect to 0-1 loss, assuming that the class of regression functions is realizable ($\eta \in \mathcal{F}$) can still be far too strong of an additional assumption. In this section, we consider a particular surrogate loss function, squared loss, and explore the conditions under which $\mathcal{E}_{0\text{-}1}(f)$ can be upper bounded in terms of excess risk under square loss with respect to $\mathcal{F}$. We formally define excess risk under squared loss with respect to $\mathcal{F}$ below

$$\mathcal{E}_{sq}(f,\mathcal{F}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] - \inf_{f'\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f'(x)-y)^2].$$

We also denote $h^* = h_{f^*}$, for $f^* := \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2]$, to be the classifier induced by the function in $\mathcal{F}$ that achieves the minimum squared error on the distribution $\mathcal{D}$.

In order to attain meaningful bounds on classification excess risk, it is useful to make the assumption that the optimal function for the regression problem induces an optimal classifier, thereby reducing classification to finding the optimal regression function. However, identifying f^* directly from data is infeasible due to the unknown distribution $\mathcal{D}$. Instead, we use regression-based function approximation techniques to identify a function $f \in \mathcal{F}$ with low excess risk under squared loss on the distribution. If a relationship between the excess risk under squared loss of a function and its distance to f^* can be established, minimizing the excess risk under squared loss provides an effective way to approximate f^*.

However, this approximation is meaningful for classification only if functions close to f^* induce similar classifiers; when f^* takes values close to $1/2$, even functions arbitrarily close to f^* can induce vastly different classifiers and risk can be large unless $\eta(x)$ is close to $1/2$. This motivates the need for a condition ensuring that f^* maintains a sufficient distance away from $1/2$ with high probability. Concretely, this is addressed by bounding the probability that $f^*(x)$ lies within a margin γ of the decision threshold $1/2$. Specifically, these techniques rely on the assumption that $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}\left[\left|f^*(x)-\frac{1}{2}\right| < \gamma\right]$ is ensured to be very small. Such a condition ensures that any good estimation of f^* induces a classifier that matches h_{f^*} with high probability over the data distribution. It is in this context that margin-based noise conditions, such as Massart's and Tsybakov's, play a critical role.

Definition 4 (Massart Noise Condition, Massart and Nédélec (2006)) *The marginal distribution $\mathcal{D}_{\mathcal{X}}$ satisfies the Massart noise condition with parameter $\gamma \in [0, \frac{1}{2}]$ if $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[|\eta(x) - 1/2| < \gamma] = 0$.*

Definition 5 (Tsybakov Noise Condition, Tsybakov (2004)) *The marginal distribution $\mathcal{D}_{\mathcal{X}}$ satisfies the Tsybakov noise condition with parameter $\beta \geq 0$ and a universal constant $c > 0$ if $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[|\eta(x) - 1/2| < \gamma] \leq c\gamma^{\beta}$ for any $\gamma > 0$.*

To connect these noise conditions to our framework and simultaneously ensure the necessary relationship between a function's excess risk under squared loss and its distance to f^*, previous work rely on the realizability assumption, requiring that the true conditional probability function η lies within the class $\mathcal{F}$. Observe that under realizability here, the optimal regression function induces an optimal classifier and the margin on f^* is equivalent to the margin on η. Moreover, realizability establishes the desired relationship between a regression function's distance to f^* and its excess risk under squared loss,

$$\mathcal{E}_{sq}(f, \mathcal{F}) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[(f(x) - f^*(x))^2].$$

Putting all this together, assuming realizability alongside a margin-based noise condition makes it possible to bound the 0-1 excess risk of a classifier h_f by the excess risk of f under squared loss, as seen in Bartlett et al. (2006).

Although it is empowering, realizability is rarely true in practice, motivating our aim to relax it in this setting. To do so, we directly assume that the optimal regression function induces an optimal classifier and impose a margin on f^* by assuming that the bias of the best-in-class function can be represented as a non-decreasing function of the bias of the conditional probability function. Formally,

Assumption 6 *For conditional probability function $\eta(x)$ and function class $\mathcal{F}$:*

1. *$h_{f^*} = h_\eta$ almost surely, and*
2. *There exists some non-decreasing function $\psi : [0, \frac{1}{2}] \to [0, \frac{1}{2}]$ such that $\left|f^*(x) - \frac{1}{2}\right| \geq \psi\left(\left|\eta(x) - \frac{1}{2}\right|\right)$ almost surely.*

Throughout this paper we assume that the function class $\mathcal{F}$ is a convex set. Convexity of $\mathcal{F}$ immediately implies that for any $f \in \mathcal{F}$:

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

and this relationship is crucial to bound the classification excess risk in terms of the excess risk under squared loss without assuming realizability. Specifically, we prove the following lemma that makes the relationship explicit.

Lemma 7 *For any convex function class $\mathcal{F}$, if Assumption 6 holds, then for any $f \in \mathcal{F}$,*

$$\mathcal{E}_{0\text{-}1}(f) \leq 2 \inf_{\gamma} \left\{ \mathcal{E}_{sq}(f) \cdot \sup_{a>\gamma} \frac{a}{\psi^2(a)} + \gamma\, \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\}\right] \right\}$$

Proof Starting with a standard analysis of excess risk of the ERM, we have:

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

(using Assumption 6.1)

$$= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \neq y\} - \mathbb{1}\{h^*(x) \neq y\}\}\right]$$
$$= \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\{h_f(x) \neq h^*(x)\} \cdot |2\eta(x) - 1|\right]$$

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:

$$\leq 2\,\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\{h_f(x) \neq h^*(x)\} \cdot \mathbb{1}\{|\eta(x) - \tfrac{1}{2}| > \gamma\} \cdot \left|\eta(x) - \tfrac{1}{2}\right|\right]$$
$$+ 2\gamma\,\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\}\right].$$

To bound the first expectation, notice that if $h_f(x) \neq h^*(x)$, then by Assumption 6.2, we know $|f^*(x) - f(x)| \geq |f^*(x) - \frac{1}{2}| \geq \psi(|\eta(x) - \frac{1}{2}|)$ and thus:

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\{h_f(x) \neq h^*(x)\} \cdot \mathbb{1}\{|\eta(x) - \tfrac{1}{2}| > \gamma\} \cdot \left|\eta(x) - \tfrac{1}{2}\right|\right]$$
$$\leq \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\{|f^*(x) - f(x)| \geq \psi\left(\left|\eta(x) - \tfrac{1}{2}\right|\right)\} \cdot \mathbb{1}\{|\eta(x) - \tfrac{1}{2}| > \gamma\} \cdot \left|\eta(x) - \tfrac{1}{2}\right|\right],$$

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

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

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. ■

Example 1 *Consider a convex class $\mathcal{F}$. If Assumption 6 holds for $\psi(x) = x$, and Tsybakov's noise condition for parameter β, then for any $f \in \mathcal{F}$:*

$$\mathcal{E}_{0\text{-}1}(f) \leq \mathcal{O}\left(\mathcal{E}_{sq}(f,\mathcal{F})^{\frac{\beta+1}{\beta+2}}\right). \tag{1}$$

Notice that even in the case where $\beta = 0$, i.e. the data distribution satisfies no nontrivial margin-based noise condition, we still get that $\mathcal{E}_{0\text{-}1}(f) \leq \mathcal{O}(\sqrt{\mathcal{E}_{sq}(f,\mathcal{F})})$. One can also easily adapt this result for Massart's noise condition for the same ψ, giving us a simpler bound of $\mathcal{O}\left(\frac{\mathcal{E}_{sq}(f,\mathcal{F})}{\gamma}\right)$.

Remark 8 *Note that Theorem 10 of Bartlett et al. (2006) instantiated with squared loss essentially provides the bound given in Eq. (1), except that the bounds on both sides are with respect to the Bayes optimal classifier. Using only Assumption 6, we are able to prove the same bound with respect to the function class $\mathcal{F}$ even when $\eta \notin \mathcal{F}$. We also remark Theorem 3 of Bartlett et al. (2006) yields the bound Eq. (1).*

The above result indicates that, by minimizing squared loss excess risk with respect to $\mathcal{F}$, we obtain an estimator that achieves good classification error. In the learning problem, we need to learn such an f from the data. Below, we define the notion of an statistical regression oracle that does this with high probability.

Definition 9 (Offline Regression Oracle) *Given a class of functions $\mathcal{F}$ an offline regression oracle is specified by mapping $\mathsf{Alg}_{\mathsf{Reg}}^{\mathsf{Off}} : \cup_{t=0}^{\infty}(\mathcal{X} \times \mathcal{Y})^t \mapsto \mathcal{F}$ 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 $\hat{f} = \mathsf{Alg}_{\mathsf{Reg}}^{\mathsf{Off}}(S)$ is such that with probability at least $1-\delta$ over draw of samples,*

$$\mathcal{E}_{sq}(\hat{f}, \mathcal{F}) \leq \frac{\mathsf{comp}(\mathcal{F}, \delta, n)}{n}$$

We remark that, for convex $\mathcal{F}$, the ERM algorithm is known to obtain the minimax optimal rate (Bartlett et al. (2005), Lee et al. (1998), Yang and Barron (1999)). In this case, the regression oracle solves a convex optimization problem with respect to the regression function class, and can be efficiently implemented in many cases. Specifically, for the ERM oracle we have that

$$\mathsf{comp}(\mathcal{F}, \delta, n) = \mathcal{O}\left(n \log \tfrac{1}{\delta}\left(\inf_{\kappa>0,\, \nu\in[0,\kappa]}\left(4\nu + \frac{12}{\sqrt{n}}\int_{\nu}^{\kappa}\sqrt{\log \mathcal{N}_2(\mathcal{F}, \beta)}\, d\beta\right) + \frac{\log \mathcal{N}_2(\mathcal{F}, \kappa) + \log\frac{1}{\delta}}{n}\right)\right) \tag{2}$$

The above bound is stated and proved precisely in Lemma 25 for completeness. The following examples illustrate various forms that this complexity takes on.

Example 2 *Consider a convex $\mathcal{F}$ with $Pdim(\mathcal{F}) < \infty$. Lee et al. (1998) show that this complexity admits the form of $\mathsf{comp}(\mathcal{F}, \delta, n) = \mathcal{O}\left(Pdim(\mathcal{F}) \cdot \log(n)\right)$ for the least squares estimator.*

Example 3 *In the setting of nonparametric regression, Liang et al. (2015) demonstrates that a covering at scale δ achieves $\log \mathcal{N}_2(\mathcal{F}, \delta) \leq \delta^{-p}$. In the regime $p \in (0, 2)$, the comp scales as $\mathsf{comp}(\mathcal{F}, \delta, n) = n^{\frac{p}{2+p}}$, while in the regime of $p \geq 2$, it scales as $\mathsf{comp}(\mathcal{F}, \delta, n) = n^{1-1/p}$, with an extra logarithmic factor at $p = 2$.*

Given access to an offline regression oracle, we obtain the following risk bound as a direct consequence of Lemma 7.

Theorem 10 *For any convex function class $\mathcal{F}$, if Assumption 6 holds, for the function $\hat{f}$ returned by the offline regression oracle in Definition 9, we have that with probability at least $1-\delta$,*

$$\mathcal{E}_{0\text{-}1}(\hat{f}) \leq 2 \inf_{\gamma}\left\{\frac{\mathsf{comp}(\mathcal{F}, \delta, n)}{n} \cdot \sup_{a>\gamma} \frac{a}{\psi^2(a)} + \gamma\, \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\}\right]\right\}$$

Proof Combining Lemma 7 with Definition 9 yields the Theorem. ■

Corollary 11 *Consider the setting of Example 1. When $\hat{f}$ is the ERM, return by the offline regression oracle defined in Definition 9, we have that with probability $1-\delta$,*

$$\mathcal{E}_{0\text{-}1}(\hat{f}) \leq \mathcal{O}\left(\frac{\mathsf{comp}(\mathcal{F}, \delta, n)}{n}\right)^{\frac{\beta+1}{\beta+2}},$$

where $\mathsf{comp}(\mathcal{F}, \delta, n)$ is defined in Equation 2.

Remark 12 *Under Tsybakov's noise condition, Theorem 12 in Bartlett et al. (2006) shows a similar bound on the excess risk under* 0-1 *loss displayed below.*

$$\mathcal{E}_{0\text{-}1}(\hat{f}) \leq \mathcal{O}\left(\frac{\mathsf{comp}(\mathcal{F},\delta,n)}{n} + \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2 - (\eta(x)-y)^2]\right)^{\frac{\beta+1}{\beta+2}}$$

Note that the bound in Bartlett et al. (2006) makes use of the local Rademacher complexity, but it can be verified that up to constants, we can replace this with $\mathsf{comp}(\mathcal{F},\delta,n)$*. The bound makes the learner pay an additive misspecification term, rendering the result useful only in the case when this term is small. That is, in non-realizable settings, this term requires that the best-in-class regression function achieves small excess risk when competing against the Bayes optimal classifier. Our result removes this misspecification error for convex function classes* $\mathcal{F}$ *under Assumption 6.*

While it is obviously true that our assumption is weaker than realizability, a reader might wonder if our assumption is qualitatively weaker than approximate versions of realizability. Specifically, we shall say a problem is ϵ-realizable if $|f^*(x) - \eta(x)| \leq \epsilon$ almost surely (see Section 4.2 of Zhu and Nowak (2022) for an example). An avid reader will immediately recognize that ϵ-realizability, along with Massart's noise condition with parameter $\gamma \geq \epsilon$ always implies our assumption with $\psi(x) = x$. More generally, even if one uses Tsybakov's noise condition instead and changes the notion of ϵ-realizability from an $L_\infty(\mathcal{D})$ style (almost sure version) to an $L_2(\mathcal{D})$ style, one can still obtain a version of Assumption 6 which instead of holding almost surely, is violated on a region of input space with probability measure at most a function of ϵ. While we state our condition as holding almost surely, our proofs easily lift to the case where the assumption is violated only on regions with small measure, and we pay this measure additively in our bounds. All of this shows that approximate realizability along with margin condition is subsumed by our results. However, one can also see that there are examples where the problem is far from being approximately realizable (ϵ is constant order), and yet, our assumption and hence our bounds still holds for these settings. We show a simple such construction in the below example.

Example 4 (Assumption 6 holds but not $\epsilon-$realizable for any $\epsilon < 1/4$) *Take* $\mathcal{X} = [0,1]$ *and let* $\mathcal{F} = \{x \mapsto w \cdot x : w \in [0,1]\}$*. Take distribution on* $\mathcal{X}$ *to be uniform on set* $[0, 0.25] \cup [0.75, 1]$ *and let*

$$\eta(x) = \begin{cases} 0 & \text{if } x \in [0, \frac{1}{4}] \\ 1 & \text{if } x \in [\frac{3}{4}, 1] \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

Now one can notice that f^* *the minimizer of expected square-loss within* $\mathcal{F}$ *is indeed given by the weight* $w^* = 1$ *and its induced classifier has same sign as the Bayes optimal classifier. However, note that this classifier is not* ϵ*-realizable for any non-constant* ϵ *either under the almost sure version or even the weaker* $L_2(\mathcal{D})$ *version of* ϵ *realizability. It can be easily checked that, in fact,* $\epsilon < 1/4$ *suffices. Finally, note that the second part of our Assumption 6 is satisfied with* $\psi(x) = 4x$ *in this example.*

The above example along with the proceeding discussion should convince the reader that our assumption is provably weaker than approximate versions of realizability and can be far from it. In fact, our assumption intuitively says that the regression function is just not allowed to be close to margin in regions where the true label is fairly decisive. On the other hand, it does not preclude cases

when the regression function is very confident in places where the true label is close to the margin. This makes sense since the only true danger is when the true label has a clear bias that the regression solution is unable to capture.

3. Active Learning via Regression

We now consider the active learning version of our problem. In this setting, the learner gets access to only the input instances $x_1, \ldots, x_n$ drawn i.i.d from the underlying distribution. The learner can then choose to ask for labels for whichever data points they choose. The goal of the learner is two fold: first, to ensure that excess risk is small and second, to also ask for as few labels as possible.

We specifically consider an online variant of the active learning problem often referred to as selective sampling where the learner sequentially receives input instances, and based on what they have seen so far, they choose whether or not to query for a label. This querying strategy naturally induces *subdistributions* of the original data distribution that are biased towards regions of uncertainty in the input space. In other words, we can view points as being drawn from these subdistributions as they are the ones on which the learner queries and obtains information.

Formally, the learner's querying strategy can be modeled as a function $q : \mathcal{X} \to \{0, 1\}$, that indicates whether or not the learner would query a given point in the input space. We then define the subset of the input space that consists of such points, $Q := \{x \in \mathcal{X} : q(x) = 1\}$. From here, the subdistribution induced by q, $\mathcal{D}_Q$, is simply the distribution $\mathcal{D}$ restricted on to set Q (and renormalized to make it a valid probability measure). That is,

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

Under realizability, the Bayes optimal classifier remains the optimal classifier over any subdistribution. However, without realizability, we can no longer assume that a single $f \in \mathcal{F}$ is the minimizer of expected squared loss under every subdistribution. Since subdistributions are induced by query conditions and therefore defined over regions of uncertainty, it is possible that two vastly different functions are the best model η on different subdistributions, respectively. To account for this, we make the following assumption (analogous to Assumption 6 made in the passive setting) to ensure the performance of our active learning algorithm.

Assumption 13 *There exists non-decreasing function $\psi : [0, \frac{1}{2}] \to [0, \frac{1}{2}]$ such that for every set $Q \subseteq \mathcal{X}$, if $\tilde{f}_Q = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}_Q} \left[(f(x) - y)^2 \right]$:*

1. *$h_{\tilde{f}_Q}(x) = h_\eta(x)$ almost surely, and*
2. *$\left| \tilde{f}_Q(x) - \frac{1}{2} \right| \geq \psi\left(\left| \eta(x) - \frac{1}{2} \right| \right)$ almost surely.*

Assumption 13 simply says that our Assumption 6 introduced in the passive learning setting holds for every subdistribution and not just on $\mathcal{D}$. Such an assumption is intuitively required because, as we restrict the points our algorithm chooses to query, the resulting distribution that the algorithm performs regression on changes.

We now present our epoch-based active learning algorithm in Algorithm 1, which is similar to the one presented in Zhu and Nowak (2022). Our algorithm departs from previous approaches

in the following crucial ways. First, on every epoch, our algorithm queries a given input if and only if all of the query conditions corresponding to previous epochs query the input. In previous approaches, the realizability assumption vacated the need for checking all previous conditions since the optimal function belonged to all previous version spaces. Next, at the end of each epoch m, an offline regression oracle, introduced in Definition 9, is used to obtain the function $\hat{f}_m \in \mathcal{F}$ on a sample of queried points *only* from epoch m. Again, unlike previous works, one cannot use samples from across multiple epochs. Then, our algorithm constructs an implicit class of regression functions $\mathcal{F}_m \subseteq \mathcal{F}$ by including every function in $\mathcal{F}$ whose cumulative empirical distance to $\hat{f}_m$ on queried points from epoch m is at most B. For every $x \in \mathcal{X}$ on which the algorithm is uncertain, the algorithm uses the class of regression functions $\mathcal{F}_m$ to obtain both a new upper confidence bound $\mathrm{ucb}_m(x) = \sup_{f \in \mathcal{F}_m} f(x)$ and lower confidence bound $\mathrm{lcb}_m(x) = \inf_{f \in \mathcal{F}_m} f(x)$ on the probability $\eta(x)$. Notice that if this interval contains $\frac{1}{2}$, then there exists a pair of functions $f, f' \in \mathcal{F}_m$ that induces classifiers $h_f, h_{f'}$ which classify x differently. From this, the algorithm creates the subsequent epoch's query condition q_m from the current query condition q_{m-1} by incorporating the information it learned in epoch m. Finally, and crucially, the algorithm is an improper learning algorithm that, for any input x, looks at the smallest epoch $i \in [M]$ for which there did not exist a pair of functions $f, f' \in \mathcal{F}_i$ whose induced classifiers $h_f, h_{f'}$ classified the x differently. If such an i exists, it outputs the classification of the consensus of the classifiers induced by the regression functions in $\mathcal{F}_i$; otherwise, it outputs 1.

For the algorithm presented, the following theorem that provides a bound on excess risk and query complexity.

Theorem 14 *For any convex function class $\mathcal{F}$, if Assumption 13 holds, then for the predictor $\hat{f}$ returned by Algorithm 1 using offline regression oracle in Definition 9 as subroutine, we have that with probability at least $1 - \delta$,*

$$\mathcal{E}_{0\text{-}1}(\hat{f}) \leq \tilde{\mathcal{O}}\left(\inf_{\gamma > 0}\left\{\frac{1}{n}\left(\sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \cdot \mathsf{comp}(\mathcal{F}, \delta, n)) \vee \log\frac{\log n}{\delta}\right) + \mathbb{P}_{x \sim D_{\mathcal{X}}}\left[\left|\eta(x) - \tfrac{1}{2}\right| \leq \gamma\right]\right\}\right),$$

and simultaneously, the number of label queries is bounded as

$$N_n \leq \tilde{\mathcal{O}}\left(\inf_{\gamma > 0}\left\{\log n\left(\frac{\mathsf{comp}(\mathcal{F}, \delta, n) \cdot \theta(\mathcal{F}, \psi(\gamma))}{\psi^2(\gamma)} \vee \log\frac{\log n}{\delta}\right) + n\,\mathbb{P}_{x \sim D_{\mathcal{X}}}\left[\left|\eta(x) - \tfrac{1}{2}\right| \leq \gamma\right]\right\}\right),$$

where $\tilde{\mathcal{O}}$ simply hides poly-logarithmic factors.

Notice that the excess risk bound in the above theorem matches up to some additive log factor the bound in the passive learning case presented in Theorem 10. In the case when $\psi(x) = x$ (which is still much weaker than realizability), the query complexity matches that of Zhu and Nowak (2022) and is minimax optimal. The proof of Theorem 14 employs a disagreement-based approach to bound the excess risk in terms of comp, which achieves a fast rate for many relevant function classes. Specifically, in Lemma 22, we bound the probability of our query condition firing in epoch m, which is conditional on all the previous query conditions firing, by the events that constitute the disagreement coefficient. By Lemma 20, our convexity assumption allows us to bound estimation error in terms of excess risk over each subdistribution induced by our condition, allowing us to claim that our current version space contains the minimizer over the previous subdistribution. This implies that there exists a function that is further from the minimizer than to $\frac{1}{2}$ itself, and thus has a margin

Algorithm 1 Active Learning in Epochs

Parameters: $\delta \in (0,1)$
Define $\tau_m = 2^m - 1, \tau_0 = 0$, and $q_0(x) = 1$ and $B := C\log^3(n) \cdot \mathsf{comp}(\mathcal{F}, \delta, n)$.
for $m = 1, \ldots, M$ **do**
  for $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
    Receive x_t for $(x_t, y_t) \sim \mathcal{D}$
    if $q_{m-1}(x_t) = 1$ **then**
      Query the label y_t of x_t
    end if
  end for
  Use offline regression oracle on $S_m = \{(x_t, y_t) : q_{m-1}(x_t) = 1, t \in [\tau_{m-1}+1, \tau_m]\}$:

$$\hat{f}_m = \mathsf{Alg}^{\mathsf{Off}}_{\mathsf{Reg}}(S_m, \mathcal{F}) \qquad \text{(From Definition 9)}$$

  Implicitly construct the set of regression functions: $\mathcal{F}_m \subseteq \mathcal{F}$ as:

$$\mathcal{F}_m := \left\{ f \in \mathcal{F} : \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t) = 1\}(f(x_t) - \hat{f}_m(x_t))^2 \leq B \right\}$$

  Construct query function $q_m(x) := \prod_{i=1}^m \mathbb{1}\left\{\frac{1}{2} \in [\mathrm{lcb}_i(x), \mathrm{ucb}_i(x)]\right\}$
end for
return $\hat{f}$ defined by:

$$\hat{f}(x) := \begin{cases} \hat{f}_i(x) & \text{if } q_M(x) = 0 \text{ and } i \text{ is the smallest index s.t. } \frac{1}{2} \notin [\mathrm{lcb}_i(x), \mathrm{ucb}_i(x)] \\ 1 & \text{otherwise} \end{cases}$$

by Assumption 13. Thus, we can reduce the event of querying to the first event in the definition of the disagreement coefficient, which focuses on bounding the difference between functions in the current version space and the subdistribution minimizer under the absolute loss. Lemma 24 provides an upper bound on the expected squared distance between $\tilde{f}_m$ and any $f \in \mathcal{F}_m$, which is the second event of the value function disagreement coefficient.

Proof Idea: To see why the output classifier $\hat{h}$ can be shown to have low excess risk, consider the high probability event of Lemma 20, in which the minimizer of the expected squared loss on $\mathcal{D}_m$, denoted by $\tilde{f}_m$, is in $\mathcal{F}_m$ for all $m \in [M]$. Under this event, for any $x \in \mathcal{X}$, we are guaranteed that $\tilde{f}_m(x)$ is in the confidence interval $[\mathrm{lcb}_m(x), \mathrm{ucb}_m(x)]$ for all $m \in [M]$. Then, the error we incur on any $x \in \mathcal{X}$ will fall into one of two cases,

1. Label of x is not queried: In this case, there must exist an $m \in [M]$ for which $\frac{1}{2} \notin [\mathrm{lcb}_m(x), \mathrm{ucb}_m(x)]$. So, $\hat{h}(x) = h_{\hat{f}_m}(x) = h_{\tilde{f}_m}(x)$. Then, by Assumption 13, we know $h_{\tilde{f}_m}(x) = h_\eta(x) = h_{f^*}(x)$ implying that such x's do not contribute to excess risk .

2. Label of x is queried: In this case, although we accumulate error, this event happens at a rate inversely proportional to the margin on the conditional probability $\eta(x)$ while the expected

loss incurred scales with the margin. This, combined with our margin-based noise condition, ensures a small error.

We show the probability of the algorithm querying a new data point can be bounded by a margin-based complexity measure known as the disagreement coefficient. This complexity measure essentially determines the maximal probability with which two functions in a function class can simultaneously have large point-wise distance while being close in expectation. However, since the optimal function $\tilde{f}_m$ on the distribution $\mathcal{D}_m$ induced by the query condition q_{m-1} simultaneously has a margin on $\mathcal{D}_m$ and, by the convexity assumption, is also a fraction as close to the functions in $\mathcal{F}_m$ as $\hat{f}_m$, if a function disagrees with $\tilde{f}_m$ on any point it must have a large point-wise difference to it. Therefore, we can neatly bound the probability of querying a new data point by this complexity measure.

4. Discussion

Our results show that for both the active and passive learning setting, with only mild assumptions, one can obtain effectively the same guarantees as the ones currently proven under the stringent assumption of realizability. Thus, this work shows that regression can be successful in providing algorithms for classification even if our class of functions is far from modeling the conditional probability distribution of the labels. Below we note some useful points a reader should take note of:

1. For the sake of presentation, we considered the simplest setting of binary labels. However, we note that our core assumptions and the results can be lifted to multiclass classification using a multiregression oracle. In this case, the first part of the assumption would be modified to say that for the minimizer of squared loss (multi-regression version) the class label $y \in [K]$ that has the maximum value coincides with $y \in [K]$ that has the maximum class probability $\mathbb{P}(Y = y|X = x)$. Further, for the second part of the assumption, we can analogously require that the margin for every class label when one uses $f^*(x)$ is lower bounded by a monotonic function of margin for that class under $\mathbb{P}(Y|X = x)$.

2. Regression based algorithms for the contextual bandit problem under realizability are provided in Foster and Rakhlin (2020) for the worst case, and for instance dependent bounds, in Foster et al. (2020). One can ask the question of whether this realizability assumption can be relaxed with milder assumptions like the one we mention above for the multiclass setting. This is an interesting future direction. We do note however that, since, in these reductions, one picks distributions over actions in an epoch based off of the current estimate of $\hat{f}_m(x)$, we do not necessarily get subdistributions as in Assumption 13. One can still change the definition to having subdistributions over $\mathcal{X}$ space but any distribution over actions. However, we are yet to carefully analyze the implications of such an assumption.

3. Finally, we note that our active learning algorithm is improper and technically can work better than the passive learning algorithm in certain scenarios. While we have toy examples that illustrate this, further principled investigation into when this happens can be interesting.

References

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. URL https://www.cs.cmu.edu/~ninamf/papers/a2.pdf.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. URL https://projecteuclid.org/euclid.aos/1121955600.

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. URL https://www.tandfonline.com/doi/abs/10.1198/016214505000000907.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 49–56, 2009. URL https://icml.cc/Conferences/2009/papers/392.pdf.

Alina Beygelzimer, Daniel Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 199–207, Red Hook, NY, USA, 2010. Curran Associates Inc.

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik, 2002.

Sanjoy Dasgupta and Yoav Freund. Active learning using region-based sampling, 2023. URL https://arxiv.org/abs/2303.02721.

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. URL https://arxiv.org/abs/2002.04926.

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. URL https://api.semanticscholar.org/CorpusID:222177499.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014. doi: 10.1561/2200000037. URL https://www.nowpublishers.com/article/Details/MAL-037.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401. doi: https://doi.org/10.1016/0890-5401(92)90010-D. URL https://www.sciencedirect.com/science/article/pii/089054019290010D.

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. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(95)90052-7. URL https://www.sciencedirect.com/science/article/pii/0097316595900527.

Yiran Huang, Jian-Feng Yang, and Haoda Fu. Efficient human-in-the-loop active learning: A novel framework for data labeling in ai systems, 2024. URL https://arxiv.org/abs/2501.00277.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers, 2004. URL https://arxiv.org/abs/math/0405343.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daume III, and John Langford. Active learning for cost-sensitive classification, 2021. URL https://arxiv.org/abs/1703.01014.

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. URL https://ieeexplore.ieee.org/document/705577.

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. URL https://proceedings.mlr.press/v40/Liang15.html.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5), October 2006. ISSN 0090-5364. doi: 10.1214/009053606000000786. URL http://dx.doi.org/10.1214/009053606000000786.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984. URL https://www.stat.yale.edu/~pollard/Books/1984book/pollard1984.pdf.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2), May 2017. ISSN 1350-7265. doi: 10.3150/14-bej679. URL http://dx.doi.org/10.3150/14-BEJ679.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression, 2023. URL https://arxiv.org/abs/2307.04998.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. URL https://projecteuclid.org/euclid.aos/1079120131.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Vladimir N Vapnik and Y Alexey. Chervonenkis. on the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. URL https://projecteuclid.org/euclid.aos/1017939142.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning, 2014. URL https://arxiv.org/abs/1407.2657.

Yinglun Zhu and Robert Nowak. Efficient active learning with abstention, 2022. URL https://arxiv.org/abs/2204.00043.

Appendix A. Proofs for Passive Learning

We provide corollaries for Theorem 10, instantiating the Massart and Tsybakov noise conditions.

Corollary 15 *Consider a convex class $\mathcal{F}$ and $\gamma \in (0, \frac{1}{2}]$. Then, if Assumption 6 holds, Massart's noise condition is true for parameter γ, and we consider some monotone increasing function $\psi : [0, \frac{1}{2}] \to [0, \frac{1}{2}]$, then for any $f \in \mathcal{F}$, we have:*

$$\mathcal{E}_{0\text{-}1}(h_f) \leq 2 \left(\mathcal{E}_{sq}(f) \cdot \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} \right)$$

Proof We begin with the result from Lemma 7,

$$\mathcal{E}_{0\text{-}1}(h_f) \leq 2 \left(\mathcal{E}_{sq}(f) \cdot \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \, \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\} \right] \right)$$

Now, Massart's noise condition implies that

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\} \right] = \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[|\eta(x) - \tfrac{1}{2}| \leq \gamma \right] = 0$$

The result follows from plugging this back into the margin-based noise term. ■

Corollary 16 *Consider a convex class $\mathcal{F}$, $\beta \geq 0$, and universal constant c. Then, if Assumption 6 holds, Tsybakov's noise condition is true for parameters β and γ and constant c, and we consider some monotone increasing function $\psi : [0, \frac{1}{2}] \to [0, \frac{1}{2}]$, then for any $f \in \mathcal{F}$, we have:*

$$\mathcal{E}_{0\text{-}1}(h_f) \leq \inf_{\gamma > 0} 2 \left(\mathcal{E}_{sq}(f) \cdot \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + c\gamma^{\beta+1} \right)$$

Proof Again, we begin with the result from Lemma 7,

$$\mathcal{E}_{0\text{-}1}(h_f) \leq 2 \left(\mathcal{E}_{sq}(f) \cdot \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + \gamma \, \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\} \right] \right)$$

Tsybakov's noise condition implies that

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\mathbb{1}\{|\eta(x) - \tfrac{1}{2}| \leq \gamma\} \right] = \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[|\eta(x) - \tfrac{1}{2}| \leq \gamma \right] \leq c\gamma^{\beta}$$

Plugging this into the margin-based noise term, we get

$$\mathcal{E}_{0\text{-}1}(h_f) \leq 2 \left(\mathcal{E}_{sq}(f) \cdot \sup_{a \in (\gamma, 1]} \frac{a}{\psi^2(a)} + c\gamma^{\beta+1} \right).$$

This inequality holds for all $\gamma > 0$, so our final result follows by optimizing the right side w.r.t γ. ■

Appendix B. Proofs for Active Learning

B.1. Proof of Main Theorem

For the sake of completeness, we will first restate Theorem 14 before providing a proof.

Theorem 17 *With probability at least $1-\delta$, Algorithm 1 returns a classifier $\hat{h}$ for which*

$$\mathcal{E}_{0\text{-}1}(\hat{h}) \leq \mathcal{O}\left(\frac{1}{n}\cdot\left(\sup_{a\in(\gamma,1]}\frac{a}{\psi^2(a)}\cdot(\mathsf{comp}(\mathcal{F},\delta,n)+C(\mathcal{F},\delta,n))\ \vee\ \log\frac{\log n}{\delta}\right)\right)+\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right],$$

and

$$N_n \leq \mathcal{O}\left(\log n\cdot\left(\frac{\mathsf{comp}(\mathcal{F},\delta,n)+C(\mathcal{F},\delta,n)}{\psi^2(\gamma)}\cdot\theta\left(\mathcal{F},\psi(\gamma)\right)\ \vee\ \log\frac{\log n}{\delta}\right)\right)+n\cdot\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right],$$

where $C(\mathcal{F},\delta,n)=\mathcal{O}\left(n\log^3(n)\cdot\mathrm{Rad}_n^2(\mathcal{F})+\log\left(\frac{\log n}{\delta}\right)\right)$, which is $\mathcal{O}\left(\log^3(n)\cdot\mathsf{comp}(\mathcal{F},\delta,n)\right)$.

Proof We start by bounding the excess risk of the classifier outputted by Algorithm 1 and then bound the number of queries it makes. Consider the classifier $\hat{h}=h_{\hat{f}}$ outputted by Algorithm 1. Its excess risk can be bounded by,

$$\begin{aligned}
&\mathcal{E}_{0\text{-}1}(h_{\hat{f}})\\
&=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_{\hat{f}}(x)\neq y\}-\mathbb{1}\{h^*(x)\neq y\}\}\right]\\
&=\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{h_{\hat{f}}(x)\neq h^*(x)\}\cdot(1-2\,\mathbb{P}_y[h^*(x)\neq y|x])\right]\\
&\leq\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{h_{\hat{f}}(x)\neq h^*(x)\}\cdot|2\eta(x)-1|\right],
\end{aligned}$$

where by decomposing on the query condition of Algorithm 1, we get:

$$=\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_M(x)=0,h_{\hat{f}}(x)\neq h^*(x)\}\cdot|2\eta(x)-1|+\mathbb{1}\{q_M(x)=1,h_{\hat{f}}(x)\neq h^*(x)\}\cdot|2\eta(x)-1|\right].$$

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 19 and Lemma 21.

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 $q_M(x)=0$, there exists an $m\in[M]$ such that $\frac{1}{2}\notin[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]$ or in other words there exists an m such that every function in $\mathcal{F}_m$ agrees on the classification of x. Now take i to be the smallest such m. By Lemma 20, we know that $\tilde{f}_i\in\mathcal{F}_i$ and therefore that $h_{\tilde{f}_i}(x)=h_{\hat{f}_i}(x)=h_{\hat{f}}(x)$. Finally, by Assumption 13, we know that $h_{\tilde{f}_i}(x)=h_\eta(x)=h_{f^*}(x)$, implying that the classifier does not incur risk when it would not have queried.

The excess risk incurred when the classifier would have chosen to query can be bounded by the probability it would have queried a data point,

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_M(x)=1,h_{\hat{f}}(x)\neq h^*(x)\}\cdot|2\eta(x)-1|\right]\\
&\leq 2\,\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_M(x)=1\}\cdot|\eta(x)-\tfrac{1}{2}|\right],
\end{aligned}$$

We have a bound on this quantity by Lemma 23, which gives us that,

$$
\begin{aligned}
&2\,\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_M(x)=1\}\cdot\left|\eta(x)-\tfrac{1}{2}\right|\right] \\
&\leq \frac{4}{n_M}\left(\sup_{a\in(\gamma,1]}\frac{a}{\psi^2\left(a\right)}(9\cdot\ C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n))\ \vee\ \log\frac{M}{\delta}\right)+\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right] \\
&=\frac{8}{n}\cdot\left(\sup_{a\in(\gamma,1]}\frac{a}{\psi^2\left(a\right)}(9\cdot\ C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n))\ \vee\ \log\frac{\log n}{\delta}\right)+\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right],
\end{aligned}
$$

where in the last line we plug in $M=\log n$, and by extension $n_M=\frac{n}{2}$, to recover our claimed bound on excess risk.

We now bound the number of queries made by Algorithm 1. We know by Lemma 21 that,

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

Applying Lemma 22,

$$
\begin{aligned}
&\leq\sum_{m=1}^{M}\frac{3}{2}\cdot n_m\cdot\frac{4}{n_{m-1}}\left(\frac{9\cdot C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n)}{\psi^2\left(\gamma\right)}\cdot\theta\left(\mathcal{F},\psi\left(\gamma\right)\right)\ \vee\ \log\frac{M}{\delta}\right) \\
&\quad+n_m\cdot\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right]+\log\frac{M}{\delta}
\end{aligned}
$$

Where, by plugging in the upper bound for $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_M(x)=1]$ we get by Lemma 22 , we get

$$
\begin{aligned}
&=\sum_{m=1}^{M}12\cdot\left(\frac{9\cdot C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n)}{\psi^2\left(\gamma\right)}\cdot\theta\left(\mathcal{F},\psi\left(\gamma\right)\right)\ \vee\ \log\frac{M}{\delta}\right) \\
&\quad+n_m\cdot\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right]+\log\frac{M}{\delta} \\
&=12\log n\cdot\left(\frac{9\cdot C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n)}{\psi^2\left(\gamma\right)}\cdot\theta\left(\mathcal{F},\psi\left(\gamma\right)\right)\ \vee\ \log\frac{\log n}{\delta}\right) \\
&\quad+n\cdot\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right]+\log\frac{\log n}{\delta},
\end{aligned}
$$

where in the last line we plug in $M=\log n$ to recover our claimed bound on label complexity. ■

B.2. Concentration Lemma

Lemma 18 (Bousquet (2002); Rakhlin et al. (2017)) *Let $\mathcal{D}\in\Delta(\mathcal{X}\times\mathcal{Y})$, $\mathcal{F}\subseteq\{f\mid f:\mathcal{X}\to[0,1]\}$, $n\geq 2$, 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 $f, f' \in \mathcal{F}$:

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

and,

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

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

Lemma 19 *Fix any $\delta\in(0,1)$. Then, for all $m\in[M]$ and any $f\in\mathcal{F}$, with probability $1-\delta$, we have:*

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[(f(x)-\tilde{f}_m(x))^2\right]\\ &\le 2\cdot\frac{1}{k_m}\sum_{t=\tau_{m-1}+1}^{\tau_m}\mathbb{1}\{q_{m-1}(x_t)=1\}\cdot(f(x_t)-\tilde{f}_m(x_t))^2 + \frac{C(\mathcal{F},\delta,n)}{k_m},\end{aligned}$$

and

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

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

Proof Take $m\in[M]$. Recall the definition of $\tilde{f}_m$ 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 $\tilde{f}_m$ 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 $t_1^m,\ldots,t_{k_m}^m$ we have:

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

Then, by applying the upper and lower bounds on this quantity from Lemma 18, union bounding over all $m\in[M]$, and re-normalizing δ, we get our desired lemma except with the additive term $C(\mathcal{F},\delta,k_m)$. Finally, we remark that since $C(\mathcal{F},\delta,k_m)$ is an increasing function in its third argument we can replace k_m with n. ■

B.3. Intermediate Lemmas

Lemma 20 *Under the high probability event of Lemma 19, with probability* $1-\delta$*, it is true that for any* $m \in [M]$*,* $\tilde{f}_m \in \mathcal{F}_m$*.*

Proof First recall that we denote t_i^m to be the i-th queried point in epoch m. Then, the empirical distance between $\tilde{f}_m$ and $\hat{f}_m$ on queried points in the m-th epoch can be rewritten as:

$$\sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{q_{m-1}(x_t)=1\}\cdot(\hat{f}_m(x_t)-\tilde{f}_m(x_t))^2$$
$$=\sum_{i=1}^{k_m}(\hat{f}_m(x_{t_i^m})-\tilde{f}_m(x_{t_i^m}))^2,$$

where, by Lemma 19, we can bound this by,

$$\leq 2k_m\,\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}[(\tilde{f}_m(x)-\hat{f}_m(x))^2]+C(\mathcal{F},\delta,n).$$

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

$$\leq 2k_m\,\mathbb{E}_{(x,y)\sim\mathcal{D}_m}[(\hat{f}_m(x)-y)^2-(\tilde{f}_m(x)-y)^2]+C(\mathcal{F},\delta,n).$$

Now, we apply Lemma 25 to get,

$$\leq 2\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,k_m)+C(\mathcal{F},\delta,n)$$
$$\leq 2\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n)+C(\mathcal{F},\delta,n),$$

where the final inequality is possible since comp is an increasing function in its third argument. Plugging this back in gives us our desired result. ■

Lemma 21 *For all* $m \in [M]$*, with probability* $1-2\delta$*,*

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

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 $\epsilon\in(0,1)$,

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

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 $m\in[M], k_m<(1-\epsilon)\cdot\mathbb{E}[k_m]$ 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 $1-2\delta$. ■

The following lemma bounds the probability that our query condition is triggered in each epoch in terms of the disagreement coefficient and serves as the crux for the query complexity analysis.

Lemma 22 *Under the high probability event of Lemma 21, for all $m \in [M]$, we have:*

$$\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1]$$
$$\leq \frac{4}{n_m}\left(\frac{9\cdot C(\mathcal{F},\delta,n)+16\,C_1\cdot \mathsf{comp}(\mathcal{F},\delta,n)}{\psi^2(\gamma)}\cdot\theta\left(\mathcal{F},\psi(\gamma)\right)\vee\log\frac{M}{\delta}\right)+\mathbb{P}_{x\sim D_\mathcal{X}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right].$$

for absolute constant C_1 coming from Lemma 25.

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 $\mathbb{E}[k_m]\leq 4\log M/\delta$, then we have,

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

which implies that,

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

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

$$\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1]$$
$$=\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],q_{m-1}(x)=1\right]$$
$$=\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],q_{m-1}(x)=1,\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right]$$
$$+\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],q_{m-1}(x)=1,\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right],$$

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

$$=\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],q_{m-1}(x)=1,\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right]+\mathbb{P}_{x\sim D_\mathcal{X}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right].$$

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

$$\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],q_{m-1}(x)=1,\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right]$$
$$=\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\,\middle|\,q_{m-1}(x)=1\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]$$
$$=\mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\,\middle|\,x\in\mathcal{X}_m\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]$$
$$=\mathbb{P}_{x\sim D_{\mathcal{X}_m}}\left[\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1].$$

To bound the first term in this product, we start by recalling that from Lemma 20, we know $\tilde{f}_m \in \mathcal{F}_m$ for all $m \in [M]$. Then, for any $x \in \mathcal{X}_m$ for which $\frac{1}{2} \in [\text{lcb}_m(x), \text{ucb}_m(x)]$, there must exist a function $f \in \mathcal{F}_m$ for which $|\tilde{f}_m(x) - f(x)| \geq |\tilde{f}_m(x) - \frac{1}{2}|$. Furthermore, we know by Assumption 13 that $|\tilde{f}_m(x) - \frac{1}{2}| > \psi(|\eta(x) - \frac{1}{2}|)$. However, since f is in $\mathcal{F}_m$, by Lemma 24 we also know an upper bound on $\|f - \tilde{f}_m\|^2_{\mathcal{D}_{\mathcal{X}_m}}$. Therefore, we can bound by the probability these two events happen simultaneously, to get:

$$\begin{aligned}
&\mathbb{P}_{x\sim D_{\mathcal{X}_m}}\left[\frac{1}{2} \in [\text{lcb}_m(x), \text{ucb}_m(x)], \left|\eta(x) - \tfrac{1}{2}\right| > \gamma\right] \\
&\leq \mathbb{P}_{x\sim D_{\mathcal{X}_m}}\left[\left|\eta(x) - \tfrac{1}{2}\right| > \gamma, \exists f \in \mathcal{F}_m : |f(x) - \tilde{f}_m(x)| > \psi\left(\left|\eta(x) - \tfrac{1}{2}\right|\right), \|f - \tilde{f}_m\|^2_{\mathcal{D}_{\mathcal{X}_m}}\right. \\
&\leq \frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m}.
\end{aligned}$$

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

$$\leq \mathbb{P}_{x\sim D_{\mathcal{X}_m}}\left[\exists f \in \mathcal{F}_m : |f(x) - \tilde{f}_m(x)| > \psi\left(\gamma\right), \|f - \tilde{f}_m\|^2_{\mathcal{D}_{\mathcal{X}_m}} \leq \frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m}\right]$$

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

$$\begin{aligned}
&\leq \left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot k_m}\right) \cdot \theta_{\tilde{f}_m}\left(\mathcal{F}_m, \psi\left(\gamma\right), \frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m}\right) \\
&\leq \left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot k_m}\right) \cdot \theta\left(\mathcal{F}, \psi\left(\gamma\right)\right).
\end{aligned}$$

Now, from Lemma 21, we know that, with probability at least $1 - \delta$, the following inequality holds:

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

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

$$\begin{aligned}
&\left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot k_m}\right) \cdot \theta\left(\mathcal{F}, \psi\left(\gamma\right)\right) \\
&\leq 4 \cdot \left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot \mathbb{E}[k_m]}\right) \cdot \theta\left(\mathcal{F}, \psi\left(\gamma\right)\right) \\
&= 4 \cdot \left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot n_m \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1]}\right) \cdot \theta\left(\mathcal{F}, \psi\left(\gamma\right)\right).
\end{aligned}$$

Plugging this upper bound back in and simplifying gives us,

$$\begin{aligned}
&\mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\frac{1}{2} \in [\text{lcb}_m(x), \text{ucb}_m(x)], q_{m-1}(x) = 1, \left|\eta(x) - \tfrac{1}{2}\right| > \gamma\right] + \mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x) - \tfrac{1}{2}\right| \leq \gamma\right] \\
&\leq 4 \cdot \left(\frac{9\cdot\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\psi^2\left(\gamma\right) \cdot n_m}\right) \cdot \theta\left(\mathcal{F}, \psi\left(\gamma\right)\right) + \mathbb{P}_{x\sim D_{\mathcal{X}}}\left[\left|\eta(x) - \tfrac{1}{2}\right| \leq \gamma\right].
\end{aligned}$$

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

The following Lemma is used when bounding the excess risk of the final classifier produced by Algorithm 1. An attentive reader will notice that the analysis follows similar steps to those of Theorem 14 for the passive learning setting.

Lemma 23 *Under the high probability event of Lemma 21, for all $m \in [M]$, we have:*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}\{q_m(x)=1\}\cdot|\eta(x)-\tfrac{1}{2}|]$$
$$\leq \frac{4}{n_m}\left(\sup_{a\in(\gamma,1]}\frac{a}{\psi^2(a)}(9\cdot C(\mathcal{F},\delta,n)+16\,C_1\cdot\mathsf{comp}(\mathcal{F},\delta,n))\vee\log\frac{M}{\delta}\right)+\mathbb{P}_{x\sim D_\mathcal{X}}\left[\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right].$$

for absolute constant C_1 coming from Lemma 25.

Proof We consider two cases and independently prove a bound on our desired quantity for each case. Then, the bound on our desired quantity is simply the greater of the two bounds.

Case 1: If $\mathbb{E}[k_m]\leq 4\log M/\delta$, then we have,

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

which implies that,

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}\{q_m(x)=1\}\cdot|\eta(x)-\tfrac{1}{2}|]\leq\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1]\leq\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\leq\frac{4\log M/\delta}{n_m}.$$

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

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\{q_m(x)=1\}]$$
$$=\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]\right\}\cdot\mathbb{1}\{q_{m-1}(x)=1\}\right],$$

then further decompose it as,

$$=\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]\right\}\cdot\mathbb{1}\{q_{m-1}(x)=1\}\cdot\mathbb{1}\left\{\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right\}\right]$$
$$+\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]\right\}\cdot\mathbb{1}\{q_{m-1}(x)=1\}\cdot\mathbb{1}\left\{\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right\}\right]$$
$$\leq\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]\right\}\cdot\mathbb{1}\{q_{m-1}(x)=1\}\cdot\mathbb{1}\left\{\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right\}\right]$$
$$+\gamma\cdot\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{1}\left\{\left|\eta(x)-\tfrac{1}{2}\right|\leq\gamma\right\}\right],$$

where the second term is just the probability $x\sim\mathcal{D}$ falls inside the margin. Now, to bound the first term, we use the fact that for any function $g(x)$ and event A, $\mathbb{E}[g(x)\mathbb{1}\{A\}]=\mathbb{E}[g(x)|A]\cdot\mathbb{P}[A]$:

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)]\right\}\cdot\mathbb{1}\{q_{m-1}(x)=1\}\cdot\mathbb{1}\left\{\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right\}\right]$$
$$=\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right\}\middle|\,q_{m-1}(x)=1\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]$$
$$=\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[|\eta(x)-\tfrac{1}{2}|\cdot\mathbb{1}\left\{\frac{1}{2}\in[\mathrm{lcb}_m(x),\mathrm{ucb}_m(x)],\left|\eta(x)-\tfrac{1}{2}\right|>\gamma\right\}\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1].$$

Then, because we know by Lemma 20 that $\tilde{f}_m \in \mathcal{F}$, for any $x \in \mathcal{X}$, if $\frac{1}{2} \in [\text{lcb}_m(x), \text{ucb}_m(x)]$ then there exists a function $f \in \mathcal{F}_m$ such that $|\tilde{f}_m(x) - f(x)| \geq |\tilde{f}_m(x) - \frac{1}{2}|$. Furthermore, by Assumption 13, we know that $|\tilde{f}_m(x) - \frac{1}{2}| > \psi(|\eta(x) - \frac{1}{2}|)$, giving us:

$$\leq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[|\eta(x) - \tfrac{1}{2}| \cdot \mathbb{1}\left\{\sup_{f\in\mathcal{F}_m} |\tilde{f}_m(x) - f(x)| > \psi(|\eta(x) - \tfrac{1}{2}|), \left|\eta(x) - \tfrac{1}{2}\right| > \gamma\right\}\right] \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1],$$

where we can bound the indicator by the ratio of the two terms in the first condition multiplied by an indicator for the second condition to get:

$$\leq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[|\eta(x) - \tfrac{1}{2}| \cdot \sup_{f\in\mathcal{F}_m} \frac{(\tilde{f}_m(x) - f(x))^2}{\psi^2(|\eta(x) - \frac{1}{2}|)} \cdot \mathbb{1}\left\{\left|\eta(x) - \tfrac{1}{2}\right| > \gamma\right\}\right] \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1]$$

$$\leq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[|\eta(x) - \tfrac{1}{2}| \cdot \sup_{f\in\mathcal{F}_m} \frac{(\tilde{f}_m(x) - f(x))^2}{\psi^2(|\eta(x) - \frac{1}{2}|)}\right] \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1]$$

$$\leq \sup_{a\in(\gamma,1]} \frac{a}{\psi^2(a)} \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[\sup_{f\in\mathcal{F}_m} (\tilde{f}_m(x) - f(x))^2\right] \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1].$$

Then, from Lemma 24, we have:

$$\leq \sup_{a\in(\gamma,1]} \frac{a}{\psi^2(a)} \cdot \left(\frac{9 \cdot C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m}\right) \cdot \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{m-1}(x) = 1]$$

$$= \sup_{a\in(\gamma,1]} \frac{a}{\psi^2(a)} \cdot \left(\frac{9 \cdot C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m}\right) \cdot \frac{\mathbb{E}[k_m]}{n_m},$$

where by applying the lower bound from Lemma 21 and using our Case 2 assumption, we get:

$$\leq \sup_{a\in(\gamma,1]} \frac{4a}{\psi^2(a)} \cdot \left(\frac{9 \cdot C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{\mathbb{E}[k_m]}\right) \cdot \left(\frac{\mathbb{E}[k_m]}{n_m}\right)$$

$$= \sup_{a\in(\gamma,1]} \frac{4a}{\psi^2(a)} \cdot \left(\frac{9 \cdot C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{n_m}\right),$$

which gives us our desired result. ■

Lemma 24 *Under the high probability event of Lemma 19, for any $m \in [M]$ and $f \in \mathcal{F}_m$, we have,*

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[(f(x) - \tilde{f}_m(x))^2\right] \leq \frac{9\, C(\mathcal{F}, \delta, n) + 16\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n)}{k_m},$$

for absolute constant C_1 coming from Lemma 19 and Lemma 25, respectively.

Proof By Lemma 19, we have,

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

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

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

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

$$\leq 8\, C_1 \cdot \mathsf{comp}(\mathcal{F}, \delta, n) + 4\, C(\mathcal{F}, \delta, n).$$

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

Lemma 25 *[Liang et al. (2015)] For any convex function class $\mathcal{F}$ with probability $1-\delta$, if $\hat{f}$ is the ERM, then*

$$\mathcal{E}_{sq}(\hat{f}) < C_1 \cdot \frac{\mathsf{comp}(\mathcal{F}, \delta, n)}{n},$$

for $\mathsf{comp}(\mathcal{F}, \delta, n) = n \log^3 n \log \frac{1}{\delta} \left(\inf_{\kappa > 0,\, \nu \in [0,\kappa]} \left(4\nu + \frac{12}{\sqrt{n}} \int_{\nu}^{\kappa} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \beta)}\, d\beta \right) + \frac{\log \mathcal{N}_2(\mathcal{F}, \kappa) + \log \frac{1}{\delta}}{n} \right)$ *and some absolute constant $C_1 > 0$.*

Proof This is a direct consequence of applying Lemma 7 to the upper bound of Theorem 4 of Liang et al. (2015), then upper bounding the resulting complexity by multiplying by an additional $\log^3 n$. ■