Journal of Machine Learning Research 20 (2019) 1-38 Submitted 3/17; Revised 1/19; Published 1/19

The Relationship Between Agnostic Selective Classification, Active Learning and the Disagreement Coefficient

Roei Gelbhart ROEIGE@CS.TECHNION.AC.IL
Department of Computer Science
Technion — Israel Institute of Technology

Ran El-Yaniv RANI@CS.TECHNION.AC.IL
Department of Computer Science
Technion — Israel Institute of Technology

Editor: Gabor Lugosi

Abstract

A selective classifier (f, g) comprises a classification function f and a binary selection function g, which determines if the classifier abstains from prediction, or uses f to predict. The classifier is called pointwise-competitive if it classifies each point identically to the best classifier in hindsight (from the same class), whenever it does not abstain. The quality of such a classifier is quantified by its rejection mass, defined to be the probability mass of the points it rejects. A "fast" rejection rate is achieved if the rejection mass is bounded from above by $\tilde{O}(1/m)$ where m is the number of labeled examples used to train the classifier (and $\tilde{O}$ hides logarithmic factors). Pointwise-competitive selective (PCS) classifiers are intimately related to disagreement-based active learning and it is known that in the realizable case, a fast rejection rate of a known PCS algorithm (called Consistent Selective Strategy) is equivalent to an exponential speedup of the well-known CAL active algorithm.

We focus on the agnostic setting, for which there is a known algorithm called LESS that learns a PCS classifier and achieves a fast rejection rate (depending on Hanneke's disagreement coefficient) under strong assumptions. We present an improved PCS learning algorithm called ILESS for which we show a fast rate (depending on Hanneke's disagreement coefficient) without any assumptions. Our rejection bound smoothly interpolates the realizable and agnostic settings. The main result of this paper is an equivalence between the following three entities: (i) the existence of a fast rejection rate for any PCS learning algorithm (such as ILESS); (ii) a poly-logarithmic bound for Hanneke's disagreement coefficient; and (iii) an exponential speedup for a new disagreement-based active learner called Active-ILESS.

Keywords: active learning, selective prediction, disagreement coefficient, selective sampling, selective classification, reject option, pointwise-competitive, selective classification, statistical learning theory, PAC learning, sample complexity, agnostic case

1. Introduction

Selective classification is a unique and extreme instance of the broader concept of confidence-rated prediction (Chow, 1970; Vovk et al., 2005; Bartlett and Wegkamp, 2008; Yuan and Wegkamp, 2010; Cortes et al., 2016a; Wiener and El-Yaniv, 2012; Kocak et al., 2016; Zhang and Chaudhuri, 2014). Given a training sample consisting of m labeled in-

stances, the learning algorithm is required to output a *selective classifier* (El-Yaniv and Wiener, 2010), defined to be a pair (f, g), where f is a prediction function, chosen from some hypothesis class $\mathcal{F}$, and $g : \mathcal{X} \to \{0, 1\}$ is a *selection function*, serving as a qualifier for f as follows: for any x, if $g(x) = 1$, the classifier predicts $f(x)$, and otherwise it abstains. The general performance of a selective classifier is quantified in terms of its *coverage* and *risk*, where coverage is the probabilistic mass of non-rejected instances, and risk is the normalized average loss of f restricted to non-rejected instances. Let f^* be any (unknown) true risk minimizer[1] in $\mathcal{F}$ for the given problem. The selective classifier (f, g) is said to be *pointwise-competitive* if, for each x with $g(x) = 1$, it must hold that $f(x) = f^*(x)$ for all $f^* \in \mathcal{F}$ (Wiener and El-Yaniv, 2015). Thus, pointwise-competitiveness w.h.p. over choices of the training sample, is a highly desirable property: it guarantees, for each non-rejected test point, the best possible classification obtainable using the best in-hindsight classifier from $\mathcal{F}$. We do not restrict g to be from any specific hypothesis class, however, because we use disagreement-based selective prediction, the selection of $\mathcal{F}$ will limit the possibilities of g. The scenario of a predefined decision functions hypothesis class is investigated in Cortes et al. (2016b).

Pointwise-competitive selective classification (PCS) was first considered in the realizable case (El-Yaniv and Wiener, 2010), for which a simple consistent selective strategy (CSS) was shown to achieve a bounded and monotonically increasing (with m) coverage in various non-trivial settings. Note that in the realizable case, any PCS strategy attains zero risk (over the sub-domain it covers). These results were recently extended to the agnostic setting (Wiener and El-Yaniv, 2015; El-Yaniv and Wiener, 2011) with a related but different algorithm called *low-error selective strategy (LESS)*, for which a number of coverage bounds were shown. These bounds relied on the fact that the underlying probability distribution and the hypothesis class $\mathcal{F}$ will satisfy the so-called "(β_1, β_2)-Bernstein property" (Bartlett et al., 2004). The coverage bounds used by Wiener and El-Yaniv (2015); El-Yaniv and Wiener (2011) are dependent on the parameters β_1, β_2. This Bernstein property assumption (as presented in Bartlett et al., 2004), which allows for better concentration, nevertheless, can be problematic. First, it is defined with respect to a unique true risk minimizer f^*, a property that is unlikely to hold in noisy agnostic settings. Moreover, for arbitrary $\mathcal{F}$, even for the 0/1 loss function,there is little knowledge about cases for which the property holds with a non-trivial β_2.[2] We removed the Bernstein assumption from our analysis.

Assuming that a selective classifier is w.h.p. pointwise-competitive, our key goal is a small rejection rate. We will say that a learner has a **fast R^* rejection rate**, if w.h.p. the rejection rate is bounded by

$$\text{polylog}\left(\frac{1}{R(f^*) + 1/m}\right) \cdot R(f^*) + \frac{d \cdot \text{polylog}(m, 1/\delta)}{m},$$

1. We assume that there exists an f^* in $\mathcal{F}$. Otherwise, we can artificially define f^* to be any function whose risk is sufficiently close to $\inf_{f \in \mathcal{F}}(R(f))$, for instance, not greater than a small additive factor from this infimum.
2. It was mentioned by Wiener and El-Yaniv (2015) that, under the Tsybakov noise condition (Tsybakov, 2004), the desired property holds, but this is guaranteed only for cases in which the Bayes classifier is within $\mathcal{F}$, which is a fairly strong assumption in itself.

where $R(f^*)$, d and δ are defined in Section 2. Selective classification is very closely related to the field of **active learning (AL)**. In active learning, the learner can actively influence the learning process by selecting the points to be labeled. The incentive for introducing this extra flexibility is to reduce labeling efforts. A key question in theoretical studies of AL is how many label requests are sufficient to learn a given (unknown) target concept to a specified accuracy, a quantity called *label complexity*. For an AL algorithm satisfying the "passive example complexity" property (consuming the same number of unlabeled (and labeled) examples, as a passive algorithm consumes labeled examples for achieving the same error; see Definition 6.2), we will say it has R^* **exponential speedup**, if w.h.p. the number of labels it requests is bounded by

$$\text{polylog}\left(\frac{1}{R(f^*)+1/m}\right)\cdot R(f^*)m + d\cdot\text{polylog}(m,1/\delta).$$

The connection between active learning and confidence-rated prediction is quite intuitive. A pointwise-competitive selective classifier P can be straightforwardly used as the querying component of an active learning algorithm. This reduction is most naturally demonstrated in the stream-based AL model: at each iteration, the active algorithm trains a selective classifier on the currently available labeled samples, and then decides to query a newly introduced (unlabeled) point x if P abstains on x.

Hanneke's **disagreement coefficient** (Hanneke, 2007), see Definition 2.1, is a well-known parameter of the hypothesis class and the marginal distribution; it is used in most of the known label complexity bounds (Hsu, 2010; Hanneke, 2007; Ailon et al., 2012). The disagreement coefficient is the supremum of the relation between the disagreement mass of functions that are r-distanced from f^* to r, over r. PCS classification is based on using generalization bounds to estimate the empirical error of f^*, and more specifically, its distance from the empirical error of the ERM. Whenever all the functions that reside within a ball around the ERM unanimously agree, the classifier chooses to classify. Thus, the abstain rate is dependent on the disagreement mass of the functions within the ball. The radius of the ball depends on the generalization bounds. The generalization bounds we use are of the form $\tilde{O}(R(f^*)+d/m)$ for the agnostic case. After observing m examples, we can bound the disagreement mass of a ball around the ERM, by multiplying the radius of the ball, which is $\tilde{O}(R(f^*)+d/m)$, with the disagreement coefficient. Thus, if for example, the disagreement coefficient is bounded by a constant, the abstain rate of some PCS algorithms can be bounded by $\tilde{O}(R(f^*)+d/m)$. This gives a basic idea of the disagreement coefficient, which will be formally presented later on.

Note that, in principle, the disagreement coefficient can be replaced by another important quantity, namely, the **version space compression set size**, recently shown to be equivalent to it (Wiener et al., 2015; El-Yaniv and Wiener, 2015). Specifically, an $O(\text{polylog}(m)log(1/\delta))$ version space compression set size minimal bound was shown by Wiener et al. (2015, Corollary 11), to be equivalent to an $O(\text{polylog}(1/r))$ disagreement coefficient.

Zhang and Chaudhuri (2014) present a new algorithm that uses LP in order to achieve better label complexity analysis than was previously known. This paper proposes a quantity, denoted φ_c, to replace the disagreement coefficient (see a streamlined definition of φ_c at the top of page 24 in Hanneke, 2016), which is smaller than the disagreement coefficient. They

show that for the case of linear classification under log-concave distributions, their analysis can improve (reduce) the argument inside the logarithm of the label complexity bound, and the dependency on the VC-dimension is also reduced by a square root. However, in our paper we focus mainly on the more basic question: *when* can we achieve an exponential speedup in terms of $1/\epsilon$. Thus, due to the simplicity of the disagreement coefficient, and its widespread use in the literature, we chose to focus on it. See a detailed discussion on the φ_c quantity in Section 9.

The first contribution of this paper is a novel selective classifier, called ILESS, which uses a tighter generalization error bound than LESS and depends on $R(f^*)$ (and interpolates the agnostic and realizable cases). Most importantly, the new strategy can be analyzed completely without the Bernstein condition.

We derive an active learning algorithm, called Active-ILESS, corresponding to our selective classifier, ILESS. Active-ILESS is constructed to work in a stream-based AL model and its querying function is extremely conservative: for each unlabeled example, the algorithm requests its label if and only if the labeling of the optimal classifier (from the same class) on this point cannot be inferred from information already acquired. This querying strategy, which is often termed "disagreement-based," has been used in a number of stream-based AL algorithms such as Agnostic CAL and Oracular CAL (Hsu, 2010), A^2 (*Agnostic Active*), developed by Balcan et al. (2006), RobustCAL, studied in Hanneke (2012, 2014b) and Hanneke and Yang (2012), or the general agnostic AL algorithm of Dasgupta et al. (2007). Paper Huang et al. (2015) presented a computationally efficient algorithm for disagreement-based AL.

We prove that Active-ILESS, despite being very similar to Oracular CAL Hsu (2010), exhibits an improved label complexity, in comparison to that proved for Oracular CAL. Specifically, Active-ILESS achieves the same label complexity as Agnostic CAL, while being simpler in the sense that its consumption of ERM computations is smaller.

The first formal relationship between PCS classification and AL was proposed by El-Yaniv and Wiener (2012); Wiener (2013), where the aforementioned CSS algorithm was shown to be equivalent to the well-known CAL AL algorithm of Cohn et al. (1994), in the sense that a fast coverage rate for CSS was proven to be equivalent to an exponential label complexity speedup for CAL. This result applies to the realizable setting only. Our first contribution is a similar equivalence relation between pointwise-competitive selective classification and AL, which applies to the more challenging agnostic case and smoothly interpolates the realizable and agnostic settings.

Our second and main contribution is to show a complete equivalence between (i) selective classification with a fast R^* rejection rate, (ii) an AL algorithm, Active-ILESS, with an R^* exponential speedup, and (iii) the existence of an f^* with a disagreement coefficient bounded by polylog$(1/r)$. This is illustrated in Figure 1, where the blue errors indicate the equivalence relationships we prove in this paper, and the red arrow indicates a previously known result (Hsu, 2010; Hanneke, 2007), and can also be deduced from the other arrows.

2. Definitions

Consider a domain $\mathcal{X}$, and a binary label set $\mathcal{Y} = \{\pm 1\}$. A learning problem is specified via a hypothesis class $\mathcal{F}$ and an unknown probability distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. Given a sequence

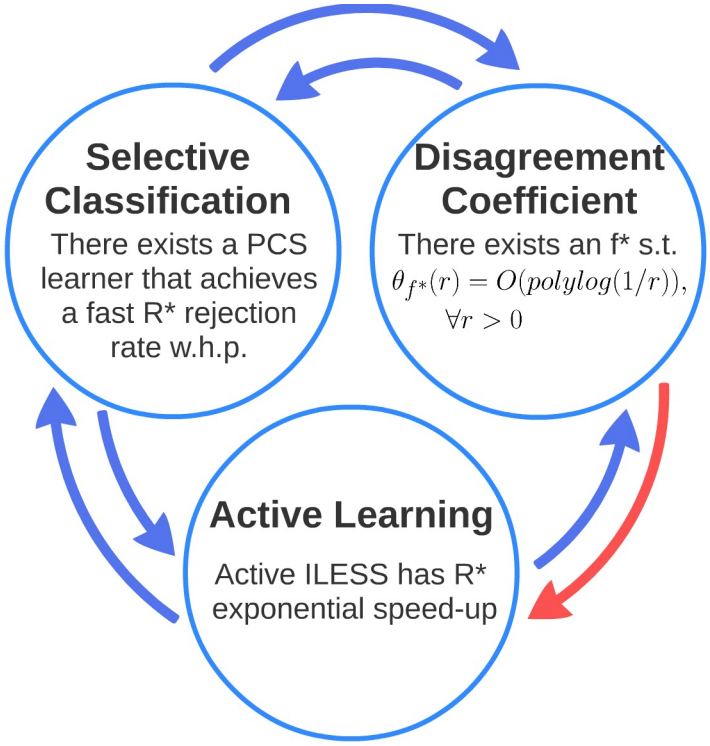

Figure 1: Main results

of labeled training examples $S_m = ((x_1, y_1), (x_2, y_2), ..., (x_m, y_m))$, such that $\forall i, (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, the empirical error of a hypothesis f over S_m is $\hat{R}(f, S_m) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function. In this paper, we focus mainly on the zero-one loss function, $\ell_{01}(y, y') \triangleq \mathbb{1}\{y \neq y'\}$. The true (zero-one) error of f is $R(f) \triangleq \mathbb{E}_{\mathcal{P}}[\ell_{01}(f(x), y)]$. An empirical risk minimizer hypothesis (henceforth, an ERM) is

$$\hat{f}(S_m) \triangleq \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f, S_m), \tag{1}$$

and a true risk minimizer is $f^* \triangleq \operatorname{argmin}_{f \in \mathcal{F}} R(f)$.[3]

We acquire the following definitions from Wiener and El-Yaniv (2015). For any hypothesis class $\mathcal{F}$, hypothesis $f \in \mathcal{F}$, distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, sample S_m, and real number $r > 0$, define the true and empirical *low-error sets*,

$$\mathcal{V}(f, r) \triangleq \left\{f' \in \mathcal{F} : R(f') \leq R(f) + r\right\} \tag{2}$$

and

$$\hat{\mathcal{V}}(f, r) \triangleq \left\{f' \in \mathcal{F} : \hat{R}(f', S_m) \leq \hat{R}(f, S_m) + r\right\}. \tag{3}$$

Let $G \subseteq \mathcal{F}$. The *disagreement set* (Hanneke, 2007) and *agreement set* (El-Yaniv and Wiener, 2010) w.r.t. G are defined, respectively, as

$$DIS(G) \triangleq \{x \in \mathcal{X} : \exists f_1, f_2 \in G \quad , \quad f_1(x) \neq f_2(x)\} \tag{4}$$

3. We assume that f^* exists, and that it need not be unique, in which case f^* refers to any one of the minimizers.

$$\text{and} \quad AGR(G) \triangleq \{x \in \mathcal{X} : \forall f_1, f_2 \in G \quad , \quad f_1(x) = f_2(x)\}. \tag{5}$$

In *selective classification* (El-Yaniv and Wiener, 2010), the learning algorithm receives S_m and is required to output a *selective classifier*, defined to be a pair (f, g), where $f \in \mathcal{F}$ is a classifier, and $g : \mathcal{X} \to \{0, 1\}$ is a *selection function*, serving as a qualifier for f as follows. For any $x \in \mathcal{X}$, $(f, g)(x) = f(x)$ iff $g(x) = 1$. Otherwise, the classifier outputs "I don't know". For any selective classifier (f, g), we define its coverage to be

$$\Phi(f, g) \triangleq \Pr_{X \sim \mathcal{P}_\mathcal{X}}(g(X) = 1),$$

and its complement, $1 - \Phi$, is called the **abstain rate**. For any $f \in \mathcal{F}$ and $r > 0$, define the set $B(f, r)$ of all hypotheses that reside within a ball of radius r around f,

$$B(f, r) \triangleq \left\{ f' \in \mathcal{F} : \Pr_{X \sim \mathcal{P}_\mathcal{X}} \left\{ f'(X) \neq f(X) \right\} \leq r \right\}.$$

For any $G \subseteq \mathcal{F}$, and distribution $\mathcal{P}_\mathcal{X}$, we denote by ΔG the volume of the disagreement set of G (see (4)), $\Delta G \triangleq \Pr\{DIS(G)\}$.

Definition 2.1 (Disagreement Coefficient) *Let $r_0 \geq 0$. Hanneke's* disagreement coefficient *(Hanneke, 2007) of a classifier $f \in \mathcal{F}$ with respect to the target distribution $\mathcal{P}_\mathcal{X}$ is*

$$\theta_f(r_0) \triangleq \sup_{r > r_0} \frac{\Delta B(f, r)}{r}, \tag{6}$$

and the general disagreement coefficient *of the entire hypothesis class $\mathcal{F}$ is*

$$\theta(r_0) \triangleq \sup_{f \in \mathcal{F}} \theta_f(r_0). \tag{7}$$

Notice that this definition of the disagreement coefficient is independent of $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Another commonly used definition of the disagreement coefficient does depend on a true risk minimizer f^*, as follows:

$$\theta'(r_0) = \sup_{r > r_0} \frac{\Delta B(f^*, r)}{r}. \tag{8}$$

Clearly, it always holds that $\theta' \leq \theta$. The independence of θ of unknown quantities such as the underlying distribution (and f^*), however, is a convenient property that sometimes allows for a direct estimation of θ, which only depends on the marginal distribution, $\mathcal{P}_\mathcal{X}$. This is, for example, the case in AL, where labels are expensive but information about the marginal distribution (provided by unlabeled examples) is cheap. Note also that the above definition of θ' implicitly assumes a unique f^*. Nevertheless, the definition can be extended to cases where f^* is not unique, in which case the infimum over all f^* can be considered (the analysis can be extended accordingly using limits). For more on the disagreement coefficient, and examples of probabilities distributions and hypothesis classes for which it is bounded, see Hanneke (2014b).

3. Convergence Bounds and LESS

We use a uniform convergence bound from Vapnik and Chervonenkis (1974); Dasgupta et al. (2007); Bousquet et al. (2003). Define convergence slacks $\sigma_{R-\hat{R}}(m,\delta,d,R,\hat{R})$ and $\sigma_{\hat{R}-R}(m,\delta,d,R,\hat{R})$, given in terms of the training sample, S_m, its size, m, the confidence parameter, δ, and the VC-dimension d of the class $\mathcal{F}$. For any $f \in \mathcal{F}$,

$$\sigma_{R-\hat{R}}(m,\delta,d,R,\hat{R}) \triangleq \min\left\{\underbrace{\frac{4d\ln(\frac{16me}{d\delta})}{m}+\sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot\hat{R}}}_{\hat{\sigma}_{R-\hat{R}}(m,\delta,d,\hat{R})},\underbrace{\sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot R}}_{\bar{\sigma}_{R-\hat{R}}(m,\delta,d,R)}\right\} \tag{9}$$

and

$$\sigma_{\hat{R}-R}(m,\delta,d,R,\hat{R}) \triangleq \min\left\{\underbrace{\frac{4d\ln(\frac{16me}{d\delta})}{m}+\sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot R}}_{\bar{\sigma}_{\hat{R}-R}(m,\delta,d,R)},\underbrace{\sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot\hat{R}}}_{\hat{\sigma}_{\hat{R}-R}(m,\delta,d,\hat{R})}\right\}. \tag{10}$$

To simplify the analysis, we further decompose the above slack terms into their empirical and non-empirical components. For (9), we thus have, respectively,

$$\hat{\sigma}_{R-\hat{R}}(m,\delta,d,\hat{R}) \triangleq \frac{4d\ln(\frac{16me}{d\delta})}{m}+\sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot\hat{R}} \tag{11}$$

and

$$\hat{\sigma}_{\hat{R}-R}(m,\delta,d,\hat{R}) \triangleq \sqrt{\frac{4d\ln(\frac{16me}{d\delta})}{m}\cdot\hat{R}}. \tag{12}$$

Similarly, the non-empirical part in these minimums are denoted by $\bar{\sigma}_{R-\hat{R}}$ and $\bar{\sigma}_{\hat{R}-R}$. With this notation, we can write, for example, $\sigma_{R-\hat{R}} = \min\{\hat{\sigma}_{R-\hat{R}}, \bar{\sigma}_{R-\hat{R}}\}$. Our Lemma 1 is taken from the work of Dasgupta et al. (2007, Lemma 1), which is based on Bousquet et al. (2003, Theorem 7) [4].

Lemma 1 *Let $\mathcal{F}$ be a hypothesis class with VC-dimension d. For any $0 < \delta < 1$, with probability of at least $1-\delta$ over the choice of S_m from $\mathcal{P}^m$, any hypothesis $f \in \mathcal{F}$ satisfies*

$$R(f) \leq \hat{R}(f) + \sigma_{R-\hat{R}}\left(m,\delta,d,R(f),\hat{R}(f)\right) \tag{13}$$

$$\hat{R}(f) \leq R(f) + \sigma_{\hat{R}-R}\left(m,\delta,d,R(f),\hat{R}(f)\right). \tag{14}$$

4. In the original lemma from Dasgupta et al. (2007), $S(\mathcal{H},n)$, the growth function, is given. We insert Sauer's Lemma, $S(\mathcal{H},n) \leq (\frac{em}{d})^d$, into Lemma 1 from Dasgupta et al. (2007) to get our lemma.

Strategy 1 is the LESS algorithm of Wiener and El-Yaniv (2015). LESS learns w.h.p. a pointwise-competitive selective classifier, (f, g), where $f \in \mathcal{F}$ and $g : \mathcal{X} \to \{0, 1\}$ is its selection function that determines whether to abstain or to classify. A *pointwise-competitive selective classifier* must satisfy the following condition: For each x with $g(x) = 1$, it must hold that $f(x) = f^*(x)$ for all $f^* \in \mathcal{F}$. A PCS learning algorithm must output a PCS classifier w.h.p. **for all** $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$; otherwise, one can consider a tailor-made trivial algorithm for each distribution, which simply returns f^*.

Remark 2 *The original definition of pointwise-competitiveness from Wiener and El-Yaniv (2015) requires a single f^*. We widen the definition to cases for which there are more than one f^*, and require that a pointwise-competitive selective classifier be equal to* all *f^*, wherever $g = 1$. This extrapolation seems a bit strict. Nevertheless, even if the requirement would have been relaxed to "any f^*", any pointwise-competitive selective classifier would still have been forced to identify with all f^*, as it is impossible to determine whether a set of functions are all f^*, or one is better than the rest.*

The main idea behind LESS is that, w.h.p., all f^* lie within a ball around an ERM hypothesis with an error radius of $2\sigma(m, \delta/4, d)$, where

$$\sigma(m, \delta, d) \triangleq 2\sqrt{\frac{2d\left(\ln\frac{2me}{d}\right) + \ln\frac{2}{\delta}}{m}} \tag{15}$$

is the slack term of a certain uniform convergence bound. Therefore, if all the functions in that ball agree over the labeling of any instance x, we know with high probability that all f^* label x the same way as the ERM. This property ensures that LESS is pointwise-competitive w.h.p.

Strategy 1 Agnostic Low-Error Selective Strategy (LESS)

Input: Sample set of size m, S_m,
Confidence level δ
Hypothesis class $\mathcal{F}$ with VC dimension d
Output: A selective classifier (h, g)
1: Set $\hat{f} = ERM(\mathcal{F}, S_m)$, i.e., $\hat{f}$ is any empirical risk minimizer from $\mathcal{F}$
2: Set $G = \hat{\mathcal{V}}\left(\hat{f}, 2\sigma(m, \delta/4, d)\right)$
3: Construct g such that $g(x) = 1 \iff x \in \{\mathcal{X} \setminus DIS(G)\}$
4: $f = \hat{f}$

4. ILESS

We now introduce an improved version of LESS, called ILESS, which uses a radius of the form $d \cdot \text{polylog}(m, 1/\delta) \cdot \left(\frac{1}{m} + \sqrt{\frac{R(f^*)}{m}}\right)$. Noting that the radius, $2\sigma(m, \delta/4, d)$, used by LESS to define $G = \hat{\mathcal{V}}$, is of the form $d \cdot \text{polylog}(m, 1/\delta)/\sqrt{m}$, we observe that in cases where $R(f^*) \approx \frac{C}{m}$, this new radius behaves as $\frac{d \cdot \text{polylog}(m, 1/\delta)}{m}$. We later show that this radius allows ILESS to achieve a faster rejection decay rate than the one achieved by LESS.

Consider the pseudo-code of ILESS given in Strategy 2. We now analyze ILESS, and begin by showing in Lemma 3 that ILESS is pointwise-competitive w.h.p., i.e., for any x

Strategy 2 Improved Low-Error Selective Strategy (ILESS)

Input: Sample set of size m, S_m,
Confidence level δ
Hypothesis class $\mathcal{F}$ with VC dimension d
Output: A selective classifier (h, g)
1: Set $\hat{f} = ERM(\mathcal{F}, S_m)$, i.e., $\hat{f}$ is any empirical risk minimizer from $\mathcal{F}$
2: Set $\sigma_{\text{ILESS}} = \hat{\sigma}_{R-\hat{R}}\left(m, \delta, d, \hat{R}(\hat{f}, S_m)\right) + \bar{\sigma}_{\hat{R}-R}\left(m, \delta, d, \hat{R}(\hat{f}, S_m) + \hat{\sigma}_{R-\hat{R}}(m, \delta, d, \hat{R}(\hat{f}, S_m))\right)$
3: Set $G = \hat{\mathcal{V}}\left(\hat{f}, \sigma_{\text{ILESS}}\right)$
4: Construct g such that $g(x) = 1 \iff x \in \{\mathcal{X} \setminus DIS(G)\}$
5: $h = \hat{f}$

for which $g(x) = 1$, $f(x) = f^*(x)$ for all f^*. The calculation of g appears to be very problematic, as for a specific x, a unanimous decision over an infinite number of functions must be ensured. This problem was shown to be reducible to finding an ERM under one constraint (Lemma 6.1 in El-Yaniv and Wiener, 2011, a.k.a. the disbelief principle). This is a difficult problem, nonetheless, albeit one that could be estimated with heuristics.

Definition 4.1 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown probability distribution. Given a sample set S_m, drawn from $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, we denote by $\mathcal{E}$ the event where both inequalities (13) and (14) of Lemma 1 simultaneously hold. We know from the lemma that $\mathcal{E}$ occurs with probability of at least $1 - \delta$.*

Lemma 3 (ILESS **is pointwise-competitive)** *Given that event $\mathcal{E}$ occurred (see Definition 4.1), for all $f^* \in \mathcal{F}$, f^* resides within G (from Strategy 2), and therefore,* ILESS *is pointwise-competitive w.h.p.*

Proof From (14), it follows that

$$\begin{aligned} \hat{R}(f^*, S_m) &\leq R(f^*) + \sigma_{\hat{R}-R}(m, \delta, d, R(f^*), \hat{R}(f^*, S_m)) \\ &\leq R(f^*) + \bar{\sigma}_{\hat{R}-R}(m, \delta, d, R(f^*)). \end{aligned} \tag{16}$$

Additionally, by the definition of f^*, we know that it has the lowest true error, and using Inequality (13) from Lemma 1 we obtain,

$$\begin{aligned} R(f^*) &\leq R(\hat{f}) \\ &\leq \hat{R}(\hat{f}, S_m) + \sigma_{R-\hat{R}}(m, \delta, d, R(\hat{f}), \hat{R}(\hat{f}, S_m)) \\ &\leq \hat{R}(\hat{f}, S_m) + \hat{\sigma}_{R-\hat{R}}(m, \delta, d, \hat{R}(\hat{f}, S_m)). \end{aligned} \tag{17}$$

Finally, by applying (17) in (16), we have,

$$\begin{aligned} \hat{R}(f^*, S_m) \leq \hat{R}(\hat{f}, S_m) &+ \hat{\sigma}_{R-\hat{R}}(m, \delta, d, \hat{R}(\hat{f}, S_m)) \\ &+ \bar{\sigma}_{\hat{R}-R}\left(m, \delta, d, \hat{R}(\hat{f}, S_m) + \hat{\sigma}_{R-\hat{R}}(m, \delta, d, \hat{R}(\hat{f}, S_m))\right), \end{aligned}$$

which means that $f^* \in G$. ■

Lemma 4 below bounds the radius σ_{ILESS} of ILESS. The lemma uses the notation

$$A \triangleq 4d\ln(\frac{16me}{d\delta}),$$

with which, by the definition of σ_{ILESS} (see Strategy 2), we have,

$$\begin{aligned}\sigma_{\text{ILESS}} &= \hat{\sigma}_{R-\hat{R}}(m,\delta,d,\hat{R}(\hat{f},S_m)) + \bar{\sigma}_{\hat{R}-R}\left(m,\delta,d,\hat{R}(\hat{f},S_m) + \hat{\sigma}_{R-\hat{R}}(m,\delta,d,\hat{R}(\hat{f},S_m))\right) \\ &= \frac{A}{m} + \sqrt{\frac{A}{m}\cdot\hat{R}(\hat{f},S_m)} + \frac{A}{m} + \sqrt{\frac{A}{m}\cdot\left[\hat{R}(\hat{f},S_m) + \frac{A}{m} + \sqrt{\frac{A}{m}\cdot\hat{R}(\hat{f},S_m)}\right]}.\end{aligned} \tag{18}$$

Lemma 4 *Given that event $\mathcal{E}$ (see Definition 4.1) occurred, the radius of* ILESS *satisfies*

$$\sigma_{\text{ILESS}} \leq 6\frac{A}{m} + 3\sqrt{\frac{A}{m}\cdot R(f^*)} = O\left(\frac{A}{m} + \sqrt{\frac{A}{m}\cdot R(f^*)}\right), \tag{19}$$

where $A \triangleq 4d\ln(\frac{16me}{d\delta})$.

Proof Under our assumption, inequalities (13) and (14) hold for every $f \in \mathcal{F}$. We thus have

$$\hat{R}(\hat{f},S_m) \leq \hat{R}(f^*,S_m) \leq R(f^*) + \frac{A}{m} + \sqrt{\frac{A}{m}\cdot R(f^*)}. \tag{20}$$

Replacing the three occurrences of $\hat{R}(f^*, S_m)$ in (18) with the R.H.S. of (20), and using the basic inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{ab} \leq a/2 + b/2$, we get,

$$\begin{aligned}
\sigma_{\text{ILESS}} &\leq \frac{A}{m} + \sqrt{\frac{A}{m} \cdot \left(R(f^*) + \frac{A}{m} + \sqrt{\frac{A}{m} \cdot R(f^*)} \right)} + \frac{A}{m} + \\
&\quad + \sqrt{\frac{A}{m} \cdot \left[R(f^*) + \frac{2A}{m} + \sqrt{\frac{A}{m} \cdot R(f^*)} + \sqrt{\frac{A}{m} \cdot \left(R(f^*) + \frac{A}{m} + \sqrt{\frac{A}{m} \cdot R(f^*)} \right)} \right]} \\
&\leq \frac{2A}{m} + \sqrt{\frac{A}{m} \cdot \left(R(f^*) + \frac{A}{m} + \frac{A}{2m} + \frac{1}{2} R(f^*) \right)} + \\
&\quad + \sqrt{\frac{A}{m} \cdot \left[R(f^*) + \frac{3A}{2m} + \frac{1}{2} R(f^*) + \frac{A}{m} + \sqrt{\frac{A}{m} \cdot \left(R(f^*) + \frac{3A}{2m} + \frac{1}{2} R(f^*) \right)} \right]} \\
&\leq \frac{2A}{m} + \frac{3A}{2m} + \frac{3}{2}\sqrt{\frac{A}{m} R(f^*)} + \sqrt{\frac{A}{m} \cdot \left[\frac{5A}{2m} + \frac{3}{2} R(f^*) + \sqrt{\frac{A}{m} \cdot \left(\frac{3A}{2m} + \frac{3}{2} R(f^*) \right)} \right]} \\
&\leq \frac{7A}{2m} + \frac{3}{2}\sqrt{\frac{A}{m} \cdot R(f^*)} + \sqrt{\frac{A}{m} \cdot \left[\frac{5A}{2m} + \frac{3}{2} R(f^*) + \frac{3A}{2m} + \sqrt{\frac{3A}{2m} \cdot R(f^*)} \right]} \\
&\leq \frac{7A}{2m} + \frac{3}{2}\sqrt{\frac{A}{m} \cdot R(f^*)} + \sqrt{\frac{A}{m} \cdot \left[\frac{5A}{2m} + \frac{3}{2} R(f^*) + \frac{3A}{2m} + \frac{3A}{4m} + \frac{3}{4} R(f^*) \right]} \\
&\leq \frac{7A}{2m} + \frac{3}{2}\sqrt{\frac{A}{m} \cdot R(f^*)} + \sqrt{\frac{19}{4}} \frac{A}{m} + \sqrt{\frac{A}{m} \cdot \frac{9}{4} R(f^*)} \\
&\leq 6\frac{A}{m} + 3\sqrt{\frac{A}{m} \cdot R(f^*)}. \qquad (21)
\end{aligned}$$

■

In comparison, the radius of LESS is of order $O(\sqrt{\frac{A}{m}})$, which can be significantly larger when $R(f^*)$ is small. This potential radius advantage translates into a potential coverage advantage of ILESS, as stated in the following theorem.

Theorem 5 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown probability distribution. Given that event $\mathcal{E}$ (see Definition 4.1) occurred, G defined in Strategy 2, holds that*

$$G \subseteq B(f^*, R_0),$$

where

$$R_0 \triangleq 2 \cdot R(f^*) + 11 \cdot \frac{A}{m} + 6 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)},$$

and thus, for all f^*, the abstain rate is bounded by

$$1 - \Phi(\text{ILESS}) \leq \theta_{f^*}(R_0) \cdot R_0.$$

This immediately implies (by definition) that

$$1 - \Phi(\text{ILESS}) \leq \theta(R_0) \cdot R_0.$$

Remark 6 *Note that* $R_0 = O\left(R(f^*) + \frac{A}{m}\right)$ *due to* $\sqrt{\frac{A}{m} \cdot R(f^*)} \leq \frac{1}{2}(\frac{A}{m} + R(f^*))$.

Proof We start by showing that G, defined in Strategy 2, resides within a ball around any specific f^*. To do so, we need to bound the true error of all functions in G.

$$\begin{aligned}
f \in G \quad &\Rightarrow \quad \hat{R}(f, S_m) \leq \hat{R}(\hat{f}, S_m) + \sigma_{\text{ILESS}} && (22)\\
&\Rightarrow \quad \hat{R}(f, S_m) \leq R(f^*) + \frac{A}{m} + \sqrt{\frac{A}{m} \cdot R(f^*)} + 6\frac{A}{m} + 3\sqrt{\frac{A}{m} \cdot R(f^*)} && (23)\\
&\Rightarrow \quad \hat{R}(f, S_m) \leq R(f^*) + 7 \cdot \frac{A}{m} + 4 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)}, && (24)
\end{aligned}$$

where Inequality (22) is explained by the definition of G, and inequality (23) follows from (20) and (21) (under event $\mathcal{E}$). We then have,

$$\begin{aligned}
R(f) \quad &\leq \quad \hat{R}(f, S_m) + \hat{\sigma}_{R-\hat{R}}(m, \delta, d, \hat{R}) && (25)\\
&\leq \quad \hat{R}(f, S_m) + \frac{A}{m} + \sqrt{\frac{A}{m} \cdot \hat{R}(f, S_m)} && (26)\\
&\leq \quad R(f^*) + 8 \cdot \frac{A}{m} + 4 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)} + \\
&\qquad + \sqrt{\frac{A}{m} \cdot \left[R(f^*) + 7 \cdot \frac{A}{m} + 4 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)}\right]} && (27)\\
&\leq \quad R(f^*) + 8 \cdot \frac{A}{m} + 4 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)} + \sqrt{\frac{A}{m} \cdot \left[3R(f^*) + 9 \cdot \frac{A}{m}\right]} && (28)\\
&\leq \quad R(f^*) + 11 \cdot \frac{A}{m} + 6 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)}, && (29)
\end{aligned}$$

where inequality (25) is (13) (which holds given $\mathcal{E}$), inequality (26) follows directly from the definition of $\hat{\sigma}_{R-\hat{R}}$, inequality (27) is obtained using (24), Inequality (28) follows from $\sqrt{ab} \leq a/2 + b/2$, and (29) from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

Using (29), for all $f \in G$, and any f^*, we have,

$$\begin{aligned}
\Pr_{X \sim \mathcal{P}_\mathcal{X}}\{f(X) \neq f^*(X)\} \quad &= \quad \Pr_{X,Y \sim \mathcal{P}_{\mathcal{X},\mathcal{Y}}}\{f(X) \neq f^*(X) \wedge f^*(X) = Y\} + \\
&\qquad \Pr_{X,Y \sim \mathcal{P}_{\mathcal{X},\mathcal{Y}}}\{f(X) \neq f^*(X) \wedge f^*(X) \neq Y\}\\
&\leq \quad \Pr_{X,Y \sim \mathcal{P}_{\mathcal{X},\mathcal{Y}}}\{f(X) \neq f^*(X) \wedge f^*(X) = Y\} + R(f^*)\\
&\leq \quad \Pr_{X,Y \sim \mathcal{P}_{\mathcal{X},\mathcal{Y}}}\{f(X) \neq Y\} + R(f^*)\\
&= \quad R(f) + R(f^*)\\
&\leq \quad 2 \cdot R(f^*) + 11 \cdot \frac{A}{m} + 6 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)}. && (30)
\end{aligned}$$

It follows that

$$f \in B\left(f^*, 2 \cdot R(f^*) + 11 \cdot \frac{A}{m} + 6 \cdot \sqrt{\frac{A}{m} \cdot R(f^*)}\right) = B(f^*, R_0),$$

and, in particular,

$$G \subseteq B(f^*, R_0),$$

so,

$$\Delta G \leq \Delta B(f^*, R_0).$$

The abstain rate of ILESS equals ΔG. We can now use the disagreement coefficient to bound the abstain rate from above,

$$\Delta G \leq \Delta B(f^*, R_0) = \frac{\Delta B(f^*, R_0)}{R_0} \cdot R_0 \leq \theta(R_0) \cdot R_0, \tag{31}$$

which concludes the proof. ■

According to Theorem 5, assuming the disagreement coefficient is $\theta(r) = O(\text{polylog}(1/r))$ for $r \geq R(f^*)$, the rejection mass of ILESS, defined as the probability that the classifier trained by ILESS will output "I don't know" is bounded w.h.p. by

$$\text{polylog}_1\left(\frac{1}{R(f^*) + 1/m}\right) \cdot R(f^*) + \frac{d \cdot \text{polylog}_2(m, 1/\delta)}{m}. \tag{32}$$

In many cases, the disagreement coefficient, $\theta(r)$, is bounded by a constant, or by $O(\text{polylog}(1/r))$ for all $r > 0$ (see Hanneke, 2014b). For example, it was shown in Wiener et al. (2015), that for linear separators under mixture of Gaussians, and for axis-aligned rectangles with probability mass bounded away from zero under product densities over $\mathbb{R}^k$, $\theta(r)$ is bounded by $O(\text{polylog}(1/r))$ for all $r > 0$. For such cases, we know that (32) always holds, regardless of the size of $R(f^*)$. The disagreement coefficient is only dependent on the marginal $\mathcal{P}_\mathcal{X}$, the hypothesis class $\mathcal{F}$, and the identity of the true risk minimizers, f^* (which is not necessarily unique). This fact motivates the following definition of a rejection rate of a selective learning algorithm, which is only dependent on $\mathcal{P}_\mathcal{X}$,$\mathcal{F}$ and f^*.

Definition 4.2 (Fast R^* Rejection Rate) *Given $\mathcal{P}_\mathcal{X}$,$\mathcal{F}$ and f^*, if for all $\delta > 0$, and all $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ for which f^* is a true risk minimizer, the rejection mass of a selective classifier learning algorithm is bounded by (32) with probability of at least $1 - \delta$, we say that the algorithm achieves a **fast** R^* **rejection rate**, with* polylog_1 *and* polylog_2 *as its parameters.*

Corollary 7 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d. Given $\mathcal{P}_\mathcal{X}$ and f^*, if $\theta_{f^*}(r)$ is bounded by $O(\text{polylog}(1/r))$ for all $r > 0$, then there exists a* PCS *learning algorithm (*ILESS*) which achieves a fast R^* rejection rate.*

Proof The proof is immediate from Theorem 5. ■

In the next section, we will show the other direction; that is, if there is a PCS learning algorithm that has a fast R^* rejection rate, then $\theta(r) = O(\text{polylog}(1/r))$ for all $r > 0$.

As long as the number of training examples that ILESS receives is not "too large" relative to $1/R(f^*)$, i.e., $m \ll \frac{1}{R(f^*)}$, the rejection mass of ILESS is

$$O\left(\frac{d \cdot \text{polylog}(m, 1/\delta)}{m}\right).$$

When m is large, and $R(f^*)$ becomes more dominant than $\frac{1}{m}$, our coverage bound is dominated by $R(f^*)$. This should not surprise us, as ILESS achieves *pointwise-competitiveness* w.h.p., and any strategy that achieves pointwise-competitiveness cannot ensure a better rejection mass than $R(f^*)$ without making more assumptions about the error or the distribution. This can be seen in the following example, in which $\theta(r) \leq 1$ for all $r > 0$, but the rejection mass of any pointwise-competitive strategy is always at least $R(f^*)$.

Example 1 *Given any* $0 < \epsilon < 0.5$*, let* $\mathcal{X} = [0, 1]$*, and* $\mathcal{F} = \{f_1, f_2\}$ *where*

$$f_1(x) = \begin{cases} 1, & x < \epsilon \\ 0, & \textit{otherwise} \end{cases}, f_2(x) = \begin{cases} 1, & x > 1 - \epsilon \\ 0, & \textit{otherwise.} \end{cases}$$

Let $\mathcal{P}_{\mathcal{X}}$ be the uniform distribution over $[0, 1]$. Assume that Y will always be zero. f_1 and f_2 are both f^*. Every pointwise-competitive classifier will have to output $g(x) = 0$ for every x in the disagreement set of f_1 and f_2. $R(f^*) = \epsilon$, and the rejection mass is $2\epsilon (= 2R(f^*))$.

5. From Selective Classification to the Disagreement Coefficient

We now turn to show a reduction from selective classification, to the disagreement coefficient.

Theorem 8 *Let* $\mathcal{F}$ *be a hypothesis class with a finite VC dimension* d*, and* $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ *be an unknown distribution. Let* PCS *be an algorithm that returns a pointwise-competitive selective classifier w.h.p. for all distributions whose marginal is* $\mathcal{P}_{\mathcal{X}}$*. If for every* $m \leq 1/R(f^*)$*,* $0 < \delta$*, with probability of at least* $1-\delta$*, the abstain rate* $1-\Phi$ *of* $\text{PCS}(S_m, \delta, \mathcal{F}, d)$ *is bounded by a monotonic*[5] *polylog as follows:*

$$1 - \Phi(\text{PCS}) \leq \frac{\text{polylog}_0(\text{m}, 1/\delta)}{\text{m}}. \tag{33}$$

Then for every f^* *(every true risk minimizer), for every* $r \geq R(f^*)$*,*

$$\theta_{f^*}(r) \leq 20\left(\text{polylog}_0(1/r, 1/r) + 3\right).$$

Proof For any $m \in \{4, 5, ..., \lfloor 1/R(f^*) \rfloor\}$, denote by $\mathcal{S}_m$ a random training sample drawn from $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. Let Z be a random variable representing a single random unlabeled example sampled from $\mathcal{P}_{\mathcal{X}}$, and let f^* to be a specific true risk minimizer.

For $Z \in DIS\left(B(f^*, \frac{1}{m})\right)$, we use the following argument from Hanneke (2012, Lemma 47). We know that there exists a function $h_Z \in \mathcal{F}$ s.t. $h_Z(Z) \neq f^*(Z)$ and $Pr(h_Z(X) \neq f^*(X)) \leq \frac{1}{m}$. We denote by $\mathcal{P}_{\mathcal{X},\mathcal{Y}_z}$ a new probability distribution, where the marginal, $\mathcal{P}_{\mathcal{X}}$,

5. If the polylog is not monotonic, the following holds: $\theta_{f^*}(r) \leq 16(\text{polylog}(\lfloor 1/r \rfloor, \lfloor 1/r \rfloor) + 3) \cdot \frac{5}{4}$; see Eq. (44).

remains the same, but Y becomes $Y \triangleq h_Z(x)$. Clearly, h_Z is a true risk minimizer (f^*) for such a distribution.

Denote by e_1 the probability event where (33) holds (for a specific $m \leq 1/R(f^*)$). Denote by e_2 the event where PCS has succeeded in returning a pointwise-competitive selective classifier (f_{s_m}, g_{s_m}) under S_m.

Define S'_m to be a modified S_m where, for every $(x, y) \in S_m$ s.t. $h_Z(x) \neq y$, y is modified to become $y = h_Z(x)$. S'_m is a random training sample drawn from $\mathcal{P}_{\mathcal{X},\mathcal{Y}_z}$. Denote by e_{3z} the event where PCS has succeeded in returning a pointwise-competitive selective classifier $(f_{s'_m}, g_{s'_m})$ under S'_m. h_Z is only defined for cases in which $Z \in DIS\left(B(f^*, \frac{1}{m})\right)$, and thus we define that e_{3z} will vacuously hold when $Z \notin DIS\left(B(f^*, \frac{1}{m})\right)$.

Under our assumptions, $\Pr(e_1), \Pr(e_2) \geq 1-\delta$. For every $Z \in DIS(B(f^*, \frac{1}{m}))$, $\Pr(e_{3z}|Z) \geq 1-\delta$, and for every $Z \notin DIS(B(f^*, \frac{1}{m}))$, $\Pr(e_{3z}|Z) = 1$, which implies that $\Pr(e_{3z}) \geq 1-\delta$. We denote by $h_Z(X_m) = Y_m$ the event where $h_Z(x) = y$ for all $(x_i, y_i) \in S_m$. The explanations for the following (in)equalities follow.

$$
\begin{aligned}
& \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m\right\} && (34)\\
= \; & \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m \wedge e_1 \wedge e_2 \wedge e_{3z}\right\} && (35)\\
& + \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m \mid \neg(e_1 \wedge e_2 \wedge e_{3z})\right\} \cdot \Pr(\neg(e_1 \wedge e_2 \wedge e_{3z}))\\
\leq \; & \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m \wedge e_1 \wedge e_2 \wedge e_{3z}\right\} + 3\delta && (36)\\
\leq \; & \Pr\{g_{s_m}(Z) = 0 \wedge e_1 \wedge e_2 \wedge e_{3z}\} + 3\delta && (37)\\
\leq \; & \Pr\{g_{s_m}(Z) = 0 \wedge e_1\} + 3\delta\\
\leq \; & \Pr\{g_{s_m}(Z) = 0 \mid e_1\} + 3\delta\\
\leq \; & \frac{\text{polylog}_0(m, 1/\delta)}{m} + 3\delta. && (38)
\end{aligned}
$$

In (34), it is convenient to view the random experiment as if we draw z first, and then S_m. If $Z \in DIS\left(B(f^*, \frac{1}{m})\right)$, then consider h_Z to be any function that holds $h_Z(Z) \neq f^*(Z)$ and $Pr(h_Z(X) \neq f^*(X) \leq \frac{1}{m})$. If $Z \notin DIS\left(B(f^*, \frac{1}{m})\right)$, then the event described in (34) does not occur, and h_Z is undefined. In (35), we use conditional probability, and in (36) we apply the union bound. Inequality (37) is justified as follows. If $h_Z(X_m) = Y_m$, then the algorithm received the same input under $\mathcal{P}_{\mathcal{X},\mathcal{Y}_z}$ and $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. Given that e_2 and e_{3z} occurred, we know that the algorithm has successfully output a pointwise-competitive selective classifier for both probabilities, which means that whenever f^* and h_Z disagree, g_{s_m} has to output zero; otherwise, it will not be pointwise-competitive for one of the distributions. By the definition of h_Z, $h_Z(Z) \neq f^*(Z)$, which explains the inequality. Inequality (38) follows from the definition of e_1. Taking $\delta = \frac{1}{m}$, we get,

$$\Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m\right\} \leq \frac{\text{polylog}_0(m, m) + 3}{m}. \qquad (39)$$

The following inequalities are derived using elementary conditional probability. In Equation (41) we use an argument taken from the proof of Hanneke (2012, Lemma 47). $h_Z \in$

$B(f^*, \frac{1}{m})$, and thus the probability that f^* and h_Z will have a different label for a specific example is bounded by $1/m$. The probability that f^* will mispredict is by definition $R(f^*)$, and thus, using the union bound, we get (41).

$$\begin{aligned}
& \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right) \wedge h_Z(X_m) = Y_m\right\} \\
= \; & \Pr\left\{h_Z(X_m) = Y_m \mid Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \cdot \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \\
\geq \; & \Pr\left\{f^*(X_m) = Y_m \wedge h_Z(X_m) = f^*(X_m) \mid Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \cdot \qquad (40) \\
& \cdot \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \\
\geq \; & \left(1 - R(f^*) - \frac{1}{m}\right)^m \cdot \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \qquad (41) \\
\geq \; & \left(1 - \frac{1}{m} - \frac{1}{m}\right)^m \cdot \Pr\left\{Z \in DIS\left(B(f^*, \frac{1}{m})\right)\right\} \\
\geq \; & \frac{1}{16} \cdot \Delta B\left(f^*, \frac{1}{m}\right). \qquad (42)
\end{aligned}$$

Combining (39) and (42), we get that for every $m \in \{4, 5, ..., \lfloor 1/R(f^*) \rfloor\}$,

$$\frac{\Delta B(f^*, 1/m)}{1/m} \leq 16\left(\text{polylog}_0(m, m) + 3\right). \qquad (43)$$

The following inequalities follow from (43), and from the fact that $\Delta B(f^*, x)$ and $\text{polylog}_0(\cdot)$ are non-decreasing. For any r in $[R(f^*), \frac{1}{5}]$, and noting that $\frac{1}{\lfloor 1/r \rfloor} \geq r$,

$$\begin{aligned}
\frac{\Delta B(f^*, r)}{r} &\leq \frac{\Delta B\left(f^*, \frac{1}{\lfloor 1/r \rfloor}\right)}{\frac{1}{\lfloor 1/r \rfloor}} \cdot \frac{1}{r \cdot \lfloor 1/r \rfloor} \\
&\leq 16(\text{polylog}_0(\lfloor 1/r \rfloor, \lfloor 1/r \rfloor) + 3) \cdot \frac{1}{r \cdot (1/r - 1)} \\
&\leq 16\left(\text{polylog}_0(\lfloor 1/r \rfloor, \lfloor 1/r \rfloor) + 3\right) \cdot \frac{1}{1 - r} \\
&\leq 16(\text{polylog}_0(\lfloor 1/r \rfloor, \lfloor 1/r \rfloor) + 3) \cdot \frac{5}{4} \qquad (44) \\
&\leq 20(\text{polylog}_0(1/r, 1/r) + 3)
\end{aligned}$$

and for r in $[\frac{1}{5}, 1]$,

$$\frac{\Delta B(f^*, r)}{r} \leq \frac{1}{1/5} = 5, \qquad (45)$$

which concludes the proof.

■

Corollary 9 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution. If for every $m \leq 1/R(f^*)$, $0 < \delta$, with probability of at least $1-\delta$, the abstain rate $1-\Phi$ of* ILESS$(S_m, \delta, \mathcal{F})$ *is bounded by a monotonic* polylog *as follows:*

$$1 - \Phi(\text{ILESS}) \leq \frac{\text{polylog}_0(m, 1/\delta)}{m}.$$

Then for every f^* (every true risk minimizer), for every $r \geq R(f^*)$,

$$\theta_{f^*}(r) \leq 20(\text{polylog}_0(1/r, 1/r) + 3).$$

Proof This is a direct result from Theorem 8, and from the fact that ILESS is PCS. ■

Given $\mathcal{P}_{\mathcal{X}}$,$\mathcal{F}$ and f^*, if any PCS learning algorithm has a fast R^* rejection rate, we can apply Theorem 8 with a deterministic $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ distribution for which $Y = f^*(X)$, and get that $R(f^*) = 0$. Thus, by definition,

$$1 - \Phi(\text{PCS}) \leq \text{polylog}_1\left(\frac{1}{\text{R(f*)} + 1/\text{m}}\right) \cdot 0 + \frac{\text{d} \cdot \text{polylog}_2(\text{m}, 1/\delta)}{\text{m}}.$$

We can now apply Theorem 8 with $R(f^*) = 0$, and get that the disagreement coefficient is bounded by polylog$(1/r)$ for all $r > 0$. Thus, together with Corollary 7, we complete a two sided equivalence from PCS with a fast R^* rejection rate to a disagreement coefficient bounded by polylog$(1/r)$ for all $r > 0$.

6. Active-ILESS

In this section in Strategy 3 we introduce an agnostic active learning algorithm called Active-ILESS. Active-ILESS is very similar to Agnostic CAL and to Oracular CAL (Dasgupta et al., 2007), which were inspired by A^2 (Balcan et al., 2006). Both Agnostic CAL and Oracular CAL use a $LEARN_{\mathcal{H}}(S,T)$ subroutine, which returns an ERM over a certain set T with the constraint that a zero training error must be achieved over a set S (see details in Hsu, 2010). Oracular CAL has an advantage over Agnostic CAL in that it only uses one such constraint while using the $LEARN_{\mathcal{H}}$ subroutine. Nevertheless, as Hsu mentions, this weakens the label complexity analysis of Oracular CAL, and makes it dependent on the square of the disagreement coefficient. Similar to Oracular CAL, Active-ILESS can be implemented with an ERM computation under only one constraint (as seen from the disbelief principle in El-Yaniv and Wiener, 2011, already discussed in Section 4), but its proven label complexity is only dependent linearly on the disagreement coefficient. In this sense, Active-ILESS enjoys the best of both worlds. The label complexity analysis of Active-ILESS appears in Section 8.

As Agnostic CAL and Oracular CAL, Active-ILESS creates artificial labels (step 1), but unlike these other algorithms, Active-ILESS works in batches (inside each batch, the decision whether to query an example is made instantly and not at the end of the batch). This allows Active-ILESS to be a bit more conservative with its deltas. This also allows us to achieve a tighter label complexity.

In Section 7 we use Active-ILESS to show an equivalence between active learning (represented by Active-ILESS) and selective classification (represented by a variant of ILESS,

Strategy 3 Agnostic Low-Error Active Learning Strategy (Active-ILESS)

Input: ϵ and/or m depending on the desired termination condition (error or labeling budget, respectively)
Confidence level δ
Hypothesis class $\mathcal{F}$ with VC dimension d
An unlabeled input sequence sampled i.i.d from $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$: $x_1, x_2, x_3, \ldots$
Output: A classifier $\hat{f}$
Initialize:
Set $\hat{S} = \emptyset$, $G_0 = \mathcal{F}$, $t = 1$
Perform for each example x_t received:
1: if $x_t \in AGR(G_{t-1})$: don't request label for x_t and set $y_t = f(x_t)$ using any $f \in G_{t-1}$
otherwise: request label y_t
2: Set $\hat{S} = \hat{S} \cup \{(x_t, y_t)\}$
3: Set $\hat{f} = \hat{f}(\hat{S})$
4: if $\log_2(t) \in \mathbb{N}$:

- Set $\sigma_{Active} = \hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2}, \frac{\delta}{2t}, d, \hat{R}(\hat{f}, \hat{S})\right) + \bar{\sigma}_{\hat{R}-R}\left(\frac{t}{2}, \frac{\delta}{2t}, d, \hat{R}(\hat{f}, \hat{S}) + \hat{\sigma}_{R-\hat{R}}(\frac{t}{2}, \frac{\delta}{2t}, d, \hat{R}(\hat{f}, \hat{S}))\right)$
- If ϵ was given as input and $\sigma_{Active} < \epsilon$, terminate and return $\hat{f}$
- If m was given as input and $t > m/2$, terminate and return $\hat{f}$
- Set $G_t = \hat{\mathcal{V}}\left(\hat{f}, \sigma_{Active}\right)$
- Set $\hat{S} = \emptyset$

otherwise:

- $G_t = G_{t-1}$

5: Set t = t +1

"Batch-ILESS"). The introduction of these new variants facilitates a straightforward proof of the equivalence relationship. This equivalence implies a novel relationship between selective and active classification in the agnostic setting.

We begin by analyzing Active-ILESS and showing that much like ILESS, $f^* \in G_t$ in each iteration t. The low-error set G, maintained by ILESS, contains all the hypotheses that have an empirical error smaller than $\hat{R}(\hat{f}) + \sigma_{\text{ILESS}}$. In Lemma 1 we showed that this condition implies that f^* resides within the low-error set G of ILESS. A proof that $f^* \in G_t$, after each iteration of Active-ILESS, cannot follow the same argument due to the fact that Active-ILESS, shown in Strategy 3, labels by itself each example whose label is not requested from the teacher. Further, since we consider an agnostic setting, these self-labels can differ from the true labels.

Active-ILESS, as seen in Strategy 3, receives as a termination condition either $\epsilon > 0$ and/or m, and terminates when the radius of its low-error set, G_t, is smaller than ϵ, or when it has processed m examples.

Active-ILESS changes its low-error set, G_t, only for t that are natural powers of 2. For each change, Active-ILESS begins to create fake labels for $x_t \in AGR(G_{t-1})$ that may or may not be equal to the real label of x_t (under the original distribution). In fact, this G_t defines a new distribution, $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)$, and this distribution changes for every t that is a natural power of 2. With respect to a run of Active-ILESS, and $t = 2^i, i \in \mathbb{N}$, we denote by $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)$, the new probability distribution implied by G_t, and the fake labels created by the algorithm. $R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f)$ will be the true risk under the new distribution, while $R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f)$ is the true risk of f under the original distribution.

Definition 6.1 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution. Given a run of Active-ILESS, we denote by $\mathcal{K}$ the event where both inequalities (46) and (47) hold simultaneously for every $f \in \mathcal{F}$, for all iterations of Active-ILESS where $t = 2^i, i \in \mathbb{N}$. $\hat{R}(f) \triangleq \hat{R}(f, \hat{S})$ for $\hat{S}$ before it was initialized:*

$$R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f) \leq \hat{R}(f) + \sigma_{R-\hat{R}}\left(\frac{t}{2}, \frac{\delta}{2t}, d, R(f), \hat{R}(f)\right) \tag{46}$$

$$\hat{R}(f) \leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f) + \sigma_{\hat{R}-R}\left(\frac{t}{2}, \frac{\delta}{2t}, d, R(f), \hat{R}(f)\right) \tag{47}$$

Lemma 10 *$\mathcal{K}$ occurs with probability of at least $1 - \delta$.*

Proof G_t changes only for iterations of the type $2^i, i \in \mathbb{N}$. We know by Lemma 1 that the probability that inequalities (46) and (47) do not hold is smaller than $\delta/(2t)$. By the union bound, the probability that one of these inequalities does not hold after any iteration is smaller than

$$\sum_{t=2^i, i\in\mathbb{N}} \frac{\delta}{2t} \leq \delta.$$

■

Lemma 11 *If f^*, a true risk minimizer under probability distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, resides within G_t, then it is also a true risk minimizer under probability distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)$.*

Proof

$$\operatorname*{argmin}_{f\in\mathcal{F}} R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f) = \operatorname*{argmin}_{f\in\mathcal{F}} \left(\underbrace{R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f)}_{A} + \underbrace{R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f) - R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f)}_{B} \right).$$

We know that f^* minimizes A, and we note that every function that resides within G_t minimizes B, because every difference in the labeling between $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ and $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)$ was done according to the label given by the unanimous decision of functions in G_t. In other words, whenever $x \in AGR(G_t)$, the labelling of x is done according to any function in G_t, for instance, f^*. Thus,

$$\begin{aligned} R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f) - R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f) &= \mathbb{E}\left[\mathbb{1}[X \in AGR(G_t)](\mathbb{1}[f(X) \neq f^*(X)] - \mathbb{1}[f(X) \neq Y])\right] \\ &= \mathbb{E}\left[\mathbb{1}[X \in AGR(G_t) \wedge f^*(X) \neq Y] \cdot (-1)^{\mathbb{1}(f(X)=f^*(X))}\right] \end{aligned}$$

and we see that f^* minimizes the last term. Hence, f^* minimizes $A + B$. ■

The proofs of the following four lemmas appear in Appendix A. They all show basic good qualities of Active-ILESS.

Lemma 12 *Given that event $\mathcal{K}$ (see Definition 6.1) occurred, each f^* of the **original** distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ resides within G_t for all t. This implies that $R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_t)}(f^*) \leq R(f^*)$, for all t, as every change in the labeling is done according to f^*.*

Lemma 13 *Given that event $\mathcal{K}$ (see Definition 6.1) occurred, and under the assumption that Active-ILESS terminated with the ϵ condition, the hypothesis returned by Active-ILESS, $\hat{f}$, holds:*

$$R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(\hat{f}) \leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f^*) + \epsilon.$$

Lemma 14 *Given that event $\mathcal{K}$ (see Definition 6.1) occurred, the final radius of Active-ILESS satisfies*

$$\sigma_{Active} = O\left(\frac{B}{m} + \sqrt{\frac{B}{m} \cdot R(f^*)}\right), \tag{48}$$

where $B \triangleq 16d\ln(\frac{16m^2e}{d\delta})$.

Lemma 15 *Given that event $\mathcal{K}$ (see Definition 6.1) occurred, the total number of examples that Active-ILESS(ϵ) processed (without necessarily requesting labels) is*

$$O\left(\frac{1}{\epsilon}\ln(\frac{1}{\epsilon}) + \frac{R(f^*)}{\epsilon^2}\ln(\frac{R(f^*)}{\epsilon^2})\right),$$

where we hide factors of $d, \ln(1/\delta)$ under the O.

Definition 6.2 *An active learner that generates a hypothesis whose true error is smaller than ϵ w.h.p., has **passive example complexity**, if it observes up to*

$$O\left(\frac{1}{\epsilon}\ln(\frac{1}{\epsilon}) + \frac{R(f^*)}{\epsilon^2}\ln(\frac{R(f^*)}{\epsilon^2})\right)$$

examples (not necessarily labeled).

By Lemmas 13 and 15 we know that Active-ILESS has passive example complexity.

The definition of a fast R^* rejection rate for selective classification induces the following related definition for the exponential speedup of active learning algorithms.

Definition 6.3 (R^* Exponential Speedup) *Given $\mathcal{P}_{\mathcal{X}}$,$\mathcal{F}$ and f^*, we say that an active learner has an R^* **exponential speedup**, with* polylog_1 *and* polylog_2 *as its parameters, if for every $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ for which f^* is a true risk minimizer, and for every $m > 0$, with probability of at least $1-\delta$, the number of labels requested by the active learner after observing m examples is not greater than*

$$\text{polylog}_1\left(\frac{1}{R(f^*) + 1/m}\right) \cdot R(f^*)m + d \cdot \text{polylog}_2(m, 1/\delta).$$

Hsu (2010) introduced the agnostic CAL algorithm and showed (Theorem 4.3, page 41) that if the disagreement coefficient is bounded, then Agnostic CAL has an R^* exponential speedup (under our new definition). Any active algorithm that has passive example complexity and achieves an R^* exponential speedup requires w.h.p. no more than $O\left(\text{polylog}(\frac{R(f^*)}{\epsilon^2})\frac{R(f^*)^2}{\epsilon^2} + \text{polylog}(\frac{1}{\epsilon})\right)$ labels to reach a true error smaller than ϵ. The proof is immediate by considering the cases $\frac{R(f^*)}{\epsilon} \geq 1$ and $\frac{R(f^*)}{\epsilon} < 1$. The leading term of this bound is $\frac{R(f^*)^2}{\epsilon^2}$, which is also the case for A^2 (Balcan et al., 2006).

7. A Reduction from Active-iLess to Batch-ILESS

In Strategy 4 we define a selective classifier, called Batch-ILESS, which uses Active-ILESS as its engine. Given a labeled sample S_m, Batch-ILESS simulates the active algorithm, by applying it over S_m in a straightforward manner (i.e., it sequentially introduces to the active algorithm an unlabeled example and reveals the label only if the active algorithm requests it). Upon termination, after the active algorithm has consumed all examples, our batch algorithm receives $\hat{f}$ from the active algorithm and uses its last low-error set G_t to define its selection function.

Lemma 12 implies that Batch-ILESS is pointwise-competitive. We note that Lemma 4, Theorem 5 and Theorem 9, which were proven for ILESS, can also be proven for Batch-ILESS. We chose to prove it for ILESS, as it is simpler than Batch-ILESS, and does not require an active algorithm as its engine. We state these ideas formally, and give sketches of their proofs, in Lemma 23 and Theorem 24 which appear in Appendix C.

Strategy 4 Batch Improved Low-Error Selective Strategy (Batch-ILESS)

Input: Sample set of size m, S_m,
Confidence level δ
Hypothesis class $\mathcal{F}$ with VC dimension d
Output: A selective classifier (h, g)
1: Simulate Active-ILESS with S_m as its input stream; let G_t be the low-error set obtained by Active-ILESS in its last round, and let $\hat{f}$ be its resulting classifier.
2: Construct g such that $g(x) = 1 \iff x \in \{\mathcal{X} \setminus DIS(G_t)\}$
3: $h = \hat{f}$

Theorem 16 shows a deep connection between the speedup of Active-ILESS and the rejection mass of Batch-ILESS for specific $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. An immediate corollary of this theorem is that if Active-ILESS has R^* exponential speedup (see Definition an 6.3), then Batch-ILESS has a fast R^* rejection rate (see Definition 4.2).

Theorem 16 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution. If for every $m \leq m_{max}$, after observing m examples, with probability of at least $1-\delta$, the number of labels requested by Active-ILESS is not greater than*

$$\text{polylog}_1\left(\frac{1}{R(f^*)+1/m}\right) \cdot R(f^*)m + d \cdot \text{polylog}_2(m, 1/\delta),$$

then the rejection mass of Batch-ILESS for every $m \leq \frac{m_{max}}{2}$ is bounded w.h.p. by

$$8 \cdot \text{polylog}_1\left(\frac{1}{R(f^*)+1/m}\right) \cdot R(f^*) + \frac{2\left(\sqrt{\ln(2/\delta)} + \sqrt{\ln(2/\delta) + 2d \cdot \text{polylog}_2(2m, 2/\delta)}\right)^2}{m}.$$

Proof Consider an application of Active-ILESS with $\delta = \delta_0$ over $m_0 \triangleq 2^{\lceil \log(m+1) \rceil}$ examples. Denote by X_i an indicator random variable for the labeling of its ith example, $1 \leq i \leq m_0$. With probability of at least $1-\delta_0$ over the choice of samples from $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$,

$$\sum_{i=1}^{m_0} X_i \leq \text{polylog}_1\left(\frac{1}{R(f^*)+1/m_0}\right) \cdot R(f^*)m_0 + d \cdot \text{polylog}_2\left(m_0, 1/\delta_0\right). \tag{49}$$

We know by the definition of Active-ILESS (Strategy 3), that the last $m_0/2$ examples had the exact same probability, $\Delta G_{m_0/2}$, of requiring a label, and that this is exactly the probability that Batch-ILESS will decide to abstain after receiving m examples, according to Strategy 4.

We now estimate $\Delta G_{m_0/2}$ using the version of the Chernoff bound given by Canny (2012). For the sake of self-containment, Canny's statement and proof of the bound are provided in Lemma 21 in Appendix B.

The statement of the lemma is as follows. Let $Z_1, Z_2, \ldots, Z_n$ be independent Bernoulli trials with $Pr[Z_i = 1] = p$, let $Z \triangleq \sum_{i=1}^{n} Z_i$, and $\mu = \mathbb{E}Z$. Then, for every $\alpha > 0$:

$$\Pr\left(Z < (1-\alpha)\mu\right) \leq \exp(-\mu\alpha^2/2).$$

Applying the Chernoff bound with the indicator variables of the last $m_0/2$ examples, we have $X = \sum_{m_0/2}^{m_0} X_i$, $\mu = p\frac{m_0}{2}$, and set $p \triangleq \Delta G_{m_0/2}$. Select α such that

$$\exp\left(-p\frac{m_0}{2}\alpha^2/2\right) = \delta_2.$$

Solving for α,

$$\alpha = \sqrt{\frac{4\ln(1/\delta_1)}{m_0 p}}.$$

We conclude that with probability of at least $1-\delta_1$,

$$\begin{aligned} & X \geq \left(1 - \sqrt{\frac{4\ln(1/\delta_1)}{m_0 p}}\right) \cdot p\frac{m_0}{2} \\ \Leftrightarrow \quad & 0 \geq \frac{pm_0}{2} - \sqrt{pm_0 \cdot \ln(1/\delta_1)} - X. \end{aligned} \tag{50}$$

Solving the quadratic equation (50) for $\sqrt{pm_0}$, we get that

$$\begin{aligned} & \sqrt{pm_0} \leq \sqrt{\ln(1/\delta_1)} + \sqrt{\ln(1/\delta_1) + 2X} \\ \Rightarrow \quad & p \leq \frac{(\sqrt{\ln(1/\delta_1)} + \sqrt{\ln(1/\delta_1) + 2X})^2}{m_0}. \end{aligned} \tag{51}$$

Combining (49) and (51), from the union bound we get that with probability of at least $1 - \delta_0 - \delta_1$,

$$\begin{aligned} & \Delta G_{m_0/2} \leq \\ & \leq \frac{\left(\sqrt{\ln(\frac{1}{\delta_1})} + \sqrt{\ln(\frac{1}{\delta_1}) + 2\text{polylog}_1\left(\frac{1}{R(f^*)+1/m_0}\right) R(f^*)m_0 + 2d \cdot \text{polylog}_2(m_0, \frac{1}{\delta_0})}\right)^2}{m_0}. \end{aligned}$$

If we take $\delta_0 = \delta_1 = \delta/2$, then, since $m \leq m_0 \leq 2m$, we can use $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $(a+b)^2 \leq 2a^2 + 2b^2$, to obtain

$$\begin{aligned}
& \Delta G_{m_0/2} \leq \\
\leq & \frac{\left(\sqrt{\ln(\frac{2}{\delta})} + \sqrt{\ln(\frac{2}{\delta}) + 4\text{polylog}_1\left(\frac{1}{R(f^*)+1/m}\right) \cdot R(f^*)m + 2d \cdot \text{polylog}_2(2m, \frac{2}{\delta})}\right)^2}{m} \\
\leq & \frac{\left(\sqrt{\ln(\frac{2}{\delta})} + \sqrt{\ln(\frac{2}{\delta}) + 2d \cdot \text{polylog}_2(2m, \frac{2}{\delta})} + \sqrt{4\text{polylog}_1\left(\frac{1}{R(f^*)+1/m}\right) R(f^*)m}\right)^2}{m} \\
\leq & \frac{2\left(\sqrt{\ln(\frac{2}{\delta})} + \sqrt{\ln(\frac{2}{\delta}) + 2d \cdot \text{polylog}_2(2m, \frac{2}{\delta})}\right)^2}{m} + \\
& + \frac{2\left(\sqrt{4\text{polylog}_1\left(\frac{1}{R(f^*)+1/m}\right) \cdot R(f^*)m}\right)^2}{m} \\
= & \frac{2\left(\sqrt{\ln(\frac{2}{\delta})} + \sqrt{\ln(\frac{2}{\delta}) + 2d \cdot \text{polylog}_2(2m, \frac{2}{\delta})}\right)^2}{m} + \\
& + 8 \cdot R(f^*) \cdot \text{polylog}_1\left(\frac{1}{R(f^*) + 1/m}\right)
\end{aligned}$$

■

Corollary 17 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution. If for every $m \leq 2/R(f^*)$, $0 < \delta$ after observing m examples, with probability of at least $1 - \delta$, the number of labels requested by Active-ILESS is not greater than*

$$\text{polylog}_1\left(\frac{1}{R(f^*) + 1/m}\right) \cdot R(f^*)m + d \cdot \text{polylog}_2(m, 1/\delta),$$

then for every $r \geq R(f^*)$,

$$\begin{aligned}
\theta_{f^*}(r) &\leq 20\left(2\left(\sqrt{\ln(2/r)} + \sqrt{\ln(2/r) + 2d \cdot \text{polylog}_2(2/r, 2/r)}\right)^2 + 8 \cdot \text{polylog}_1(1/r) + 3\right) \\
&= O\left(d \cdot \text{polylog}(1/r)\right).
\end{aligned}$$

Proof The proof follows from Theorems 16 and 8. Applying Theorem 16, we know that for $m \leq 1/R(f^*)$, the rejection mass of Batch-ILESS is bounded w.h.p. by,

$$\frac{2\left(\sqrt{\ln(2/\delta)} + \sqrt{\ln(2/\delta) + 2d \cdot \text{polylog}_2(2m, 2/\delta)}\right)^2 + 8 \cdot \text{polylog}_1(\frac{1}{R(f^*)+1/m})}{m}.$$

We know by Lemma 12 that $f^* \in G_t$, and thus Batch-ILESS is PCS. We apply Theorem 8, and get that for every $r \geq R(f^*)$,

$$\begin{aligned}\theta_{f^*}(r) &\leq 20\left(2\left(\sqrt{\ln(\frac{2}{r})}+\sqrt{\ln(\frac{2}{r})+2d\cdot\text{polylog}_2(\frac{2}{r},\frac{2}{r})}\right)^2+8\cdot\text{polylog}_1\left(\frac{1}{R(f^*)+r}\right)+3\right)\\ &\leq 20\left(2\left(\sqrt{\ln(\frac{2}{r})}+\sqrt{\ln(\frac{2}{r})+2d\cdot\text{polylog}_2(\frac{2}{r},\frac{2}{r})}\right)^2+8\cdot\text{polylog}_1(1/r)+3\right).\end{aligned}$$

■

8. From the Disagreement Coefficient to Active Learning

In this section we show that when $\theta'(r)$ is bounded by $\text{polylog}_1(1/r)$ for all $r > R(f^*)$ for some specific $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, then the label complexity of Active-ILESS under the same $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ is bounded by

$$\text{polylog}_2\left(\frac{1}{R(f^*)+1/m}\right)\cdot R(f^*)m + d\cdot\text{polylog}_3(m,1/\delta), \tag{52}$$

where the parameters of polylog_2 and polylog_3 are only dependent on $\text{polylog}_1(1/r)$. Thus, if $\theta'(r) \leq \text{polylog}_1(1/r)$ for all $r > 0$, we get that Active-ILESS has an R^* exponential speedup. This direction has been shown before in Hsu (2010); Hanneke (2007) for Oracular CAL, Agnostic CAL and A^2. For the sake of self-containment, we show it here for Active-ILESS. Due to the fact that Active-ILESS relies on ILESS, for which we already have bounds, the proof is straightforward.

As a preparation for Theorem 19, we present Lemma 18 (shown before in Hanneke, 2014b, Theorem 7.1), which introduces a small feature of the disagreement coefficient that will serve us later.

Lemma 18 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, and let $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution. For every $f \in \mathcal{F}$ and $0 < r \leq 1$, $\theta_f(r)\cdot r$ is a non-decreasing function.*

Theorem 19 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution, and f^*, a true risk minimizer of $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. The label complexity of Active-ILESS$(m,\delta/2)$ is bounded w.h.p. by*

$$\theta'\left(5R(f^*)+14\frac{A}{m}\right)\cdot 2e\cdot mR(f^*)+\log_2(2/\delta)+56e\cdot\log_2 m\cdot A\cdot\theta'\left(5R(f^*)+14\frac{A}{m}\right).$$

Proof Each run of Active-ILESS$(m,\delta/2)$ simulates $\lceil\log_2 m\rceil$ runs of ILESS. Denote by T indices of iterations where $\log_2(t) \in \mathbb{N}$. We know by Lemma 10 that with probability of at least $1-\delta/2$, inequalities (46) and (47) hold for each run. Recall that we denoted by $\mathcal{K}$ the event where both inequalities hold throughout all runs of ILESS, which is exactly the definition of event $\mathcal{E}$ per run (see Definition 4.1). Under event $\mathcal{K}$, Lemma 12 implies that

all f^* of the original distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ reside within G_T for all T. This also implies that all f^* of the original distribution remain the true risk minimizers under $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_T)$, for all T, as they always benefit from the creation of the artificial labels.

Because the marginal of the distribution does not change during the run of Active-ILESS, and because event $\mathcal{E}$ holds for each iteration of ILESS, we can apply Theorem 5 for all the runs of ILESS. We thus get that for every run of ILESS, which we denote by ILESS_T,

$$G_T \subseteq B(f^*, R_T),$$

and thus, the rejection mass is bounded as follows

$$1 - \Phi(\text{ILESS}_T) \leq \Delta B(f^*, R_T) \leq \theta(R_T) \cdot R_T,$$

where

$$R_T \triangleq 2 \cdot R(f^*) + 11 \cdot \frac{A}{T} + 6 \cdot \sqrt{\frac{A}{T} \cdot R(f^*)}.$$

We denoted by $R(f^*)$ the true error according to the original distribution, which might be larger than the true error implied by the fake label distributions that the algorithm induces. According to Lemma 18, enlarging R_T can only weaken the bound, and thus, there is no problem doing so. We additionally bound R_T using $\sqrt{ab} \leq a/2 + b/2$ to get

$$R_T \leq 5 \cdot R(f^*) + 14 \cdot \frac{A}{T}.$$

and thus,

$$\Delta B(f^*, R_T) \leq \theta' \left(5 \cdot R(f^*) + 14 \cdot \frac{A}{T}\right) \cdot \left(5 \cdot R(f^*) + 14 \cdot \frac{A}{T}\right).$$

In this proof we bound the sum $\sum_{t=1}^{m} X_t$, where $X_t \triangleq \mathbb{1}[x_t \in \Delta G_t]$, as ΔG_t is by definition the probability that Active-ILESS requests a label. To do so, we will bound the sum $\sum_{t=1}^{m} P_t$ where $P_t \triangleq \mathbb{1}[x_t \in \Delta B(f^*, R_t)]$, as we know that when $\mathcal{K}$ holds, $\Delta G_t \subseteq \Delta B(f^*, R_t)$.

According to the definition of G_t in Strategy 3, the probability distribution of the artificial labeling done by Active-ILESS changes only when t is a natural power of 2. Thus, R_t only changes when $\log_2(t) \in \mathbb{N}$, and we get that

$$P_t \leq \theta' \left(5 \cdot R(f^*) + 14 \cdot \frac{A}{T}\right) \cdot 5R(f^*) + \frac{14A \cdot \theta'\left(5 \cdot R(f^*) + 14 \cdot \frac{A}{T}\right)}{T}, \tag{53}$$

where $T = 2^{\lfloor \log_2(t-1) \rfloor - 1}$.

We now have a series of Poisson trials, $P_1, P_2, \ldots, P_m$. We use a version of the Chernoff bound (as shown by Canny, 2012) to bound the label complexity.[6] The statement and a sketch of the proof of this bound are provided in Lemma 22 in Appendix B.

For independent Poisson variables $P_1, P_2, \ldots, P_m$, where $Pr[P_i = 1] = p_i$, $P \triangleq \sum_{i=1}^{n} P_i$, and $\mu = \mathbb{E}P$, for every $\alpha > 2e - 1$:

$$\Pr(P > (1 + \alpha)\mu) \leq 2^{-\mu\alpha}.$$

6. This useful bound was used by Hanneke (2014a, Theorem 5.4).

To bound $\mu = \mathbb{E}P$ from above, we use Inequality (53) and plug it into the definition of μ.

$$\begin{aligned}
\mu &= P_1 + P_2 + \sum_{i=3}^{m} P_t \\
&\leq 2 + \sum_{k=1}^{\lceil \log_2 m \rceil - 1} 2^k P_{2^{k+1}} \\
&\leq 2 + m \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right) \cdot R(f^*) + \sum_{k=1}^{\lceil \log_2 m \rceil - 1} 2^k \frac{14A \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right)}{2^{k-1}} \\
&\leq 2 + m \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right) \cdot R(f^*) + 28 \log_2 m \cdot A \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right).
\end{aligned} \tag{54}$$

We need to choose an α that satisfies both $2^{-\mu\alpha} \leq \delta/2$, and $\alpha > 2e - 1$. Clearly, $\alpha = \frac{\log_2(2/\delta)}{\mu} + 2e - 1$ suffices. Hence, we get that with probability of at least $1 - \delta/2$,

$$\begin{aligned}
P &\leq \left(1 + \frac{\log_2(\frac{2}{\delta})}{\mu} + 2e - 1 \right) \mu \\
&= \log_2(\frac{2}{\delta}) + 2e\mu.
\end{aligned}$$

Inequality (54) holds with probability of at least $1 - \delta/2$, and using the union bound, we get that with probability of at least $1 - \delta$,

$$\begin{aligned}
P &\leq \log_2(\frac{2}{\delta}) + \\
&\quad + 2e \left(2 + m \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right) \cdot R(f^*) + 28 \log_2 m \cdot A \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right) \right) \\
&= \theta' \left(5R(f^*) + 14\frac{A}{m} \right) 2e \cdot mR(f^*) + \log_2(\frac{2}{\delta}) + 56e \cdot \log_2 m \cdot A \cdot \theta' \left(5R(f^*) + 14\frac{A}{m} \right)
\end{aligned} \tag{55}$$

■

Corollary 20 *Let $\mathcal{F}$ be a hypothesis class with a finite VC dimension d, $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown distribution, and f^* a true risk minimizer of $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$. If for all $r > R(f^*)$,*

$$\theta'(r) \leq \text{polylog}_1(1/r),$$

then the label complexity of Active-ILESS$(m, \delta/2)$ is bounded w.h.p. by

$$\text{polylog}_1 \left(\frac{1}{5R(f^*) + 14\frac{A}{m}} \right) 2e \cdot mR(f^*) + \log_2(2/\delta) + 56e \cdot \log_2 m \cdot A \cdot \text{polylog}_1 \left(\frac{1}{5R(f^*) + 14\frac{A}{m}} \right),$$

which has the same form as Equation (52).

Hsu (2010) calculated the label complexity of Agnostic CAL and Oracular CAL under the assumption that the disagreement coefficient is bounded by a constant, θ for all $r > 0$. If we adopt this assumption and use it in our analysis, we get that the label complexity is bounded by

$$O\left(\theta m R(f^*) + \ln(1/\delta) + \theta d \ln m \ln\left(\frac{m}{d\delta}\right)\right).$$

This result is nearly identical (up to some logarithmic factors) to the result shown by Hsu (2010, Theorem 4.3) for Agnostic CAL. Agnostic CAL, however, as mentioned before, requires calculation of ERM under several constraints. Oracular CAL only requires one constrained ERM calculation, and its label complexity depends on $\theta^2 \ln^3 m$ (Hsu, 2010, Theorem 5.2). Our algorithm thus enjoys the good properties of both algorithms.

The dominant factor of Equation (55), if we ignore the logarithmic factors, is $mR(f^*)$. Active-ILESS has passive example complexity (see Definition 6.2), which means that the total sample complexity is bounded by $\tilde{O}(\frac{1}{\epsilon} + \frac{R(f^*)}{\epsilon^2})$, where $\tilde{O}(\cdot)$ hides logarithmic factors. Plugging the sample complexity into m in (55), we get that the total label complexity is bounded by $\tilde{O}(\frac{R(f^*)^2}{\epsilon^2})$, in cases for which ILESS has a fast R^* rejection rate. Kääriäinen (2006, Theorem 3) showed that for every AL algorithm, under a specific (non-trivial) hypothesis class $\mathcal{F}$, there exists a deterministic target function g, and a marginal distribution $\mathcal{P}_\mathcal{X}$, s.t. the label complexity is $\tilde{\Omega}(\frac{R(f^*)^2}{\epsilon^2})$ (where $\tilde{\Omega}(\cdot)$ hides logarithmic factors).

9. Concluding Remarks

In this paper we focused on disagreement-based methods. Namely, we always required that f^* remains inside a low-error subset of hypotheses w.h.p., and made decisions based on disagreement considerations. We have always chosen to abstain (in selective classification) or require a label (in AL) whenever there was no consensus in the low error set for a given example. However, this approach is not necessary when you have an epsilon budget for error. Zhang and Chaudhuri (2014) presented an AL algorithm that uses a confidence-rated predictor to decide which labels to query. This predictor (see top of page 6 in their paper) uses LP to minimize the query probability under a constraint on the total error. They also provided a new complexity measure, which is nicely presented in Hanneke (2016), and termed φ_c. Up to certain constants, φ_c replaces the disagreement coefficient in the error bounds of Equation (16) and Theorem 16 in that paper. φ_c is a potential improvement because it is shown that $\varphi_c(r) \leq \theta_f(r)$. Zhang and Chaudhuri show an example where this improvement can be as large as $\sqrt{d}$ for linear classification under the log-concave distribution.

In our paper we focused only on consensus based decisions. This approach did not affect our results significantly as our main concern was the dependency on $1/\epsilon$ (logarithmic or linear). Consensus was an important element in the proof of Theorem 8. Pointwise-competitive selective classifiers are based on consensus decisions, and thus the LP idea is not an appropriate AL equivalent for PCS. Nevertheless, our reduction from PCS to $\theta_{f^*}(r)$ (Theorem 8), is immediately translated to $\varphi_c(r)$, as $\varphi_c(r) \leq \theta_{f^*}(r)$. It is interesting, however, to define a selective classifier that is not PCS, has an error budget, and apply the LP technique. We leave this for future work.

We introduced a new selective classification algorithm, called ILESS, whose rejection "engine" uses sharp generalization bounds (which depend on $R(f^*)$). Our analysis proves

that ILESS has sometimes significantly better rejection guarantees relative to the best known pointwise-competitive selective strategy of Wiener and El-Yaniv (2015). Moreover, the guarantees we provide for ILESS do not depend at all on the Bernstein assumption. For the general agnostic setting, we showed an equivalence relation between pointwise-competitive selective classification, active learning with Active-ILESS, and the disagreement coefficient (see Figure 1). This equivalence is formulated in terms of a fast R^* rejection rate and an R^* exponential speedup (Definitions 4.2 and 6.3).

Theorems 8 and 5 show that selective classification with a fast R^* rejection rate is completely equivalent to having a disagreement coefficient bounded by polylog$(1/r)$ for $r > 0$. In Section 6, in Strategy 3, we define Active-ILESS using ILESS implicitly as its engine (see Statement 4 in Strategy 3). We can replace ILESS with another pointwise-competitive selective algorithm, and thus construct a new active learner, that queries a label whenever the selective classifier abstains, and create a fake label according to the decision of the classifier whenever it decides to predict. Because the selective predictor is pointwise-competitive, we know that the underlying distribution induced by its fake labels is equivalent to a distribution defined by a deterministic labeling according to f^* and the same $\mathcal{P}_{\mathcal{X}}$. The algorithm will terminate using the exact same termination condition as Active-ILESS (when $\sigma_{Active} < \epsilon$), and thus the total sample complexity (labeled and unlabeled examples) will remain the same. The change will only be in the labeling criterion. Lemmas 11, 12, 13, 14, and 15 can all be generalized to such an algorithm.

Going in the other direction to create a selective classifier from a general active learner is more challenging. If the active learner, however, follows the Active-ILESS paradigm, and in particular, uses a pointwise-competitive selective classifier to decide on label requests, then a new pointwise-competitive selective classifier can be created in the same way that Batch-ILESS was created. We can then restate Theorem 16, providing a reduction from an R^* exponential speedup of the active algorithm to a fast R^* rejection rate of the selective classifier.

Disagreement-based decision making in active and selective learning leads to "defensive" algorithms. For example, in the active learning case, this means that a defensive algorithm will ask for more labels than a more aggressive algorithm. In selective classification, this defensiveness provides the power to be pointwise-competitive, but at a cost of an increased rejection rate. It would be interesting to consider more aggressive algorithms that could, for example, take into consideration an estimation of $\mathcal{P}_{\mathcal{X}}$ in order to ignore examples that cause disagreement only between functions that are very similar to each other (in terms of the probability mass of their difference). Such algorithms can be seen in Dasgupta (2005); Freund et al. (1997); Gonen et al. (2013); Zhang and Chaudhuri (2014), for the realizable and the low-error scenarios. We believe that there is still work to be done for the agnostic scenario.

Many aggressive algorithms could be devised under assumptions about knowledge of $\mathcal{P}_{\mathcal{X}}$ (that could be acquired during the algorithm run, and is given in the transductive case), or in a Bayesian setting where a prior distribution on $\mathcal{F}$ exists. When researching this direction, one might also want to define a cost over unlabeled examples, and discuss the trade-off between labeled and unlabeled examples. The main open question inspired by our results would be to identify similar correspondence between aggressive selective classification algorithms and aggressive active learners.

Another aspect of selective classification and active learning, which was not addressed in this paper, is differentiating between more and less noisy areas of the distribution. If we define the noise to be the error of the Bayes classifier, and we assume that the noise behaves similarly in close areas (for instance, we assume that the derivative of the noise is bounded), we can derive an active learner that will estimate the noise, and take it into consideration. A noisy area could be defined as an area for which even the best classifier in the class could not achieve a low-error. This motivates a new type of labeling for selective prediction, where one can abstain for two reasons: (i) lack of knowledge in a specific region of $\mathcal{X}$, i.e., not enough examples were observed in that region, and the generalization bounds are not sufficiently tight. (ii) The region was well explored, but even the best classifier performs poorly, and thus the answer is unknown (the region is noisy). In our paper, an active learner will query for both scenarios; however, a more clever active learner might only query examples of the first type, as examples of the second type cannot reduce its error.

Consider the following experiment in Active Learning; Suppose we are trying to learn a hyperspace hypothesis over R^2, where we have a Gaussian distribution of positive points located at $(-1, 0)$, and another Gaussian distribution of negative points located at $(1, 0)$. A simple active learner that is based on distance from the margin, will quickly converge to querying points that have $x \sim 0$, but will have no preferation over the value of y. Now consider another Gaussian distribution of noisy points, located at $(0, 1)$, where each point is -1 or 1 with equal probability. Querying in areas where $y \sim 1$ is almost completely unbeneficial, however, this is still very close to the margin. We devised a simple active learner that creates a noise hit map over R^2, and consider the noise as part of its heuristics. Thus, achieving faster convergence over the usual margin based algorithm. Of course, this is only a toy example, and when running this experiment over some few real world data sets, we got no improvement. We believe that some more experimental work in this area, might prove to be beneficial.

There has been a lot of progress on the theoretical level, in the field of active learning with noise estimation (Locatelli et al., 2017; Hanneke, 2017; Minsker, 2012; Hanneke and Yang, 2015) However most of the work requires either the Tsybakov noise condition (Tsybakov, 2004), a finite VC dimension, or some other strong smoothing assumptions. It would be interesting to see some experimental work on real data sets, accompanied with some theoretical work with a simplistic noise locality assumption.

Acknowledgments

We are indebted to an anonymous referee for extremely useful comments and for fixing an error in an earlier version of this manuscript. This research was supported by The Israel Science Foundation (grant No. 1890/14).

Appendix A. Proofs for Lemmas in Section 6

Proof of Lemma 12 We prove the claim by induction over t for which G_t is different from G_{t-1}. The base case of the induction is clear. We now show that functions that are true risk minimizers of $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})$ reside within G_t. According to Lemma 11, f^* is a true risk minimizer under $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})$ (given the induction hypothesis), and hence will also be within

G_t. We refer by f^* to a true risk minimizer according to $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})$. Using Inequality (47) and the definition of $\bar{\sigma}_{\hat{R}-R}$,

$$\begin{aligned}\hat{R}(f^*,\hat{S}) &\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*)+\sigma_{\hat{R}-R}\left(\frac{t}{2},\frac{\delta}{2t},d,R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*),\hat{R}(f^*,\hat{S})\right)\\ &\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*)+\bar{\sigma}_{\hat{R}-R}\left(\frac{t}{2},\frac{\delta}{2t},d,R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*)\right),\end{aligned} \tag{56}$$

and by Inequality (46) and the definition of $\hat{f}$ we get,

$$\begin{aligned}R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*) &\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(\hat{f})\\ &\leq \hat{R}(\hat{f},\hat{S})+\sigma_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(\hat{f}),\hat{R}(\hat{f},\hat{S})\right)\\ &\leq \hat{R}(\hat{f},\hat{S})+\hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right).\end{aligned} \tag{57}$$

Plugging (57) into (56) we get,

$$\begin{aligned}\hat{R}(f^*,\hat{S}) \leq\ & \hat{R}(\hat{f},\hat{S})+\hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right)\\ &+\bar{\sigma}_{\hat{R}-R}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})+\hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right)\right)\\ &\Rightarrow f^*\in G_t.\end{aligned}$$

■

Proof of Lemma 13 Let G_{t-1} be the final low-error set of Active-ILESS, and $\hat{S}$ be the final set of examples. The following inequalities are derived from Lemma 10 and inequalities

(56) and (57).

$$\begin{aligned}
R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(\hat{f}) &\leq \hat{R}(\hat{f},\hat{S}) + \hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right) \\
&\leq \hat{R}(f^*,\hat{S}) + \hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right) \\
&\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*) + \bar{\sigma}_{\hat{R}-R}\left(\frac{t}{2},\frac{\delta}{2t},d,R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*)\right) \\
&\quad +\hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right) \\
&\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*) + \bar{\sigma}_{\hat{R}-R}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S}) + \hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right)\right) \\
&\quad +\hat{\sigma}_{R-\hat{R}}\left(\frac{t}{2},\frac{\delta}{2t},d,\hat{R}(\hat{f},\hat{S})\right) \\
&\leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*) + \epsilon.
\end{aligned}$$

By Lemma 12 we know that f^* resides within G_{t-1}, which implies that any change in $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})$ in comparison to $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ reduces the true error of f^*. This also means that for every $f \in \mathcal{F}$,

$$R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f) - R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f) \leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f^*) - R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*),$$

which results in

$$R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(\hat{f}) \leq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}}(f^*) + \epsilon.$$

■

Proof of Lemma 14 The proof is similar to the proof of Lemma 4. We consider the last modification of G_t as a run of ILESS, under $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})$, with $m_0 \triangleq 2^{\lfloor log_2 m \rfloor - 1}$ examples and delta equal to $\frac{\delta}{4m_0}$.

Under event $\mathcal{K}$, the conditions of Lemma 4 hold, and by Lemma 12, $R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*) \leq R(f^*)$. We simply apply Lemma 4 with these parameters to get A' (A in Lemma 4).

$$A' = 4d\ln\left(\frac{16m_0e}{d\delta/4m_0}\right) = 4d\ln\left(\frac{64m_0^2e}{d\delta}\right).$$

The fact that $m/4 \leq m_0 \leq m/2$ completes the proof. ■

Proof of Lemma 15 We know by Lemma 14 that there exist constants C_1,C_2 that depend only on $\ln(\frac{1}{\delta})$ and d, and are independent of m, s.t.

$$\sigma_{Active} \leq C_1\frac{\ln m}{m} + C_2\sqrt{\frac{\ln m}{m}\cdot R(f^*)}.$$

We also know by the definition of Active-ILESS (Strategy 3), that it terminates when σ_{Active} is smaller than the given ϵ. We will find m large enough s.t.

$$C_1 \frac{\ln m}{m} \leq \epsilon/2, \tag{58}$$

$$C_2 \sqrt{\frac{\ln m}{m} \cdot R(f^*)} \leq \epsilon/2. \tag{59}$$

We assume that $\epsilon \leq 1/e$, as it is easy to find a proper m for $\epsilon > 1/e$. Starting with Equation (58), we want to show that $m = O(\frac{1}{\epsilon}\ln(\frac{1}{\epsilon}))$ satisfies it. Thus, we find k_1 s.t.

$$\begin{aligned} & C_1 \frac{\ln(k_1 \frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right))}{k_1 \frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right)} \leq \frac{\epsilon}{2} \\ \Leftrightarrow \quad & \frac{\ln\left(k_1 \frac{1}{\epsilon} \cdot \ln(\frac{1}{\epsilon})\right)}{\ln\left(\frac{1}{\epsilon}\right)} \leq \frac{k_1}{2C_1}. \end{aligned} \tag{60}$$

Bounding the left-hand side of Inequality (60) for $\epsilon \leq 1/e$ gives us,

$$\begin{aligned} \frac{\ln\left(k_1 \frac{1}{\epsilon} \cdot \ln(\frac{1}{\epsilon})\right)}{\ln(\frac{1}{\epsilon})} &\leq \frac{\ln\left(k_1 \frac{1}{\epsilon} \cdot \frac{1}{\epsilon}\right)}{\ln\left(\frac{1}{\epsilon}\right)} \\ &\leq 2 + \ln k_1. \end{aligned}$$

We need to find k_1 that will satisfy

$$2 + \ln k_1 \leq \frac{k_1}{2C_1}.$$

$k_1 = 16C_1^2$ will work for $C_1 \geq 1$; otherwise, we take $k_1 = 10$.

We use the same procedure to show that $m = O\left(\frac{R(f^*)}{\epsilon^2}\ln\left(\frac{R(f^*)}{\epsilon^2}\right)\right)$ satisfies Equation (59). We rewrite the equation in the following way:

$$\frac{\ln m}{m} \leq \frac{\epsilon^2}{4C_2^2 R(f^*)} \triangleq \epsilon_0.$$

We assume that $\epsilon_0 \leq 1/e$ ($m = 4$ holds otherwise) and find k_2 s.t.

$$\frac{\ln\left(k_2 \frac{1}{\epsilon_0} \ln\left(\frac{1}{\epsilon_0}\right)\right)}{k_2 \frac{1}{\epsilon_0} \ln\left(\frac{1}{\epsilon_0}\right)} \leq \epsilon_0.$$

As before, we reduce the problem to finding k_2 that satisfies

$$2 + \ln(k_2) \leq k_2.$$

$k_2 = 4$ suffices. We thus get that $m = O\left(\frac{1}{\epsilon_0^2}\ln\left(\frac{1}{\epsilon_0^2}\right)\right) = O\left(\frac{R(f^*)}{\epsilon^2}\ln\left(\frac{R(f^*)}{\epsilon^2}\right)\right)$ satisfies Equation (59). This implies that there exists a function

$$m(1/\epsilon, R(f^*)) = O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right) + \frac{R(f^*)}{\epsilon^2}\ln\left(\frac{R(f^*)}{\epsilon^2}\right)\right)$$

that bounds the total number of labels processed by Active-ILESS. ■

Appendix B. Proofs of Chernoff Bounds

Lemma 21 *Let $X_1, X_2, ..., X_n$ be independent Bernoulli trials with $Pr[X_i = 1] = p$, $X \triangleq \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}X$. Then, for every $\alpha \geq 0$:*

$$\Pr(X < (1-\alpha)\mu) \leq \exp(-\mu\alpha^2/2).$$

Proof This proof is taken from the work of John Canny Canny (2012).

For $t > 0$, we have

$$\Pr\left(X < (1-\alpha)\mu\right) = \Pr\left(\exp(-tX) > \exp(-t(1-\alpha)\mu)\right). \tag{61}$$

We use Markov's inequality. For a nonnegative random variable X, and $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

We apply the inequality for the right-hand side of Equation (61), to get

$$\Pr(X < (1-\alpha)\mu) \leq \frac{\mathbb{E}(\exp(-tX))}{\exp(-t(1-\alpha)\mu)}. \tag{62}$$

$X_1, X_2, ..., X_n$ are independent and thus

$$\mathbb{E}(\exp(-tX)) = \prod_{i=1}^{n} \mathbb{E}(\exp(-tX_i)).$$

For each X_i

$$\mathbb{E}(\exp(-tX_i)) = pe^{-t} + (1-p) = 1 - p(1-e^{-t}).$$

We use the fact that $1 - x < \exp(-x)$ for all x, with $x = p(1-e^{-t})$, to get

$$\mathbb{E}(\exp(-tX_i)) \leq \exp(-p(1-e^{-t})),$$

and conclude that

$$\begin{aligned}\mathbb{E}(\exp(-tX)) &= \prod_{i=1}^{n} \mathbb{E}\left(\exp(-tX_i)\right) \leq \prod_{i=1}^{n} \exp\left(-p(1-e^{-t})\right) \\ &= \exp\left(\sum_{i=1}^{n} p(e^{-t}-1)\right) = \exp\left(\mu(e^{-t}-1)\right).\end{aligned} \tag{63}$$

Going back to Equation (62), we have,

$$\Pr(X < (1-\alpha)\mu) \leq \frac{\exp\left(\mu(e^{-t}-1)\right)}{\exp\left(-t(1-\alpha)\mu\right)} = \exp\left(\mu(e^{-t}-1+t-t\alpha)\right). \tag{64}$$

We choose $t > 0$ to make the right-hand side of the equation as small as possible. After derivation, we get that the best t is $t = \ln\left(\frac{1}{1-\alpha}\right)$, and plugging it into Equation (64) gives

us,

$$\begin{aligned}\Pr(X < (1-\alpha)\mu) &\leq \exp\left(\mu(1-\alpha-1+\ln\left(\frac{1}{1-\alpha}\right)-\ln\left(\frac{1}{1-\alpha}\right)\alpha)\right)\\ &= \exp\left(\mu(-\alpha+\ln\left(\frac{1}{1-\alpha}\right)(1-\alpha))\right)\\ &= \left(\frac{e^{-\alpha}}{(1-\alpha)^{1-\alpha}}\right)^{\mu}. \end{aligned} \tag{65}$$

We now simplify this bound to get the desired result. We know that $(1-\alpha)^{1-\alpha} = e^{(1-\alpha)ln(1-\alpha)}$, and by Taylor expansion

$$ln(1-\alpha) = -\alpha - \frac{\alpha^2}{2} - \frac{\alpha^3}{3}...,$$

which multiplied by $(1-\alpha)$, gives us

$$(1-\alpha)ln(1-\alpha) = -\alpha + \frac{\alpha^2}{2} + \text{positive terms} > -\alpha + \frac{\alpha^2}{2}. \tag{66}$$

Plugging (66) into Equation (65), we finally get,

$$\begin{aligned}\Pr(X < (1-\alpha)\mu) &\leq \left(\frac{e^{-\alpha}}{(1-\alpha)^{1-\alpha}}\right)^{\mu}\\ &= \left(\frac{e^{-\alpha}}{e^{(1-\alpha)ln(1-\alpha)}}\right)^{\mu}\\ &\leq \left(\frac{e^{-\alpha}}{e^{-\alpha+\frac{\alpha^2}{2}}}\right)^{\mu}\\ &= e^{-\mu\alpha^2/2}\end{aligned}$$

■

Lemma 22 *Let $X_1, X_2, ..., X_n$ be independent Poisson trials with $Pr[X_i = 1] = p_i$, $X \triangleq \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}X$. Then, for every $\alpha \geq 2e - 1$:*

$$\Pr(X > (1+\alpha)\mu) \leq 2^{-\mu\alpha}.$$

Proof Sketch This sketch is taken from the work of John Canny Canny (2012). It is almost identical to the proof of Lemma 21.

We start by showing that

$$\Pr(X > (1+\alpha)\mu) \leq \left(\frac{e^{\alpha}}{(1+\alpha)^{1+\alpha}}\right)^{\mu}.$$

For every $t > 0$,

$$\Pr(X > (1+\alpha)\mu) \quad = \quad \Pr[\exp(tX) > \exp\left(t(1+\alpha)\mu\right)].$$

As we did in Lemma 21, we compute the Markov bound,

$$\Pr(X > (1+\alpha)\mu) \quad \leq \quad \frac{\mathbb{E}(\exp(tX))}{\exp(t(1+\alpha)\mu)},$$

and use the fact that X_i are independent, just like in (63), to get that

$$\mathbb{E}(\exp(tX)) \leq \exp\left(\mu(e^t - 1)\right).$$

Thus we get that

$$\Pr(X > (1+\alpha)\mu) \quad \leq \quad \frac{\exp(\mu(e^t-1))}{\exp(t(1+\alpha)\mu)} = \exp\left(\mu(e^t - 1 - t - \alpha t)\right).$$

From deviation, we choose $t = \ln(1+\alpha)$ to get

$$\Pr(X > (1+\alpha)\mu) \quad \leq \quad \left(\frac{e^\alpha}{(1+\alpha)^{1+\alpha}}\right)^\mu.$$

For $\alpha \geq 2e - 1$:

$$\Pr(X > (1+\alpha)\mu) \quad \leq \quad \left(\frac{e^\alpha}{(1+\alpha)^{1+\alpha}}\right)^\mu \leq \left(\frac{e^\alpha}{(2e)^{1+\alpha}}\right)^\mu \leq \left(\frac{e^\alpha}{(2e)^\alpha}\right)^\mu = 2^{-\mu\alpha}.$$

■

Appendix C. The Rejection Rate of Batch-ILESS

Lemma 23 *Given that event $\mathcal{K}$ (see Definition 6.1) occurred, the radius of Batch-ILESS, as defined in Strategy 3, Stage 4, satisfies*

$$\sigma_{Active} = O\left(\frac{B}{m} + \sqrt{\frac{B}{m} \cdot R(f^*)}\right),$$

where $B \triangleq 4d \ln\left(\frac{8m^2 e}{d\delta}\right)$.

Proof The proof is very similar to the proof of Lemma 14. ■

Theorem 24 *Let $\mathcal{F}$ be a hypothesis class with VC-dimension d, and $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ be an unknown probability distribution. Assume that event $\mathcal{K}$ (see Definition 6.1) occurred. Then, for all f^*, the abstain rate is bounded by*

$$1 - \Phi(\textit{Batch-ILESS}) \leq \theta_{f^*}(R_0) \cdot R_0,$$

where

$$R_0 \triangleq 2 \cdot R(f^*) + 44 \cdot \frac{B}{m} + 12 \cdot \sqrt{\frac{B}{m} \cdot R(f^*)}.$$

where $B \triangleq 4d \ln\left(\frac{8m^2 e}{d\delta}\right)$. *This immediately implies (by definition) that*

$$1 - \Phi(\textit{Batch-ILESS}) \leq \theta(R_0) \cdot R_0.$$

Proof Sketch The proof is very similar to the proof of Lemma 23. We observe the last modification of G_T, and notice that the change was made according to a run of ILESS, on the implied probability distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{T-1})$. Then we simply activate Theorem 5 with the relevant parameters plugged into it.

Note that by Lemma 12, all f^* of the original distribution reside within G_t for all t, and thus, by Lemma 11, they are all true risk minimizers of $\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{T-1})$. This also implies that $R(f^*) \geq R_{\mathcal{P}_{\mathcal{X},\mathcal{Y}}(G_{t-1})}(f^*)$ and thus can be used to bound Equation (30) of the original theorem that was proven for ILESS. θ_f is independent of $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ for all f, and thus the change of the labels does not affect it.

■

References

N. Ailon, R. Begleiter, and E. Ezra. Active learning using smooth relative regret approximations with applications. In 25th *Annual Conference on Learning Theory (COLT)*, 2012.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the* 23rd *International Conference on Machine Learning*, pages 65–72. ACM, 2006.

P.L. Bartlett and M.H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

P.L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 2004.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, 2003. ISBN 3-540-23122-6.

John Canny. Lecture 10, chernoff bounds. *Notes from CS174 offered at UC Berkeley*, 2012.

C.K. Chow. On optimum recognition error and reject trade-off. *IEEE Trans. on Information Theory*, 16:41–36, 1970.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2016a.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016b.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, pages 235–242, 2005.

S. Dasgupta, D. J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, pages 353–360, 2007.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *Neural Information Processing Systems (NIPS)*, pages 1665–1673, 2011.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13:255–279, 2012.

Ran El-Yaniv and Yair Wiener. On the version space compression set size and its applications. In *Measures of Complexity*, pages 341–357. Springer International Publishing, 2015.

Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

Alon Gonen, Sivan Sabato, and Shai Shalev-Shwartz. Efficient active learning of halfspaces: an aggressive approach. *Journal of Machine Learning Research*, 14(1):2583–2615, 2013.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24^{th} International Conference on Machine Learning (ICML)*, pages 353–360, 2007.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *The Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of Active Learning. `http://www.stevehanneke.com`, 2014a.

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014b. ISSN 1935-8237. doi: 10.1561/2200000037. URL `http://dx.doi.org/10.1561/2200000037`.

Steve Hanneke. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016.

Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions, 2017.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015.

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2755–2763. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5939-efficient-and-parsimonious-agnostic-active-learning.pdf`.

Mustafa A Kocak, Elza Erkip, and Dennis E Shasha. Conjugate conformal prediction for online binary classification. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 347–356. AUAI Press, 2016.

M. Kääriäinen. Active learning in the non-realizable case. *Lecture Notes in Computer Science, vol 4264. Springer, Berlin, Heidelberg*, 2006.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. An adaptive strategy for active learning with smooth decision boundary. *arXiv preprint arXiv:1711.09294*, 2017.

Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(Jan):67–90, 2012.

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

Vladimir N Vapnik and Alexey J Chervonenkis. Theory of pattern recognition, 1974.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Y. Wiener. *Theoretical Foundations of Selective Prediction*. PhD thesis, the Technion — Israel Institute of Technology, 2013.

Y. Wiener and R. El-Yaniv. Agnostic pointwise-competitive selective classification. Journal of AI Research, 52:171–201, 2015.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16:713–745, 2015. URL `http://jmlr.org/papers/v16/wiener15a.html`.

Yair Wiener and Ran El-Yaniv. Pointwise tracking the optimal regression function. In *Advances in Neural Information Processing Systems*, pages 2042–2050, 2012.

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(Jan):111–130, 2010.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.