Journal of Artificial Intelligence Research 52 (2015) 179-201 Submitted 5/14; published 1/15

Agnostic Pointwise-Competitive Selective Classification

Yair Wiener WYAIR@TX.TECHNION.AC.IL
Ran El-Yaniv RANI@CS.TECHNION.AC.IL
Computer Science Department
Technion – Israel Institute of Technology
Haifa 32000, Israel

Abstract

A *pointwise competitive* classifier from class $\mathcal{F}$ is required to classify identically to the best classifier in hindsight from $\mathcal{F}$. For noisy, agnostic settings we present a strategy for learning pointwise-competitive classifiers from a finite training sample provided that the classifier can abstain from prediction at a certain region of its choice. For some interesting hypothesis classes and families of distributions, the measure of this rejected region is shown to be diminishing at rate $\beta_1 \cdot O\left((\text{polylog}(m) \cdot \log(1/\delta)/m)^{\beta_2/2}\right)$, with high probability, where m is the sample size, δ is the standard confidence parameter, and β_1, β_2 are smoothness parameters of a Bernstein type condition of the associated excess loss class (related to $\mathcal{F}$ and the 0/1 loss). Exact implementation of the proposed learning strategy is dependent on an ERM oracle that is hard to compute in the agnostic case. We thus consider a heuristic approximation procedure that is based on SVMs, and show empirically that this algorithm consistently outperforms a traditional rejection mechanism based on distance from decision boundary.

1. Introduction

Given a labeled training set and a class of models $\mathcal{F}$, is it possible to select from $\mathcal{F}$, based on a finite training sample, a model whose predictions are always identical to best model in hindsight? While classical results from statistical learning theory surely preclude such a possibility within the standard model, when changing the rules of the game it is possible. Indeed, consider a game where our classifier is allowed to abstain from prediction without penalty in some region of its choice (a.k.a classification with a reject option). For this game, and assuming a noise free "realizable" setting, it was shown by El-Yaniv and Wiener (2010) that one can train a "perfect classifier" that never errs whenever it is willing to predict. While always abstaining will render such perfect classification vacuous, it was shown that for a quite broad set of problems (each specified by an underlying distribution family and a hypothesis class), perfect realizable classification is achievable with a rejection rate that diminishes quickly to zero with the training sample size.

In general, perfect classification cannot be achieved in a noisy setting. In this paper, our objective is to achieve *pointwise competitiveness*, a property ensuring that the prediction at every non-rejected test point is identical to the prediction of the best predictor in hindsight from the same class. Here we consider pointwise-competitive selective classification and generalize the results of El-Yaniv and Wiener (2010) to the agnostic case. In particular, we show that pointwise-competitive classification is achievable with high probability by a learning strategy called *low error selective strategy (LESS)*. Given a training sample S_m

and a hypothesis class $\mathcal{F}$, LESS outputs a pointwise-competitive selective classifier (f, g), where $f(x)$ is a standard classifier, and $g(x)$ is a selection function that qualifies some of the predictions as "don't knows" (see definitions in Section 2). The classifier f is simply taken to be the empirical risk minimizer (ERM) classifier, $\hat{f}$. Pointwise competitiveness is achieved through g as follows. Using standard concentration inequalities, we show that the true risk minimizer, f^*, achieves empirical error that is close that of $\hat{f}$. Thus, with high probability f^* belongs to the class of low empirical error hypotheses. Now all that is left to do is set $g(x)$ such that it allows the prediction of the label of x, as $\hat{f}(x)$, if and only if all the hypotheses in this low error class unanimously agree on the label of x. In the simpler, realizable setting (El-Yaniv & Wiener, 2010), this low error class simply reduces to the version space.

The bulk of our analysis (in Sections 3, 4 and 5) concerns coverage bounds for LESS, namely, showing that the measure of the region where the classifier (f, g) refuses to classify, diminishes quickly, with high probability, as the training sample size grows (see Section 2 for a formal definition). We provide several general and distribution-dependent coverage bounds. In particular, we show (in Corollaries 12 and 14, respectively) high probability bounds for the coverage $\Phi(f, g)$ of the classifier $(f, g,)$ of the form,

$$\Phi(f, g) \geq 1 - \beta_1 \cdot O\left((\text{polylog}(m) \cdot \log(1/\delta)/m)^{\beta_2/2}\right),$$

for linear models under (unknown) distribution $P(X, Y)$, where X are feature space points and Y are labels, whose marginal $P(X)$ is any finite mixture of Gaussians, and for axis aligned rectangles under $P(X, Y)$ whose marginal $P(X)$ is a product distribution, where β_1, β_2 are Bernstein class smoothness parameters depending on the hypothesis class and the underlying distribution (and the loss function, 0/1 in our case).

At the outset, efficient implementation of LESS seems to be out of reach as we are required to track the supremum of the empirical error over a possibly infinite hypothesis subset, which in general might be intractable. To overcome this computational difficulty, we propose a reduction of this problem to a problem of calculating (two) constrained ERMs. For any given test point x, we calculate the ERM over the training sample S_m with a constraint on the label of x (one positive label constraint and one negative). We show that thresholding the difference in empirical error between these two constrained ERMs is equivalent to tracking the supremum over the entire (infinite) hypothesis subset. Based on this reduction we introduce in Section 6 a "disbelief principle" that motivates a heuristic implementation of LESS, which relies on constrained SVMs, and mimics the optimal behavior.

In Section 7 we present some numerical examples over medical classification problems and examine the empirical performance of the new algorithm and compare its performance with that of the widely used selective classification method for rejection, based on distance from decision boundary.

2. Pointwise-Competitive Selective Classification: Preliminary Definitions

Let $\mathcal{X}$ be some feature space, for example, d-dimensional vectors in $\mathbb{R}^d$, and $\mathcal{Y}$ be some output space. In standard classification, the goal is to learn a classifier $f : \mathcal{X} \to \mathcal{Y}$, using a finite training sample of m labeled examples, $S_m = \{(x_i, y_i)\}_{i=1}^m$, assumed to be sampled i.i.d. from some *unknown* underlying distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. The classifier is to be selected from some hypothesis class $\mathcal{F}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a bounded loss function.

In *selective classification* (El-Yaniv & 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 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."

The general performance of a selective predictor is characterized in terms of two quantities: *coverage* and *risk*. The *coverage* of (f, g) is $\Phi(f, g) \triangleq \mathbb{E}_P[g(x)]$. The true risk of (f, g), with respect to some loss function ℓ, is the average loss of f restricted to its region of activity as qualified by g, and normalized by its coverage, $R(f, g) \triangleq \mathbb{E}_P[\ell(f(x), y) \cdot g(x)] / \Phi(f, g)$. It is easy to verify that if $g \equiv 1$ (and therefore $\Phi(f, g) = 1$), then $R(f, g)$ reduces to the familiar risk functional $R(f) \triangleq \mathbb{E}_P[\ell(f(x), y)]$. For a classifier f, let $\hat{R}(f) \triangleq \frac{1}{m}\sum_{i=1}^m \ell(f(x_i), y_i)$, the standard empirical error of f over the sample S_m. Let $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f)$ be the empirical risk minimizer (ERM), and let $f^* = \operatorname{argmin}_{f \in \mathcal{F}} R(f)$ be the true risk minimizer with respect to unknown distribution $P(X, Y)$.[1] Clearly, the true risk minimizer f^* is unknown. A selective classifier (f, g) is called *pointwise-competitive* if for any $x \in \mathcal{X}$, for which $g(x) > 0$, $f(x) = f^*(x)$.

3. Low Error Selective Strategy (LESS)

For any hypothesis class $\mathcal{F}$, hypothesis $f \in \mathcal{F}$, distribution P, sample S_m, and real number $r > 0$, define the true and empirical *low-error sets*,

$$\mathcal{V}(f, r) \triangleq \left\{f' \in \mathcal{F} : R(f') \le R(f) + r\right\} \tag{1}$$

and

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

Throughout the paper we denote by $\sigma(m, \delta, d)$ the slack of a standard uniform deviation bound, given in terms of the training sample size, m, the confidence parameter, δ, and the VC-dimension, d, of the class $\mathcal{F}$,

$$\sigma(m, \delta, d) \triangleq 2\sqrt{\frac{2d\left(\ln \frac{2me}{d}\right) + \ln \frac{2}{\delta}}{m}}. \tag{3}$$

The following theorem is a slight extension of the statement made by Bousquet, Boucheron, and Lugosi (2004, p. 184).

Theorem 1 (Bousquet et al., 2004)**.** *Let ℓ be the 0/1 loss function and $\mathcal{F}$, a hypothesis class whose VC-dimension is d. For any $0 < \delta < 1$, with probability of at least $1 - \delta$ over the choice of S_m from P^m, any hypothesis $f \in \mathcal{F}$ satisfies*

$$R(f) \le \hat{R}(f) + \sigma(m, \delta, d).$$

Similarly, $\hat{R}(f) \le R(f) + \sigma(m, \delta, d)$ under the same conditions.

[1] More formally, f^* is a classifier such that $R(f^*) = \inf_{f \in \mathcal{F}} R(f)$ and $\inf_{f \in \mathcal{F}} P((x, y) : f(x) \ne f^*(x)) = 0$. The existence of such a (measurable) f^* is guaranteed under sufficient considerations (see Hanneke, 2012, pp. 1511-2).

Remark 2. *The use of Theorem 1 and, in particular, VC bounds for classification problems (0/1 loss) is not mandatory for developing the theory presented in this paper. Similar theories can be developed using other types of bounds (e.g., Rademacher or compression bounds) for other learning problems.*

Let $G \subseteq \mathcal{F}$. The *disagreement set* (Hanneke, 2007a; El-Yaniv & Wiener, 2010) w.r.t. G is defined as

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

Let us now motivate the low-error selective strategy (LESS) whose pseudo-code appears in Strategy 1. The strategy is define whenever the empirical risk minimizer (ERM) exists, for example, in the case of the 0/1 loss. Using a standard uniform deviation bound, such as the one in Theorem 1, one can show that the training error of the true risk minimizer, f^*, cannot be "too far" from the training error of the empirical risk minimizer, $\hat{f}$. Therefore, we can guarantee, with high probability, that the empirical low error class $\hat{\mathcal{V}}\left(\hat{f}, r\right)$ (applied with appropriately chosen r) includes the true risk minimizer f^*. The selection function g is now constructed to accept a subset of the domain $\mathcal{X}$, on which all hypotheses in the empirical low-error set unanimously agree. Strategy 1 formulates this idea. We call it a 'strategy' rather then an 'algorithm' because it lacks implementation details. Indeed, it is not clear at the outset if this strategy can be implemented.

Strategy 1 Agnostic low-error selective strategy (LESS)

Input: S_m, m, δ, d

Output: a pointwise-competitive selective classifier (h, g) w.p. $1 - \delta$

1: Set $\hat{f} = ERM(\mathcal{F}, S_m)$, i.e., $\hat{f}$ is any empirical risk minimizer from $\mathcal{F}$ w.r.t. S_m

2: Set $G = \hat{\mathcal{V}}\left(\hat{f}, 2\sigma(m, \delta/4, d)\right)$ (see Eq. (2) and (3))

3: Construct g such that $g(x) = 1 \iff x \in \{\mathcal{X} \setminus DIS\,(G)\}$

4: $f = \hat{f}$

We now begin the analysis of LESS. The following lemma establishes its pointwise competitiveness. In Section 4 we develop general coverage bounds in terms of an undetermined disagreement coefficient. Then, in Section 5 we present distribution-dependent bounds that do not rely on the disagreement coefficient.

Lemma 3 (pointwise competitiveness)**.** *Let ℓ be the 0/1 loss function and $\mathcal{F}$, a hypothesis class whose VC-dimension is d. Let $\delta > 0$ be given and let (f, g) be the selective classifier chosen by LESS. Then, with probability of at least* $1 - \delta/2$, (f, g) *is a* pointwise competitive *selective classifier.*

Proof. By Theorem 1, with probability of at least $1 - \delta/4$,

$$\hat{R}(f^*) \leq R(f^*) + \sigma\left(m, \delta/4, d\right).$$

Clearly, since f^* minimizes the true error, $R(f^*) \leq R(\hat{f})$. Applying again Theorem 1, we know that with probability of at least $1 - \delta/4$,

$$R(\hat{f}) \leq \hat{R}(\hat{f}) + \sigma\left(m, \delta/4, d\right).$$

Using the union bound, it follows that with probability of at least $1 - \delta/2$,

$$\hat{R}(f^*) \leq \hat{R}(\hat{f}) + 2\sigma\left(m, \delta/4, d\right).$$

Hence, with probability of at least $1 - \delta/2$,

$$f^* \in \hat{\mathcal{V}}\left(\hat{f}, 2\sigma\left(m, \delta/4, d\right)\right) \triangleq G.$$

By definition, LESS constructs the selection function $g(x)$ such that it equals one iff $x \in \mathcal{X} \setminus DIS\left(G\right)$. Thus, for any $x \in \mathcal{X}$, for which $g(x) = 1$, all the hypotheses in G agree, and in particular f^* and $\hat{f}$ agree. Therefore (f, g) is pointwise-competitive with high probability. □

4. General Coverage Bounds for LESS in Terms of the Disagreement Coefficient

We require the following definitions to facilitate the coverage analysis. For any $f \in \mathcal{F}$ and $r > 0$, define the set $B(f, r)$ of all hypotheses that reside in a ball of radius r around f,

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

For any $G \subseteq \mathcal{F}$, and distribution P, we denote by ΔG the volume of the disagreement set of G (see (4)), $\Delta G \triangleq \Pr\left\{DIS(G)\right\}$. Let $r_0 \geq 0$. The *disagreement coefficient* (Hanneke, 2009) of the hypothesis class $\mathcal{F}$ with respect to the target distribution P is

$$\theta(r_0) \triangleq \theta_{f^*}(r_0) = \sup_{r > r_0} \frac{\Delta B(f^*, r)}{r}. \tag{5}$$

The disagreement coefficient will be utilized later on in our analysis. See also a discussion on its characteristics after Corollary 7. The associated *excess loss class* of the class $\mathcal{F}$ and the loss function ℓ (Massart, 2000; Mendelson, 2002; Bartlett, Mendelson, & Philips, 2004) is defined as

$$XL(\mathcal{F}, \ell)(x, y) \triangleq \left\{\ell(f(x), y) - \ell(f^*(x), y) : f \in \mathcal{F}\right\}.$$

Whenever $\mathcal{F}$ and ℓ are fixed we abbreviate $XL = XL(\mathcal{F}, \ell)(x, y)$. XL is said to be a (β_1, β_2)-*Bernstein class* with respect to P (where $0 < \beta_2 \leq 1$ and $\beta_1 \geq 1$), if every $h \in XL$ satisfies

$$\mathbb{E}h^2 \leq \beta_1 (\mathbb{E}h)^{\beta_2}. \tag{6}$$

Bernstein classes arise in many natural situations (see, e.g., Koltchinskii, 2006; Bartlett & Mendelson, 2006; Bartlett & Wegkamp, 2008). For example, if the conditional probability $P(Y|X)$ is bounded away from 1/2, or it satisfies Tsybakov's noise conditions[2], then the excess loss function is a Bernstein class (Bartlett & Mendelson, 2006; Tsybakov, 2004).[3]

[2] If the data was generated from *any* unknown deterministic hypothesis with limited noise then $P(Y|X)$ is bounded away from 1/2.

[3] Specifically, for the 0/1 loss, Assumption A in Proposition 1 in the work of Tsybakov (2004), is equivalent to the Bernstein class condition of Equation (6) above with $\beta_2 = \frac{\alpha}{1+\alpha}$, where α is the Tsybakov noise parameter.

In the following sequence of lemmas and theorems we assume a binary hypothesis class $\mathcal{F}$ with VC-dimension d, an underlying distribution P over $\mathcal{X} \times \{\pm 1\}$, and that ℓ is the 0/1 loss function. Also, XL denotes the associated excess loss class. Our results can be extended to loss functions other than 0/1 by similar techniques to those used by Beygelzimer, Dasgupta, and Langford (2009).

In Figure 1 we schematically depict the hypothesis class $\mathcal{F}$ (the gray area), the target hypothesis (filled black circle outside $\mathcal{F}$), and the best hypothesis in the class f^*. The distance of two points in the diagram relates to the distance between two hypothesis under the marginal distribution $P(X)$. Our first observation is that if the excess loss class is (β_1, β_2)-Bernstein class, then the set of low true error (depicted in Figure 1 (a)) resides within a larger ball centered around f^* (see Figure 1 (b)).

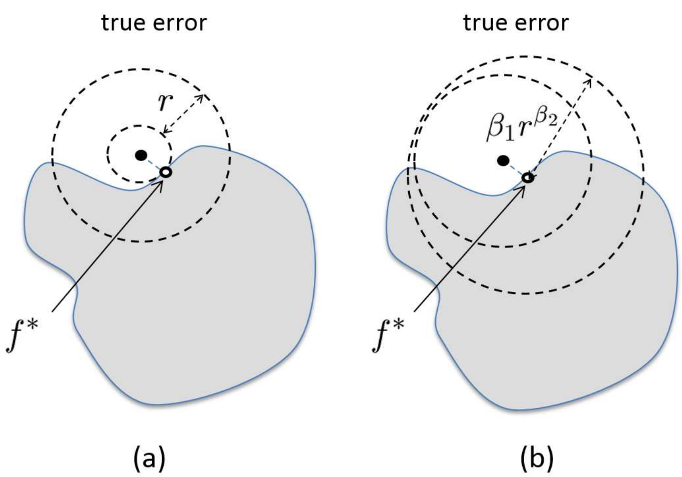

Figure 1: The set of low true error (a) resides within a ball around f^* (b).

Lemma 4. *If XL is a (β_1, β_2)-Bernstein class with respect to P, then for any $r > 0$*

$$\mathcal{V}(f^*, r) \subseteq B\left(f^*, \beta_1 r^{\beta_2}\right).$$

Proof. If $f \in \mathcal{V}(f^*, r)$ then, by definition, $\mathbb{E}\{\mathbb{I}(f(X) \neq Y)\} \leq \mathbb{E}\{\mathbb{I}(f^*(X) \neq Y)\} + r$. By linearity of expectation we have,

$$\mathbb{E}\{\mathbb{I}(f(X) \neq Y) - \mathbb{I}(f^*(X) \neq Y)\} \leq r. \tag{7}$$

Since XL is (β_1, β_2)-Bernstein,

$$\begin{aligned}
\mathbb{E}\{\mathbb{I}(f(X) \neq f^*(X))\} &= \mathbb{E}\{|\mathbb{I}(f(X) \neq Y) - \mathbb{I}(f^*(X) \neq Y)|\} \\
&= \mathbb{E}\left\{(\ell(f(X), Y) - \ell(f^*(X), Y))^2\right\} \triangleq \mathbb{E}h^2 \leq \beta_1 (\mathbb{E}h)^{\beta_2} \\
&\triangleq \beta_1 \left(\mathbb{E}\{\mathbb{I}(f(X) \neq Y) - \mathbb{I}(f^*(X) \neq Y)\}\right)^{\beta_2}.
\end{aligned}$$

By (7), $\mathbb{E}\{\mathbb{I}(f(X) \neq f^*(X))\} \leq \beta_1 r^{\beta_2}$. Therefore, by definition, $f \in B\left(f^*, \beta_1 r^{\beta_2}\right)$. $\square$

So far we have seen that the set of low true error resides within a ball around f^*. Now we would like to prove that with high probability the set of low empirical error (depicted in Figure 2 (a)) resides within the set of low true error (see Figure 2 (b)). We emphasize that the distance between hypotheses in Figure 2 (a) is based on the empirical error, while the distance in Figure 2 (b) is based on the true error.

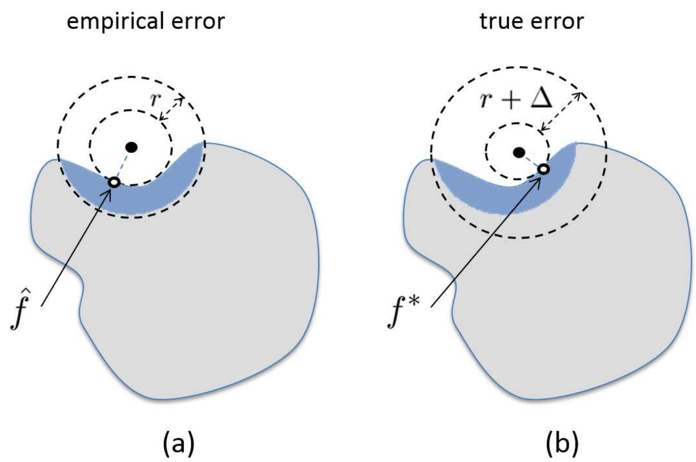

Figure 2: The set of low empirical error (a) resides within the set of low true error (b).

Lemma 5. *For any $r > 0$, and $0 < \delta < 1$, with probability of at least $1 - \delta$,*

$$\hat{\mathcal{V}}(\hat{f}, r) \subseteq \mathcal{V}\left(f^*, 2\sigma\left(m, \delta/2, d\right) + r\right).$$

Proof. If $f \in \hat{\mathcal{V}}(\hat{f}, r)$, then, by definition, $\hat{R}(f) \leq \hat{R}(\hat{f}) + r$. Since $\hat{f}$ minimizes the empirical error, we know that $\hat{R}(\hat{f}) \leq \hat{R}(f^*)$. Using Theorem 1 twice, and applying the union bound, we see that with probability of at least $1 - \delta$,

$$R(f) \leq \hat{R}(f) + \sigma(m, \delta/2, d) \quad \wedge \quad \hat{R}(f^*) \leq R(f^*) + \sigma(m, \delta/2, d).$$

Therefore,

$$R(f) \leq R(f^*) + 2\sigma\left(m, \delta/2, d\right) + r,$$

and

$$f \in \mathcal{V}\left(f^*, 2\sigma\left(m, \delta/2, d\right) + r\right).$$

□

We have shown that, with high probability, the set of low empirical error is a subset of a certain ball around f^*. Therefore, the probability that at least two hypotheses in the set of low empirical error will disagree with each other is bounded by the probability that at least two hypotheses in that ball around f^* will disagree with each other. Fortunately, the latter is bounded by the disagreement coefficient as established in the following lemma.

Lemma 6. *For any $r > 0$ and $0 < \delta < 1$, with probability of at least $1 - \delta$,*

$$\Delta\hat{\mathcal{V}}(\hat{f}, r) \leq \beta_1 \cdot (2\sigma(m, \delta/2, d) + r)^{\beta_2} \cdot \theta(r_0),$$

where $\theta(r_0)$ is the disagreement coefficient of $\mathcal{F}$ with respect to P, applied with $r_0 = (2\sigma(m, \delta/2, d))^{\beta_2}$ (see (5)).

Proof. Applying Lemmas 5 and 4 we get that with probability of at least $1 - \delta$,

$$\hat{\mathcal{V}}(\hat{f}, r) \subseteq B\left(f^*, \beta_1 (2\sigma(m, \delta/2, d) + r)^{\beta_2}\right).$$

Therefore,

$$\Delta\hat{\mathcal{V}}(\hat{f}, r) \leq \Delta B\left(f^*, \beta_1 (2\sigma(m, \delta/2, d) + r)^{\beta_2}\right).$$

By the definition of the disagreement coefficient (5), for any $r' > r_0$, $\Delta B(f^*, r') \leq \theta(r_0)r'$. Recalling that $\beta_1 \geq 1$ and thus observing that $r' = \beta_1 (2\sigma(m, \delta/2, d) + r)^{\beta_2} > (2\sigma(m, \delta/2, d))^{\beta_2} = r_0$, the proof is complete. $\square$

We are now in a position to state our first coverage bound for the selective classifier constructed by LESS. This bound is given in terms of the disagreement coefficient.

Corollary 7. *Let $\mathcal{F}$ be a hypothesis class as in Theorem 1, and assume that XL is a (β_1, β_2)-Bernstein class w.r.t. P. Let (f, g) be the selective classifier constructed by LESS. Then, with probability of at least $1 - \delta$, (f, g) is a* pointwise competitive *selective classifier and*

$$\Phi(f, g) \geq 1 - \beta_1 \cdot (4\sigma(m, \delta/4, d))^{\beta_2} \cdot \theta(r_0),$$

where $\theta(r_0)$ is the disagreement coefficient of $\mathcal{F}$ with respect to P, and $r_0 = (2\sigma(m, \delta/4, d))^{\beta_2}$.

Proof. By Lemma 3, with probability of at least $1 - \delta/2$, (f, g) is pointwise-competitive. Set $G \triangleq \hat{\mathcal{V}}\left(\hat{f}, 2\sigma(m, \delta/4, d)\right)$. By construction, $f = \hat{f}$, and the selection function $g(x)$ equals one iff $x \in \mathcal{X} \setminus DIS(G)$. Thus, by the definition of coverage, $\Phi(f, g) = \mathbb{E}\{g(X)\} = 1 - \Delta G$. Therefore, applications of Lemma 6 and the union bound imply that with probability of at least $1 - \delta$, (f, g) is pointwise-competitive and its coverage satisfies,

$$\Phi(f, g) = \mathbb{E}\{g(X)\} = 1 - \Delta G \geq 1 - \beta_1 \cdot (4\sigma(m, \delta/4, d))^{\beta_2} \cdot \theta(r_0),$$

$\square$

Noting that $\theta(r)$ is monotone non-increasing with r, we know that the coverage bound of Corollary 7 clearly applies with $\theta(0)$. The quantity $\theta(0)$ has been discussed in numerous papers and has been shown to be finite in various settings including thresholds in $\mathbb{R}$ under any distribution ($\theta(0) = 2$) (Hanneke, 2009), linear separators through the origin in $\mathbb{R}^d$ under uniform distribution on the sphere ($\theta(0) \leq \sqrt{d}$) (Hanneke, 2009), and linear separators in $\mathbb{R}^d$ under smooth data distribution bounded away from zero ($\theta(0) \leq c(f^*)d$, where $c(f^*)$ is an unknown constant that depends on the target hypothesis) (Friedman, 2009). For these cases, an application of Corollary 7 is sufficient to guarantee pointwise-competitiveness with bounded coverage that converges to one. Unfortunately for many hypothesis classes and distributions the disagreement coefficient $\theta(0)$ is infinite (Hanneke, 2009). Fortunately, if the disagreement coefficient $\theta(r)$ grows slowly with respect to $1/r$ (as shown in Wang, 2011, under sufficient smoothness conditions), Corollary 7 is sufficient to guarantee bounded coverage.

5. More Distribution-Dependent Coverage Bounds for LESS

In this section we establish distribution-dependent coverage bounds for LESS. The starting point of these bounds is the following corollary.

Corollary 8. *Let $\mathcal{F}$ be a hypothesis class as in Theorem 1, and assume that $\mathcal{F}$ has disagreement coefficient*

$$\theta(r_0) = O\left(\text{polylog}\left(1/r_0\right)\right) \tag{8}$$

w.r.t. distribution P, and that XL is a (β_1, β_2)-Bernstein class w.r.t. the same distribution. Let (f, g) be the selective classifier chosen by LESS. Then, with probability of at least $1-\delta$, (f, g) is pointwise competitive *and its coverage satisfies,*

$$\Phi(f,g) \geq 1 - \beta_1 \cdot O\left(\left(\frac{\text{polylog}(m)}{m} \cdot \log\frac{1}{\delta}\right)^{\beta_2/2}\right).$$

Proof. Plugging in (8) in the coverage bound of Corollary 7 immediately yields the result. □

Corollary 8 states fast coverage bounds for LESS in cases where the disagreement coefficient grows slowly with respect to $1/r_0$.[4] Recent results on disagreement-based active learning and selective prediction (Wiener et al., 2014; Wiener, 2013) established tight relations between the disagreement coefficient and an empirical quantity called the *version space compression set size*. This quantity has been analyzed by El-Yaniv and Wiener (2010) in the context of realizable selective classification, and there are known distribution-dependent bounds for it. Our plan for the rest of this section is to introduce the version space compression set size, discuss its relation to the disagreement coefficient, and then show how to apply those results in the agnostic setting.

While we are interested in solving the agnostic case, we will now consider for a moment the realizable setting and utilize known results that will be used in our analysis. Specifically, we now assume that $\exists f^* \in \mathcal{F}$ with $\mathbb{P}(Y = f^*(x)|X = x) = 1$ for all $x \in \mathcal{X}$, where $(X, Y) \sim P$. Given a training sample S_m, let $\text{VS}_{\mathcal{F},S}$ be the induced *version space*, i.e., the set of all hypotheses consistent with the given sample S_m. The *version space compression set size*, denoted $\hat{n}(S_m) = \hat{n}(\mathcal{F}, S_m)$, is defined to be the size of the smallest subset of S_m inducing the same version space (Hanneke, 2007b; El-Yaniv & Wiener, 2010). Being a function of S_m, clearly $\hat{n}(S_m)$ is a random variable, and for any specific realization S_m its value is unique.

For any m and $\delta \in (0, 1]$, define the *version space compression set size minimal bound* as

$$\mathcal{B}_{\hat{n}}(m,\delta) \triangleq \min\left\{b \in \mathbb{N} : \mathbb{P}(\hat{n}(S_m) \leq b) \geq 1 - \delta\right\}. \tag{9}$$

We rely on the following lemma (Wiener et al., 2014). For the sake of self-containment we provide its proof in the appendix.

---

[4] When the disagreement coefficient does not grow ploy-logarithmically with $1/r_0$ but is still $o(1/r_0)$, it is still possible to prove a lower bound on the coverage. Specifically, if $\theta(r_0) = O\left((1/r_0)^{\alpha}\right)$ with $\alpha < 1$, one can show that $\Phi(f, g) \geq 1 - O(1/(\sqrt{m})^{\beta_2(1-\alpha)})$.

Lemma 9 (Wiener et al., 2014)**.** *In the realizable case, if* $\mathcal{B}_{\hat{n}}(m,\delta) = O\left(\text{polylog}(m)\log(1/\delta)\right)$, *or* $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O\left(\text{polylog}(m)\right)$, *then* $\theta(r_0) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right)$.

Obviously, the statement of Lemma 9 only holds (and is well defined) within a realizable setting (the version space compression set size is only defined for this setting). We now turn back to the agnostic setting and consider an arbitrary underlying distribution P over $\mathcal{X}\times\mathcal{Y}$.

Recall that in the agnostic setting, we let $f^* : \mathcal{X} \to \mathcal{Y}$ denote a (measurable) classifier such that $R(f^*) = \inf_{f\in\mathcal{F}} R(f)$ and $\inf_{f\in\mathcal{F}} P((x,y) : f(x) \neq f^*(x)) = 0$, which is guaranteed to exist under sufficient assumptions (see Hanneke, 2012, Section 6.1); We call f^* an *infimal (best) hypothesis* (of $\mathcal{F}$, w.r.t. P). Clearly there can be several different infimal hypotheses. We note, however, that if XL is a (β_1,β_2)-Bernstein class with respect to P (as we assume in this paper), then Lemma 4 ensures that all infimal hypotheses are identical up to measure zero.

The definitions of version space and version space compression set size can be naturally generalized to the agnostic setting with respect to an infimal hypothesis (Wiener et al., 2014) as follows. Let f^* be an infimal hypothesis of $\mathcal{F}$ w.r.t. P. The *agnostic version space* of S_m is

$$\text{VS}_{\mathcal{F},S_m,f^*} \triangleq \{f \in \mathcal{F} : \forall(x,y) \in S_m, f(x) = f^*(x)\}.$$

The *agnostic version space compression set size*, denoted $\hat{n}(S_m) = \hat{n}(\mathcal{F}, S_m, f^*)$, is defined to be the size of the smallest subset of S_m inducing the agnostic version space $\text{VS}_{\mathcal{F},S_m,f^*}$. Finally, extend also the definition of the version space compression set minimal bound to the agnostic setting as follows.

$$\mathcal{B}_{\hat{n}}(m,\delta,f^*) \triangleq \min\{b \in \mathbb{N} : \mathbb{P}(\hat{n}(\mathcal{F},S_m,f^*) \leq b) \geq 1-\delta\}.$$

The key observation that allows for surprisingly easy utilization of Lemma 9 in the agnostic setting is that the disagreement coefficient depends only on the hypothesis class $\mathcal{F}$ and the marginal distribution $P(X)$. Using an infimal hypothesis f^* we can therefore take any agnostic learning problem and consider its realizable "projection," whereby points are labeled by f^* and it has the same marginal distribution $P(X)$. These two problems will have (essentially) the same disagreement coefficients. This idea was initially observed by Hanneke (2013) and Wiener (2013). Here we formulate it as a slight variation of the formulation in the work of Wiener, Hanneke, and El-Yaniv (2014).

We define the disagreement in the agnostic setting as in (5) with respect to an infimal hypothesis f^*. For any agnostic learning problem $(\mathcal{F}, P)$ we define its *realizable projection* $(\mathcal{F}', P')$ as follows. Let $\mathcal{F}' \triangleq \mathcal{F} \cup \{f^*\}$ where f^* is an infimal hypothesis of the agnostic problem. Define P' to be a distribution with marginal $P'(X) = P(X)$, and $\mathbb{P}(Y = f^*(x)|X = x) = 1$ for all $x \in \mathcal{X}$. It is easy to verify that $(\mathcal{F}', P')$ is a realizable learning problem, i.e., $\exists f^* \in \mathcal{F}'$ with $\mathbb{P}_{P'(X,Y)}(Y = f^*(x)|X = x) = 1$ for all $x \in \mathcal{X}$.

Lemma 10 (Realizable projection)**.** *Given any agnostic learning problem,* $(\mathcal{F}, P)$*, let* $(\mathcal{F}', P')$ *be its realizable projection. Let* $\theta(r_0)$ *and* $\theta'(r_0)$ *be the associated disagreement coefficients of the agnostic and realizable projection problems, respectively. Then,* $\theta(r_0) \leq \theta'(r_0)$.

Proof. First note that θ and θ' depend, respectively, on P and P' only via f^* and the marginal distributions $P(X) = P'(X)$. Since $\mathcal{F} \subseteq \mathcal{F} \cup \{f^*\} = \mathcal{F}'$, we readily get that $\theta(r_0) \leq \theta'(r_0)$. □

Let us summarize the above derivation. Given an agnostic problem $(\mathcal{F}, P)$, consider its realizable projection $(\mathcal{F}', P')$. If $\mathcal{B}_{\hat{n}}(m, \delta) = O\left(\text{polylog}(m) \log\left(1/\delta\right)\right)$ (or $\mathcal{B}_{\hat{n}}(m, 1/20) = O\left(\text{polylog}(m)\right)$) for the realizable problem, then by Lemma 9, $\theta(r_0) = O\left(\text{polylog}\left(1/r_0\right)\right)$, which, by Lemma 10, also holds in the original agnostic problem. Therefore, Corollary 7 applies and we obtain a fast coverage bound for LESS w.r.t. $(\mathcal{F}, P)$.

New agnostic coverage bounds for LESS are obtained using the following known bounds for the (realizable) version space compression set size. The first one, by El-Yaniv and Wiener (2010), applies to the problem of learning linear separators under a mixture of Gaussian distributions. The following theorem is a direct application of Lemma 32 in the work of El-Yaniv and Wiener (2010).

Theorem 11 (El-Yaniv & Wiener, 2010)**.** *For any $d, n \in \mathbb{N}$, let $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{F}$ be the space of linear separators on $\mathbb{R}^d$, and P be any distribution with marginal over $\mathbb{R}^d$ that is a mixture of n multivariate normal distributions. Then, there is a constant $c_{d,n} > 0$ (depending on d, n, but otherwise independent of P) such that $\forall m \geq 2$,*

$$\mathcal{B}_{\hat{n}}(m, 1/20) \leq c_{d,n} (\log(m))^{d-1}.$$

Applying Theorem 11, together with Lemma 10, Lemma 9 and Corollary 8, immediately yields the following result.

Corollary 12. *Assume the conditions of Theorem 11. Assume also that XL is a (β_1, β_2)-Bernstein class w.r.t. $P(X, Y)$. Let (f, g) be the selective classifier constructed by LESS. Then, with probability of at least $1 - \delta$, (f, g) is a* pointwise competitive *selective classifier and*

$$\Phi(f, g) \geq 1 - \beta_1 \cdot O\left((\text{polylog}(m) \cdot \log(1/\delta)/m)^{\beta_2/2} \right).$$

The second version space compression set size bound concerns realizable learning of axis-aligned rectangles under product densities over $\mathbb{R}^n$. Such bounds have been previously proposed by Wiener, Hanneke, and El-Yaniv (2014) and El-Yaniv and Wiener (2010, 2012). We now state (without proof) a recent bound (Wiener, Hanneke, & El-Yaniv, 2014) giving version space compression set size bound for this learning problem (whose positive class is bounded away from zero).

Theorem 13 (Wiener et al., 2014)**.** *For $d, m \in \mathbb{N}$ and $\lambda, \delta \in (0, 1)$, let $\mathcal{X} \subseteq \mathbb{R}^d$. For any P with marginal distribution over $\mathbb{R}^d$ that is a product of densities over $\mathbb{R}^d$ with marginals having continuous CDFs, and for $\mathcal{F}$ the space of axis-aligned rectangles f on $\mathbb{R}^d$ with $P((x, y) : f(x) = 1) \geq \lambda$,*

$$\mathcal{B}_{\hat{n}}(m, \delta) \leq \frac{8d}{\lambda} \ln\left(\frac{8d}{\delta}\right).$$

Here again, an application of Theorem 13, together with Lemma 10, Lemma 9 and Corollary 8 yields the following corollary.

Corollary 14. *For $d, m \in \mathbb{N}$ and $\lambda, \delta \in (0, 1)$, let $\mathcal{X} \subseteq \mathbb{R}^d$. Let $P(X, Y)$ be an underlying distribution with marginal $P(X)$ that is a product of densities over $\mathbb{R}^d$ with marginals having continuous CDFs. Let $\mathcal{F}$ the space of axis-aligned rectangles f on $\mathbb{R}^d$ with $P((x, y) : f(x) = 1) \geq \lambda$, Assume that XL is a (β_1, β_2)-Bernstein class w.r.t. $P(X, Y)$. Let (f, g) be the* *selective classifier constructed by LESS. Then, with probability of at least* $1-\delta$, (f,g) *is a* pointwise competitive *selective classifier and*

$$\Phi(f,g) \geq 1 - \beta_1 \cdot O\left((\text{polylog}(m) \cdot \log(1/\delta)/m)^{\beta_2/2}\right).$$

6. ERM Oracles and the Disbelief principle

At the outset, efficient construction of the selection function g prescribed by LESS seems to be out of reach as we are required to verify, for each point x in question, whether all hypotheses in the low error class agree on its label. Moreover, g should be computed for the entire domain. Luckily, it is possible to compute g in a "lazy" manner and we now show how to compute $g(x)$ by calculating (two) constrained ERMs. For any given test point x, we calculate the ERM over the training sample S_m with a constraint on the label of x (one positive label constraint and one negative). We show that thresholding the difference in empirical error between these two constrained ERMs is equivalent to tracking the supremum over the entire (infinite) hypothesis subset. The following lemma establishes this reduction.

Lemma 15. *Let* (f,g) *be a selective classifier chosen by LESS after observing the training sample* S_m*. Let* $\hat{f}$ *be an empirical risk minimizer over* S_m*. Let* x *be any point in* $\mathcal{X}$ *and define*

$$\tilde{f}_x \triangleq \operatorname*{argmin}_{f\in\mathcal{F}} \left\{\hat{R}(f) \quad | \quad f(x) = -sign\left(\hat{f}(x)\right)\right\},$$

i.e., an empirical risk minimizer forced to label x *the opposite from* $\hat{f}(x)$*. Then*

$$g(x) = 0 \quad \Longleftrightarrow \quad \hat{R}(\tilde{f}_x) - \hat{R}(\hat{f}) \leq 2\sigma\left(m, \delta/4, d\right). \tag{10}$$

Proof. First note that according to the definition of $\hat{\mathcal{V}}$ (see Eq (2)),

$$\hat{R}(\tilde{f}_x) - \hat{R}(\hat{f}) \leq 2\sigma\left(m, \delta/4, d\right) \quad \Longleftrightarrow \quad \tilde{f}_x \in \hat{\mathcal{V}}\left(\hat{f}, 2\sigma\left(m, \delta/4, d\right)\right). \tag{11}$$

To prove the first direction ($\Longleftarrow$) of (10), assume that the RHS of (10) holds. By (11), we get that both $\hat{f}, \tilde{f}_x \in \hat{\mathcal{V}}$. However, by construction, $\hat{f}(x) = -\tilde{f}_x(x)$, so $x \in DIS(\hat{\mathcal{V}})$ and $g(x) = 0$.

To prove the other direction ($\Longrightarrow$), assume that $\hat{R}(\tilde{f}_x) - \hat{R}(\hat{f}) > 2\sigma\left(m, \delta/4, d\right)$. Under this assumption, we will prove that for any $f' \in \hat{\mathcal{V}}$, $f'(x) = \hat{f}(x)$, and therefore, $x \in \mathcal{X} \setminus DIS(\hat{\mathcal{V}})$, entailing that $g(x) = 1$. Indeed, assume by contradiction that there exists $f' \in \hat{\mathcal{V}}$ such that $f'(x) = \tilde{f}_x(x) \neq \hat{f}(x)$. By construction, it holds that

$$\hat{R}(f') \geq \hat{R}(\tilde{f}_x) > \hat{R}(\hat{f}) + 2\sigma\left(m, \delta/4, d\right),$$

so $f' \notin \hat{\mathcal{V}}$. Contradiction. $\square$

Lemma 15 tells us that in order to decide if point x should be rejected we need to measure the empirical error $\hat{R}(\tilde{f}_x)$ of a special empirical risk minimizer, $\tilde{f}_x$, which is constrained to label x the opposite from $\hat{h}(x)$. If this error is sufficiently close to $\hat{R}(\hat{h})$, our classifier cannot be too sure about the label of x and we must reject it. Thus, provided we can compute these ERMs, we can decide whether to predict or reject any individual test point $x \in \mathcal{X}$, without actually constructing g for the entire domain $\mathcal{X}$. Figure 3 illustrates this principle for a 2-dimensional example. The hypothesis class is the class of linear classifiers in $\mathbb{R}^2$ and the source distribution is two normal distributions. Negative samples are represented by blue circles and positive samples by red squares. As usual, $\hat{f}$ denotes the empirical

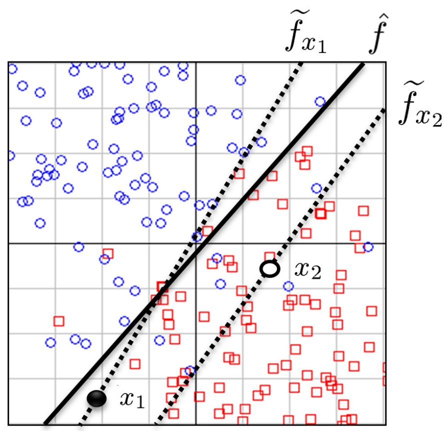

Figure 3: Constrained ERM.

risk minimizer. Let us assume that we want to classify point x_1. This point is classified positive by $\hat{f}$. Therefore, we force this point to be negative and calculate the restricted ERM (depicted by doted line marked $\tilde{f}_{x_1}$). The difference between the empirical risk of $\hat{f}$ and $\widetilde{f}_{x_1}$ is not large enough, so point x_1 will be rejected. However, if we want to classify point x_2, the difference between the empirical risk of $\hat{f}$ and $\tilde{f}_{x_2}$ is quite large and the point will be classified as positive.

Equation (11) motivates the following definition of a "disbelief index" $D_{\mathcal{F}}(x, S_m)$ for each individual point in $\mathcal{X}$. Specifically, for any $x \in \mathcal{X}$, define its *disbelief index* w.r.t. S_m and $\mathcal{F}$,

$$D(x) \triangleq D_{\mathcal{F}}(x, S_m) \triangleq \hat{R}(\tilde{f}_x) - \hat{R}(\hat{f}).$$

Observe that $D(x)$ is large whenever our model is sensitive to the label of x in the sense that when we are forced to bend our best model to fit the opposite label of x, our model substantially deteriorates, giving rise to a large disbelief index. This large $D(x)$ can be interpreted as our disbelief in the possibility that x can be labeled so differently. In this case we should definitely predict the label of x using our unforced model. Conversely, if $D(x)$ is small, our model is indifferent to the label of x and in this sense, is not committed to its label. In this case we should abstain from prediction at x. Notice that LESS is a specific application of thresholded disbelief index.

We note that a similar technique of using an ERM oracle that can enforce an arbitrary number of example-based constraints was used by Dasgupta, Hsu, and Monteleoni (2007a) and Beygelzimer, Hsu, Langford, and Zhang (2010), in the context of active learning. As in our disbelief index, the difference between the empirical risk (or importance weighted empirical risk, see Beygelzimer et al., 2010) of two ERM oracles (with different constraints) is used to estimate prediction confidence.

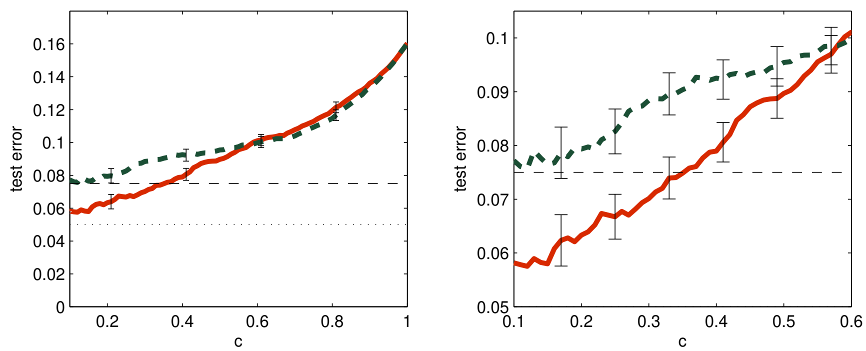

Figure 4: RC curve of our technique (depicted in red) compared to rejection based on distance from decision boundary (depicted in dashed green line). The RC curve in right figure zooms into the lower coverage regions of the left curve.

In practical applications of selective prediction it is desirable to allow for some control over the trade-off between risk and coverage; in other words, it is desirable to be able to develop the entire *risk-coverage (RC) curve* for the classifier at hand (see, e.g., El-Yaniv & Wiener, 2010) and let the user choose the cutoff point along this curve in accordance with other practical considerations and constraints. The disbelief index facilitates an exploration of the risk-coverage trade-off curve for our classifier as follows. Given a pool of test points we can rank these test points according to their disbelief index, and points with low index should be rejected first. Thus, this ranking provides the means for constructing a risk-coverage trade-off curve. Ignoring for the moment implementation details (which are discussed in Section 7), a typical RC curve generated by LESS is depicted in Figure 4 (red curve)[5]. The dashed green RC curve was computed using the traditional distance-based techniques for rejection (see discussion of this common technique in Section 8) The right graph is a zoom in section of the entire RC curve (depicted on the left graph). The dashed horizontal line is the test error of f^* on the entire domain and the dotted line is the Bayes error. While for high coverage values the two techniques are statistically indistinguishable, for any coverage less than 60% we get a significant advantage for LESS. It is clear that in this case not only the estimation error was reduced, but also the test error goes significantly below the optimal test error of f^* for low coverage values.

Interestingly, the disbelief index generates rejection regions that are fundamentally different than those obtained by the traditional distance-based techniques for rejection (see Section 8). To illustrate this point (and still ignoring implementation details), consider Figure 5 where we depict the rejection regions for a training sample of 150 points sampled from a mixture of two identical normal distributions (centered at different locations). The height map in this figure, which correspond to disbelief index magnitude (a), and distance from decision boundary (b), reflect the "confidence regions" of each technique according to its own confidence measure.

[5] The learning problem is the same synthetic problem used for generating Figure 6.

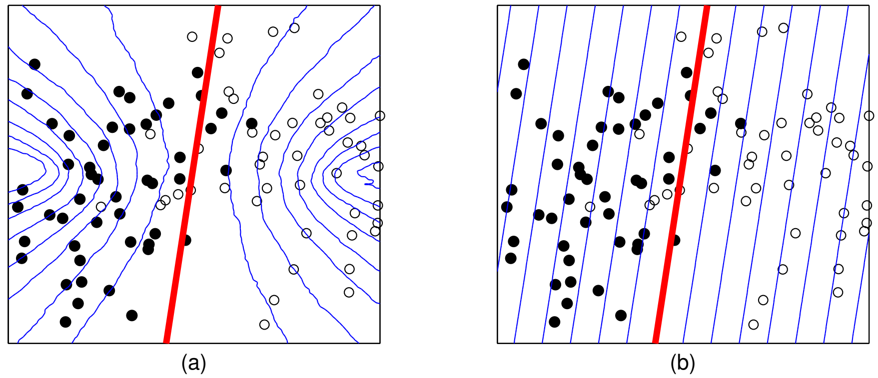

Figure 5: Linear classifier. Confidence height map using (a) disbelief index; (b) distance from decision boundary.

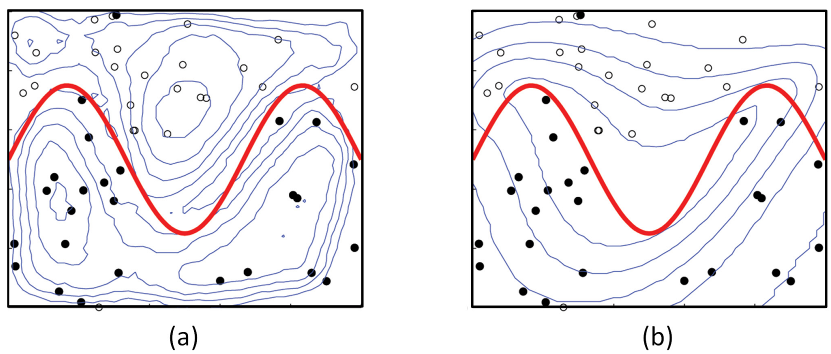

Figure 6: SVM with polynomial kernel. Confidence height map using (a) disbelief index; (b) distance from decision boundary.

To intuitively explain the height map of Figure 5(a), recall that the disbelief index is the difference between the empirical error of the ERM and the restricted ERM. If a test point resides in a high density region, we expect that forcing the wrong label for that point will result in a large increase of the training error. As a result, the denser the area is, the larger the disbelief index, and therefore, the higher the classification confidence.

The second synthetic $2D$ source distribution we consider is even more striking. Here X is distributed uniformly over $[0, 3\pi] \times [-2, 2]$ and the labels are sampled according to the following conditional distribution

$$P(Y = 1|X = (x_1, x_2)) \triangleq \begin{cases} 0.95, & x_2 \geq sin(x_1); \\ 0.05, & \text{else.} \end{cases}$$

The thick red line depicts the decision boundary of the Bayes classifier. The hight maps in Figure 6 depict the rejection regions obtained by (our approximation of) LESS and by the traditional (distance from decision boundary) technique for a training sample of 50 points sampled from this distribution (averaged over 100 iterations). Here the hypothesis class used for training was SVM with a polynomial kernel of degree 5. The qualitative difference between these two techniques, and in particular, the nice fit of the disbelief principle technique compared to SVM is quite surprising.

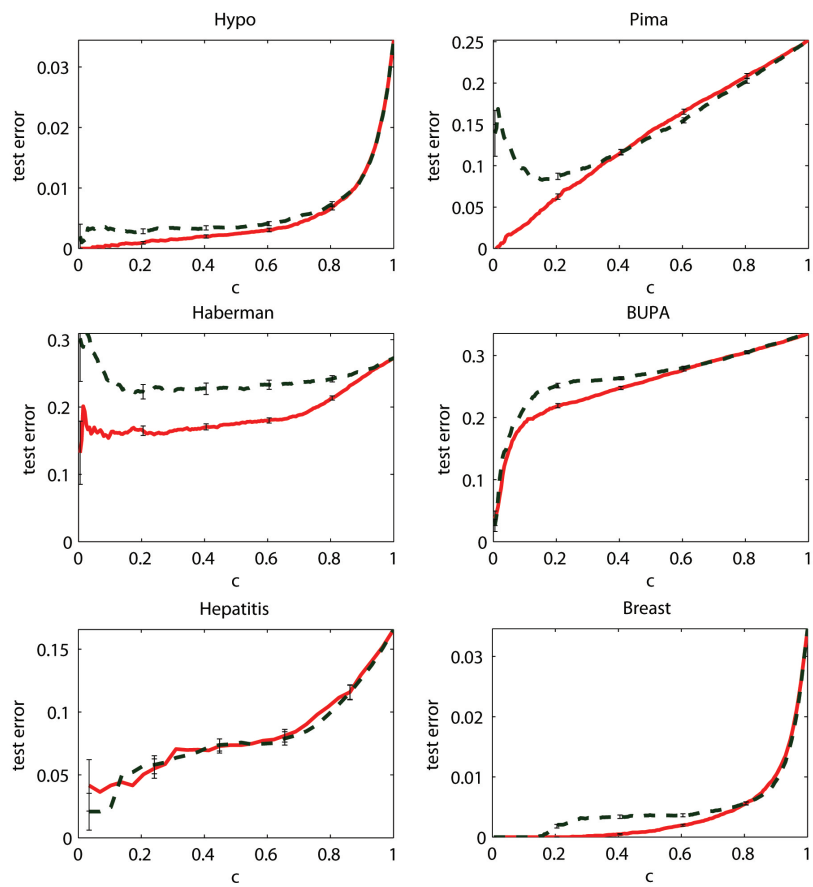

Figure 7: RC curves for SVM with linear kernel. Our method in solid red, and rejection based on distance from decision boundary in dashed green. Horizontal axis (c) represents coverage.

7. Heuristic Procedure Using SVM and its Empirical Performance

The computation of a (constrained) ERM oracle can be efficiently achieved in the case of realizable learning with linear models (see, e.g., El-Yaniv & Wiener, 2010) and in the case of linear regression (Wiener & El Yaniv, 2012). However, in a noisy setting the computation of the linear ERM oracle can be reduced to a variant of the MAX FLS and C MAX FLS problems (with strict and non-strict inequalities) (Amaldi & Kann, 1995). Unfortunately, MAX FLS is APX-complete (within a factor 2). C MAX FLS is MAX IND SET-hard, and cannot be approximated efficiently at all. Moreover, there are extensions of these results to other classes, including axis-aligned hyper-rectangles, showing that approximating ERM for these classes is NP-hard (Ben-David et al., 2003).

While at present it is not known if these hardness results (and other related lower bounds) hold for half spaces under nice distributions such as Gaussian (mixtures), we note that Tauman Kalai et al. (2008) studied the problem of agnostically learning halfspaces under distributional assumptions. In particular, they showed that if the data distribution is uniform over the d-dimensional unit sphere (or hyper-cube, and other related distributions), then it is possible to agnostically learn ϵ-accurate halfspaces in time $poly(d^{1/\epsilon^4})$. However, it is known that these particular distributions do not elicit effective pointwise competitive learning. On the contrary, the uniform distribution over the unit sphere is among the worst possible distributions for pointwise-competitive classification (and disagreement-based active learning) unless one utilizes *homogeneous* halfspaces (see discussion in, e.g., El-Yaniv & Wiener, 2010).

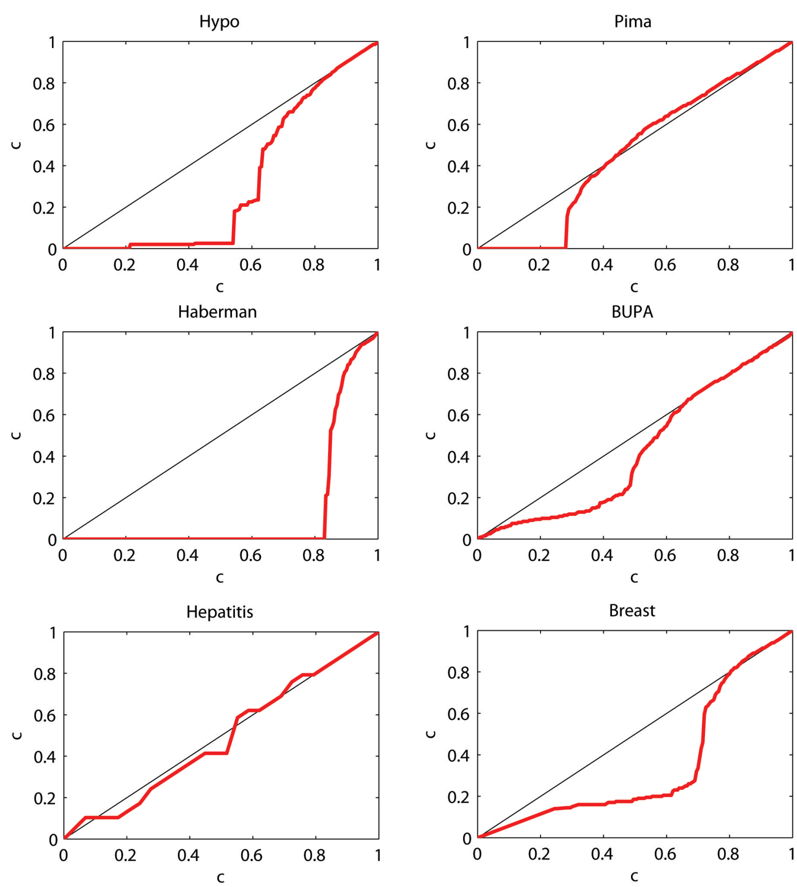

Figure 8: SVM with linear kernel. The maximum coverage for a distance-based rejection technique that allows the same error rate as our method with a specific coverage.

Having discussed these computational hurdles, we should recall that much of applied machine learning research and many of its applications are doing quite well with heuristic approximations (rather than formal ones). When practical performance is the objective, clever heuristics and tricks can sometimes make the difference. At this point in the paper we therefore switch from theory to practice, aiming at implementing a rejection method inspired by the disbelief principle and see how well they work on real world problems.

We "approximate" the ERM as follows. Using support vector machines (SVMs) we use a high C value (10^5 in our experiments) to penalize more on training errors than on small margin (see definitions of the SVM parameters in, e.g. Chang & Lin, 2011). In this way the solution to the optimization problem tend to get closer to the ERM. In order to estimate $\hat{R}(\tilde{f}_x)$ we have to restrict the SVM optimizer to only consider hypotheses that classify the point x in a specific way. To accomplish this we use a weighted SVM for unbalanced data. We add the point x as another training point with weight 10 times larger than the weight of all training points combined. Thus, the penalty for misclassification of x is very large and the optimizer finds a solution that doesn't violate the constraint.

Another problem we face is that the disbelief index is a noisy statistic that highly depends on the sample S_m. To overcome this noise we use robust statistics. First we generate an odd number k of different samples $(S_m^1, S_m^2, \ldots S_m^k)$ using bootstrap sampling (we used $k = 11$). For each sample we calculate the disbelief index for all test points and for each point take the median of these measurements as the final index. We also note that for any finite training sample the disbelief index is a discrete variable. It is often the case that several test points share the same disbelief index. In those cases we can use any confidence measure as a tie breaker. In our experiments we use distance from decision boundary to break ties. Focusing on SVMs with a linear kernel we compared the RC (Risk-Coverage) curves achieved by the proposed method with those achieved by SVM with rejection based on distance from decision boundary. This latter approach is very common in practical applications of selective classification. For implementation we used LIBSVM (Chang & Lin, 2011).

We tested our algorithm on standard medical diagnosis problems from the UCI repository, including all datasets used by Grandvalet, Rakotomamonjy, Keshet, and Canu (2008). We transformed nominal features to numerical ones in a standard way using binary indicator attributes. We also normalized each attribute independently so that its dynamic range is $[0, 1]$. No other preprocessing was employed. In each iteration we choose uniformly at random non-overlapping training set (100 samples) and test set (200 samples) for each dataset.[6] The SVM was trained on the entire training set, and test samples were sorted according to confidence (either using distance from decision boundary or disbelief index).

Figure 7 depicts the RC curves of our technique (red solid line) and rejection based on distance from decision boundary (green dashed line) for linear kernel on all 6 datasets. All results are averaged over 500 iterations (error bars show standard error). With the exception of the Hepatitis dataset, in which both methods were statistically indistinguishable, in all other datasets the proposed method exhibits significant advantage over the traditional approach. We would like to highlight the performance of the proposed method on the Pima dataset. While the traditional approach cannot achieve error less than 8% for any

[6] Due to the size of the Hepatitis dataset the test set was limited to 29 samples.

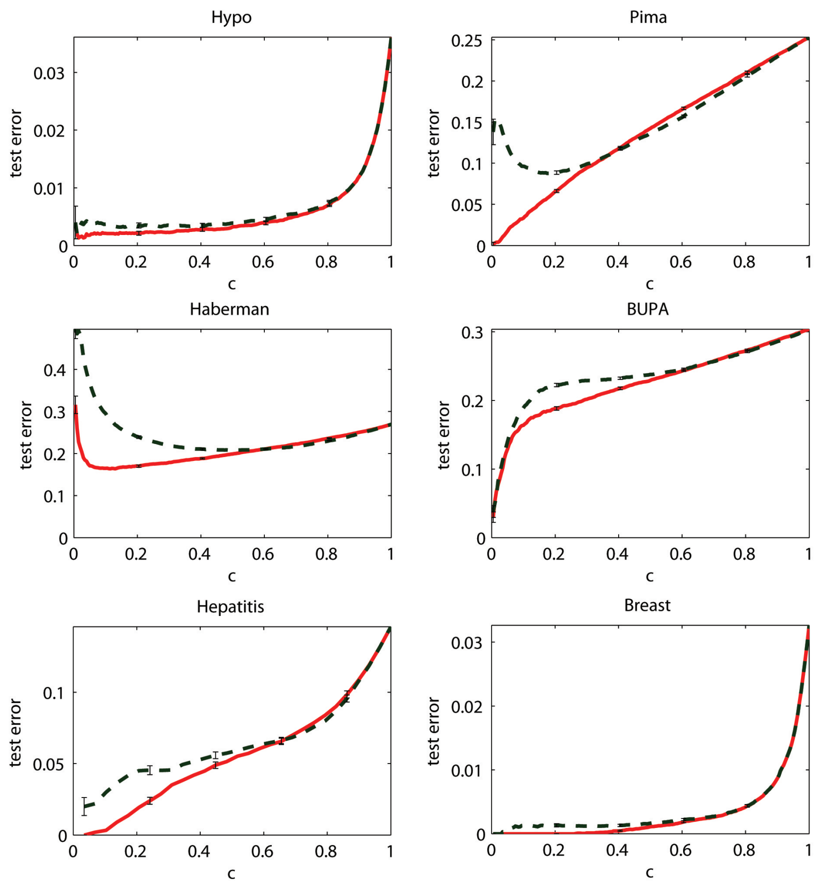

Figure 9: RC curves for SVM with RBF kernel. Our method in solid red and rejection based on distance from decision boundary in dashed green.

rejection rate, in our approach the test error decreases monotonically to zero with rejection rate. Furthermore, a clear advantage for our method over a large range of rejection rates is evident in the Haberman dataset.[7].

For the sake of fairness, we note that the running time of our algorithm (as presented here) is substantially longer than the traditional technique. The performance of our algorithm can be substantially improved when many unlabeled samples are available. In this case the rejection function can be evaluated on the unlabeled samples to generate a new "labeled" sample. Then a new rejection classifier can be trained on this sample.

Figure 8 depicts the maximum coverage for a distance-based rejection technique that allows the same error rate as our method with a specific coverage. For example, let us assume that our method can have an error rate of 10% with coverage of 60% and the

7 The Haberman dataset contains survival data of patients who had undergone surgery for breast cancer. With estimated 207,090 new cases of breast cancer in the united states during 2010 (Society, 2010) an improvement of 1% affects the lives of more than 2000 women.

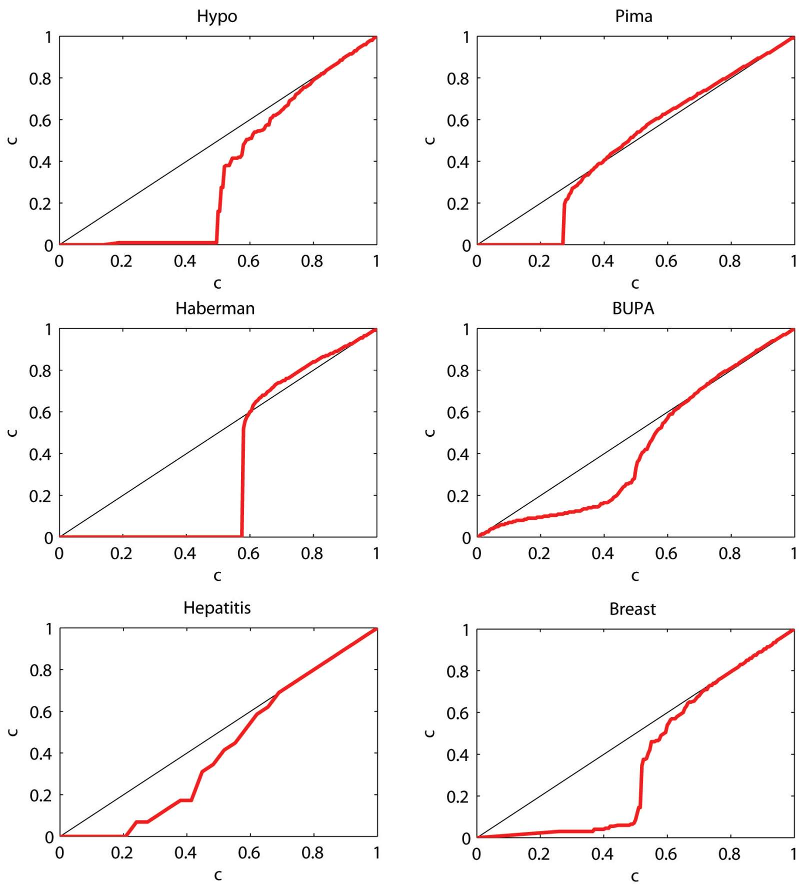

Figure 10: SVM with RBF kernel. The maximum coverage for a distance-based rejection technique that allows the same error rate as our method with a specific coverage.

distance-based rejection technique achieves the same error with maximum coverage of 40%. Then the point $(0.6, 0.4)$ will be on the red line. Thus, if the red line is bellow the diagonal then our technique has an advantage over distance-based rejection and visa versa. As an example, consider the Haberman dataset, and observe that regardless of the rejection rate, distance-based technique cannot achieve the same error as our technique with coverage lower than 80%.

Figures 9 and 10 depict the results obtained with RBF kernel. In this case a statistically significant advantage for our technique was observed for all datasets.

8. Related Work

Pointwise-competitive classification is a unique and extreme instance of classification with an abstention option, an idea which emerged from the pattern recognition community, was first proposed and studied 50 years ago by Chow (1957, 1970), and generated lots of interest

(Fumera et al., 2001; Tortorella, 2001; Santos-Pereira & Pires, 2005; Fumera & Roli, 2002; Pietraszek, 2005; Bounsiar et al., 2006; Landgrebe et al., 2006; Herbei & Wegkamp, 2006; Hellman, 1970; El-Yaniv & Pidan, 2011; Bartlett & Wegkamp, 2008; Wegkap, 2007; Freund et al., 2004). Taking a broader perspective, pointwise-competitive selective prediction (and in particular, classification) is a particular instance of the broader concept of *confidence-rated learning*, whereby the learner must formally quantify confidence in its prediction. Achieving effective confidence-rated prediction (including abstention) is a longstanding and challenging goal in a number of disciplines and research communities. Let us first discuss some of the most prominent approaches to confidence-rated prediction and note how they related to the present work.

In the 'knows-what-it-knows' (KWIK) framework studied in reinforcement-learning (Li, Littman, & Walsh, 2008; Strehl & Littman, 2007; Li & Littman, 2010) a similar notion to pointwise competitiveness is studied, and coverage rates are analyzed (Li et al., 2008; Li, 2009). However, KWIK was limited to the realizable model and is concerned with an adversarial setting where both the target hypothesis and the training data are selected by an adversary. While all positive results for the KWIK adversarial setting apply to the statistical pointwise-competitive prediction setting (where training examples are sampled i.i.d.), this adversarial setting precludes non trivial coverage for all the interesting hypothesis classes currently addressed by pointwise-competitive prediction. This deficiency comes as no surprise because the KWIK adversarial setting is much more challenging than the statistical pointwise-competitive prediction assumptions.

The *conformal prediction* framework (Vovk, Gammerman, & Shafer, 2005) provides hedged predictions by allowing the possibility of multi-labeled predictions and guarantees a user-desired confidence rate in an asymptotic sense. Conformal prediction is mainly concerned with an online probabilistic setting. Rather than predicting a single label for each sample point, a conformal predictor can assign multiple labels. Any user-defined confidence level ϵ for the error rate can be asymptotically guaranteed. When interpreting these multi-labeled predictions as rejection, we can compare it to pointwise-competitive prediction. In this sense, conformal prediction can construct online predictors with a reject option that have asymptotic performance guarantees. A few important differences between conformal predictions and pointwise-competitive prediction can be pointed out. While both approaches provide "hedged" predictions, they use different notions of hedging. Whereas in pointwise-competitive prediction the goal is to guarantee that with high probability *over the training sample* our predictor agrees with the best predictor in the same class over *all* points in the accepted domain, the goal in conformal predictions is to provide guarantees for the *average* error rate, where the average is taken over all possible samples and test points.[8] In this sense, conformal prediction cannot achieve pointwise competitiveness. In addition, conformal prediction also utilizes a different notion of error than the one used in the pointwise-competitive model. While pointwise-competitive prediction is focused on performance guarantees for the error rate only on the covered (accepted) examples, conformal prediction provides a guarantee for all examples (including those that have multiple predictions or none at all). By increasing the multi-labeled prediction rate (uncertain prediction),

[8] As noted by Vovk et al.: "It is impossible to achieve the conditional probability of error equal to ϵ given the observed examples, but it is the unconditional probability of error that equals ϵ. Therefore, it implicitly involves averaging over different data sequences..." (Vovk et al., 2005, p. 295).

the error rate can be decreased to any arbitrarily small value. This is not the case with the pointwise-competitive prediction error notion on the covered examples, which is bounded below by the Bayes error on the covered region. Finally, conformal prediction mentions a notion of *efficiency*, which is similar to coverage but, to the best of our knowledge, no finite sample results have been established. Another interesting scheme in the vicinity of confidence-rated learning is the the *guaranteed error machine (GEM)* (Campi, 2010). In the GEM model the reject option is considered as a correct answer, which means that risk can be reduced arbitrarily (as in conformal prediction).

Pointwise-competitive classification is a special case of pointwise-competitive *prediction* (El-Yaniv & Wiener, 2010, 2011; Wiener & El Yaniv, 2012; El-Yaniv & Wiener, 2012; Wiener, 2013; Wiener et al., 2014). Pointwise-competitive selective classification was first addressed by El-Yaniv and Wiener (2010) where the realizable case was studied (in that paper pointwise-competitiveness was termed "perfect classification"). The present article extends pointwise-competitive classification to noisy problems

There are also a number of theoretical studies of (general) selective classification (not pointwise-competitive). Freund et al. (2004) studied a simple ensemble method for binary classification. Given a hypothesis class $\mathcal{F}$, the method outputs a weighted average of all the hypotheses in $\mathcal{F}$, where the weight of each hypothesis exponentially depends on its individual training error. Their algorithm abstains from prediction whenever the weighted average of all individual predictions is close to zero. They were able to bound the probability of misclassification by $2R(f^*) + \epsilon(m)$ and, under some conditions, they proved a bound of $5R(f^*) + \epsilon(\mathcal{F}, m)$ on the rejection rate. The LESS strategy can be viewed as an extreme variation of the Freund et al. method. We include in our "ensemble" only hypotheses with sufficiently low empirical error and we abstain if the weighted average of all predictions is not definitive ($\neq \pm 1$). Our risk and coverage bounds are asymptotically tighter.

Excess risk bounds were developed by Herbei and Wegkamp (2006) for a model where each rejection incurs a cost in $[0, 1/2]$. Their bound applies to any empirical risk minimizer over a hypothesis class of ternary hypotheses (whose output is in $\{\pm 1, \text{reject}\}$). See also various extensions by Wegkap (2007) and Bartlett and Wegkamp (2008).

A rejection mechanism for SVMs based on distance from decision boundary is perhaps the most widely known and used rejection technique. It is routinely used in medical applications (Mukherjee et al., 1998; Guyon et al., 2002; Mukherjee, 2003). Few papers proposed alternative techniques for rejection in the case of SVMs. Those include taking the reject area into account during optimization (Fumera & Roli, 2002), training two SVM classifiers with asymmetric cost (Sousa, Mora, & Cardoso, 2009), and using a hinge loss (Bartlett & Wegkamp, 2008). Grandvalet et al. (2008) proposed an efficient implementation of SVM with a reject option using a double hinge loss. They empirically compared their results with two other selective classifiers: the one proposed by Bartlett and Wegkamp (2008) and the traditional rejection based on distance from decision boundary. In their experiments there was no statistically significant advantage to either method compared to the traditional approach for high rejection rates.

Pointwise selective classification is strongly tied to disagreement-based active learning. For the realizable case, El-Yaniv and Wiener (2012) presented a reduction of stream-based active learning with the CAL algorithm of Cohn et al. (1994) to pointwise-competitive classification. This reduction roughly states that if the rejection rate (the reciprocal of coverage) of LESS is $O(polylog(m/\delta)/m)$ then the problem $(\mathcal{F}, P)$ is actively learnable by CAL with exponential speedup. A consequence of this reduction resulted in the first exponential speedup bounds for CAL with general linear models under any finite mixture of Gaussians. The other direction, showing that exponential speedup for CAL implies the above rejection rate for LESS (in the realizable setting) was recently established by Wiener (2013) and by Wiener, Hanneke, and El-Yaniv (2014) (using two different techniques).

The version space compression set size, which is extensively utilized in the present work, has been introduced implicitly by Hanneke (2007b) as a special case of the *extended teaching dimension*, and in that context, the version space compression *set* is called the *minimal specifying set*. It was introduced explicitly by El-Yaniv and Wiener (2010) in the context of selective classification, and was proved by El-Yaniv and Wiener (2012) to be a special case of the extended teaching dimension of Hanneke (2007b). Relations between the disagreement coefficient and the version space compression set size were first discussed El-Yaniv and Wiener (2012). Sharp ties between these two quantities, such as those stated in Lemma 9, and others were very recently developed by Wiener, Hanneke, and El-Yaniv (2014).

9. Concluding Remarks

We find the existence of pointwise-competitive classification quite fascinating. The striking feature of such a classifier is that, by definition, a pointwise-competitive predictor is free of estimation error and cannot overfit. This means that our hypothesis class can be as expressive as we like and still we will be protected from overfitting. However, without effective coverage bounds our pointwise-competitive classifier may refuse to predict at all times.

The current paper, and recent studies on both selective prediction (El-Yaniv & Wiener, 2015) and active learning (Wiener, Hanneke, & El-Yaniv, 2014), place the version space compression set size at the center of stage, as a leading quantity that can drive results and intuition in both domains. At present, this is the only known technique able to prove fast coverage for pointwise-competitive classification and exponential label complexity speedup for disagreement-based active learning for both general linear models under a fixed mixture of Gaussians and axis aligned rectangles under product distributions.. Is it possible to extend these results beyond linear classifiers and axis aligned rectangles under interesting distribution families? For example, it is plausible that existing results for axis-aligned rectangles can be extended to decision trees.

The formal relationship between active learning and pointwise-competitive classification (El-Yaniv & Wiener, 2012; Wiener, 2013; Wiener et al., 2014) created a powerful synergy that allows for migrating results between these two models. Currently, this formal connection is manifested via two links. The first, within a realizable setting, is the equivalence of LESS-based classification with fast coverage to CAL-based active learning with exponential speedup. The second link consists of bounds that relate the underlying complexity measures: the disagreement coefficient in active learning, and version space compression set size in pointwise-competitive classification. A number of other non-established relations that can significantly substantiate the interaction between the two problems could be considered. For example, is it possible to prove a direct equivalence between LESS-based pointwise-competitive agnostic classification with fast coverage rates and LESS-based active learning with exponential speedup? We expect that a resolution of this question will have various interesting implications. For example, such a relationship could potentially facilitate the migration of very interesting algorithms and techniques devised for active learning to the pointwise-competitive framework. An immediate candidate is the algorithm of Beygelzimer et al. (2010), which builds on ideas of Dasgupta et al. (2007b) and Beygelzimer et al. (2009). Resembling the implementation proposed for LESS via calls to (a constrained) ERM oracle, this algorithm works without tracking the version space for both the final choice of the hypothesis as well as the querying component. Instead, for querying, it relies on an ERM oracle that enforces at most one example-based constrain. Thus, the importance-weighting technique on which it is based resembles the disbelief principle we outline here. In this regard, it will be very interesting to also consider and migrate ideas from active learning algorithms emerging from the online learning branch (Orabona & Cesa-Bianchi, 2011; Cesa-Bianchi et al., 2009; Dekel et al., 2010) while using, as required, online to batch conversion techniques (Zhang, 2005; Kakade & Tewari, 2009; Cesa-Bianchi & Gentile, 2008; Dekel, 2008).

The LESS strategy requires a unanimous vote among all hypotheses in a low empirical error subset of the hypothesis class. When considering, e.g., linear models, this subset of hypotheses is uncountable, and in any case (even if it is finite) its size can be huge. Clearly, LESS is an extremely radical and defensive strategy. An immediate question that arises is whether the LESS unanimity requirement can be relaxed to a majority vote. Can we achieve pointwise competitiveness with only a (strong) majority vote instead of unanimity? Besides the greater flexibility of a general voting scheme, which may lead to different types of interesting learning algorithms, such a relaxation can potentially ease the computational complexity of implementing LESS (which, as discussed above, is a bottleneck in agnostic classification). For example, with a relaxed voting scheme we might utilize *hypothesis sampling*, for which a classical example in a related context is the celebrated query-by-committee (QBC) strategy (Seung et al., 1992; Freund et al., 1997; Fine et al., 2002; Gilad-Bachrach, 2007; Gilad-Bachrach et al., 2005). However, if strict pointwise competitiveness is advocated, it is easy to see any strong majority vote is not sufficient. Indeed, consider an f^* that differs from all other hypotheses in $\mathcal{F}$ on a single point in $\mathcal{X}$. Unless the probability of this point is very large (not the typical case), with high probability this point is not part of the training set S_m, and therefore, any majority vote (even very strong) will label it the opposite of f^*. Hence, in the worst case, even a strong majority is not sufficient for pointwise competitiveness. As a natural compromise for the pointwise competitiveness objective, one can revert to standard excess-risk bounds (Bartlett et al., 2006) whereby we compare the overall *average* performance of our predictor, $R(f)$, to that of the optimal predictor, $R(f^*)$ (not pointwise). In this regard, the work of Freund, Mansour, and Schapire (2004) discussed in Section 8, is such a result with its excess-risk bound $R(f,g) \leq 2R(f^*) + O\left(1/(m^{1/2-\theta})\right)$ (θ is a hyper-parameter) and coverage bound $\Phi(f,g) \geq 1 - 5R(f^*) - O\left(\ln|\mathcal{F}|/\sqrt{m^{1/2-\theta}}\right)$. Considering excess-risk bounds against f^*, is it possible to beat the above risk and coverage bounds using a relaxed voting scheme for rejection? What would be the optimal bounds in a fully agnostic setting? Can better bounds be devised for specific distributions like Gaussian mixtures? We note that the Freund et al. strategy is also interesting because the final aggregated predictor is in general outside of $\mathcal{F}$ and can, in principle, significantly outperform $f^* \in F$ (the above bound does not elicit such a behavior). This emphasizes the potential usefulness of ensembles, applied not only in the rejection scheme, but also in the final predictor. Recall that in the LESS strategy the final predictor always belongs to $\mathcal{F}$. Thus, when considering ensembles and allowing excess-risk bounds, there can be even more ambitious goals, such as strictly beating f^* on average.

Acknowledgments

We thank the anonymous referees for their good comments, and are grateful to Steve Hanneke for helpful and insightful discussions. Also, we warmly thank the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI), and Israel Science Foundation (ISF) for their generous support.

Appendix A. Some Proofs

The proof of Lemma 9 below relies on the following Lemma 16 (Wiener et al., 2014), whose proof is also provided here for the sake of self-containment.

Lemma 16 (Wiener et al., 2014)**.** *In the realizable case, for any* $r_0 \in (0, 1)$,

$$\theta(r_0) \leq \max\left\{\max_{r\in(r_0,1)} 16\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right), 512\right\}.$$

Proof. We will prove that, for any $r \in (0, 1)$,

$$\frac{\Delta\mathrm{B}(f^*, r)}{r} \leq \max\left\{16\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right), 512\right\}. \tag{12}$$

The result then follows by taking the supremum of both sides over $r \in (r_0, 1)$.

Fix $r \in (0, 1)$, let $m = \lceil 1/r \rceil$, and for $i \in \{1, \ldots, m\}$, define $S_{m\setminus i} = S_m \setminus \{(x_i, y_i)\}$. Also define $D_{m\setminus i} = \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}} \cap \mathrm{B}(f^*, r))$ and $\Delta_{m\setminus i} = \mathbb{P}(x_i \in D_{m\setminus i}|S_{m\setminus i}) = P(D_{m\setminus i} \times \mathcal{Y})$. If $\Delta\mathrm{B}(f^*, r)m \leq 512$, (12) clearly holds. Otherwise, suppose $\Delta\mathrm{B}(f^*, r)m > 512$. If $x_i \in \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})$, then we must have $(x_i, y_i) \in \hat{\mathcal{C}}_{S_m}$. So

$$\hat{n}(S_m) \geq \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})}(x_i).$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}\left\{\hat{n}(S_m) \le (1/16)\Delta\mathrm{B}(f^*,r)m\right\} \\
&\le \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})}(x_i) \le (1/16)\Delta\mathrm{B}(f^*,r)m\right\} \\
&\le \mathbb{P}\left\{\sum_{i=1}^{m} {}_{D_{m\setminus i}}(x_i) \le (1/16)\Delta\mathrm{B}(f^*,r)m\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - {}_{D_{m\setminus i}}(x_i) \ge \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - (1/16)\Delta\mathrm{B}(f^*,r)m\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - {}_{D_{m\setminus i}}(x_i) \ge \right. \\
&\qquad \left. \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \frac{1}{16}\Delta\mathrm{B}(f^*,r)m, \quad \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < \frac{7}{8}\Delta\mathrm{B}(f^*,r)m\right\} \\
&+ \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - {}_{D_{m\setminus i}}(x_i) \ge \right. \\
&\qquad \left. \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \frac{1}{16}\Delta\mathrm{B}(f^*,r)m, \quad \sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) \ge \frac{7}{8}\Delta\mathrm{B}(f^*,r)m\right\} \\
&\le \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < (7/8)\Delta\mathrm{B}(f^*,r)m\right\} \\
&\quad + \mathbb{P}\left\{\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - {}_{D_{m\setminus i}}(x_i) \ge (13/16)\Delta\mathrm{B}(f^*,r)m\right\}.
\end{aligned}
$$

Since we are considering the case $\Delta\mathrm{B}(f^*,r)m > 512$, a Chernoff bound implies

$$\mathbb{P}\left(\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < (7/8)\Delta\mathrm{B}(f^*,r)m\right) \le \exp\left\{-\Delta\mathrm{B}(f^*,r)m/128\right\} < e^{-4}.$$

Furthermore, Markov's inequality implies

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{m} {}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - {}_{D_{m\setminus i}}(x_i) \ge (13/16)\Delta\mathrm{B}(f^*,r)m\right) \\
&\qquad\qquad \le \frac{m\Delta\mathrm{B}(f^*,r) - \mathbb{E}\left[\sum_{i=1}^{m} {}_{D_{m\setminus i}}(x_i)\right]}{(13/16)m\Delta\mathrm{B}(f^*,r)}.
\end{aligned}
$$

Since the x_i values are exchangeable,

$$\mathbb{E}\left[\sum_{i=1}^{m} {}_{D_{m\setminus i}}(x_i)\right] = \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{E}\left[{}_{D_{m\setminus i}}(x_i) \middle| S_{m\setminus i}\right]\right] = \sum_{i=1}^{m} \mathbb{E}\left[\Delta_{m\setminus i}\right] = m\mathbb{E}\left[\Delta_{m\setminus m}\right].$$

it can be shown (Hanneke, 2012) that this is at least

$$m(1-r)^{m-1}\Delta\mathrm{B}(f^*,r).$$

In particular, when $\Delta\mathrm{B}(f^*,r)m > 512$, we must have $r < 1/511 < 1/2$, which implies $(1-r)^{\lceil 1/r \rceil - 1} \geq 1/4$, so that we have

$$\mathbb{E}\left[\sum_{i=1}^{m} \quad {}_{D_{m\setminus i}}(x_i)\right] \geq (1/4)m\Delta\mathrm{B}(f^*,r).$$

Altogether, we have established that

$$\begin{aligned}\mathbb{P}\left(\hat{n}(S_m) \leq (1/16)\Delta\mathrm{B}(f^*,r)m\right) < \frac{m\Delta\mathrm{B}(f^*,r) - (1/4)m\Delta\mathrm{B}(f^*,r)}{(13/16)m\Delta\mathrm{B}(f^*,r)} + e^{-4} \\ = \frac{12}{13} + e^{-4} < 19/20.\end{aligned}$$

Thus, since $\hat{n}(S_m) \leq \mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right)$ with probability at least 19/20, we must have that

$$\mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) > (1/16)\Delta\mathrm{B}(f^*,r)m \geq (1/16)\frac{\Delta\mathrm{B}(f^*,r)}{r}. \qquad \square$$

Proof of Lemma 9. Assuming that $\mathcal{B}_{\hat{n}}(m,\delta) = O\left(\mathrm{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$ holds, there exists a constant $\delta_1 \in (0, 1/20)$ for which $\mathcal{B}_{\hat{n}}(m,\delta_1) = O\left(\mathrm{polylog}(m)\right)$. Because $\mathcal{B}_{\hat{n}}(m,\delta)$ is non-increasing with δ, $\mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) \leq \mathcal{B}_{\hat{n}}(m,\delta_1)$, and thus $\mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) = O\left(\mathrm{polylog}(m)\right)$. Therefore,

$$\max_{m \leq 1/r_0} \mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) = O\left(\max_{m \leq 1/r_0} \mathrm{polylog}(m)\right) = O\left(\mathrm{polylog}\left(\frac{1}{r_0}\right)\right),$$

and using Lemma 16 we have,

$$\begin{aligned}\theta(r_0) &\leq \max\left\{\max_{m \leq \lceil 1/r_0 \rceil} 16\mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right), 512\right\} \\ &\leq 528 + 16 \max_{m \leq 1/r_0} \mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) = O\left(\mathrm{polylog}\left(\frac{1}{r_0}\right)\right).\end{aligned}$$

$\square$

References

Amaldi, E., & Kann, V. (1995). The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical computer science*, *147*(1), 181–210.

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Bartlett, P., & Mendelson, S. (2006). Discussion of "2004 IMS medallion lecture: Local rademacher complexities and oracle inequalities in risk minimization" by V. koltchinskii. *Annals of Statistics*, *34*, 2657–2663.

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

Bartlett, P., & Wegkamp, M. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, *9*, 1823–1840.

Ben-David, S., Eiron, N., & Long, P. (2003). On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, *66*(3), 496–514.

Beygelzimer, A., Dasgupta, S., & Langford, J. (2009). Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 49–56. ACM.

Beygelzimer, A., Hsu, D., Langford, J., & Zhang, T. (2010). Agnostic active learning without constraints. *Advances in Neural Information Processing Systems 23*.

Beygelzimer, A., Dasgupta, S., & Langford, J. (2009). Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56. ACM.

Beygelzimer, A., Hsu, D., Langford, J., & Zhang, T. (2010). Agnostic active learning without constraints. *arXiv preprint arXiv:1006.2588*.

Bounsiar, A., Grall, E., & Beauseroy, P. (2006). A kernel based rejection method for supervised classification. *International Journal of Computational Intelligence*, *3*, 312–321.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, Vol. 3176 of *Lecture Notes in Computer Science*, pp. 169–207. Springer.

Campi, M. (2010). Classification with guaranteed probability of error. *Mach. Learn.*, *80*(1), 63–84.

Cesa-Bianchi, N., & Gentile, C. (2008). Improved risk tail bounds for on-line algorithms. *Information Theory, IEEE Transactions on*, *54*(1), 386–390.

Cesa-Bianchi, N., Gentile, C., & Orabona, F. (2009). Robust bounds for classification via selective sampling. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 121–128. ACM.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. Software available at "http://www.csie.ntu.edu.tw/ cjlin/libsvm".

Chow, C. (1957). An optimum character recognition system using decision function. *IEEE Trans. Computer*, *6*(4), 247–254.

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

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

Dasgupta, S., Hsu, D., & Monteleoni, C. (2007a). A general agnostic active learning algorithm. In *NIPS*.

Dasgupta, S., Monteleoni, C., & Hsu, D. J. (2007b). A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pp. 353–360.

Dekel, O. (2008). From online to batch learning with cutoff-averaging.. NIPS.

Dekel, O., Gentile, C., & Sridharan, K. (2010). Robust selective sampling from single and multiple teachers.. In *COLT*, pp. 346–358.

El-Yaniv, R., & Pidan, D. (2011). Selective prediction of financial trends with hidden markov models. In *NIPS*, pp. 855–863.

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

El-Yaniv, R., & Wiener, Y. (2011). Agnostic selective classification. In *Neural Information Processing Systems (NIPS)*.

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

El-Yaniv, R., & Wiener, Y. (2015). On the version space compression set size and its applications. In Vovk, V., Papadopoulos, H., & Gammerman, A. (Eds.), *Measures of Complexity: Festschrift for Alexey Chervonenkis*. Springer, Berlin.

Fine, S., Gilad-Bachrach, R., & Shamir, E. (2002). Query by committee, linear separation and random walks. *Theoretical Computer Science*, *284*(1), 25–51.

Freund, Y., Mansour, Y., & Schapire, R. (2004). Generalization bounds for averaged classifiers. *Annals of Statistics*, *32*(4), 1698–1722.

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

Friedman, E. (2009). Active learning for smooth problems. *In Proceedings of the* 22^{nd} *Annual Conference on Learning Theory*.

Fumera, G., & Roli, F. (2002). Support vector machines with embedded reject option. In *Pattern Recognition with Support Vector Machines: First International Workshop*, pp. 811–919.

Fumera, G., Roli, F., & Giacinto, G. (2001). Multiple reject thresholds for improving classification reliability. *Lecture Notes in Computer Science*, *1876*.

Gilad-Bachrach, R. (2007). *To PAC and Beyond*. Ph.D. thesis, the Hebrew University of Jerusalem.

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2005). Query by committee made real. In *NIPS*.

Grandvalet, Y., Rakotomamonjy, A., Keshet, J., & Canu, S. (2008). Support vector machines with a reject option. In *NIPS*, pp. 537–544. MIT Press.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines.. *Machine Learning*, 389–422.

Hanneke, S. (2007a). A bound on the label complexity of agnostic active learning. In *ICML*, pp. 353–360.

Hanneke, S. (2007b). Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, Vol. 4539 of *Lecture Notes in Artificial Intelligence*, pp. 66–81.

Hanneke, S. (2009). *Theoretical Foundations of Active Learning*. Ph.D. thesis, Carnegie Mellon University.

Hanneke, S. (2013). A statistical theory of active learning. *Unpublished*.

Hanneke, S. (2012). Activized learning: Transforming passive to active with improved label complexity. *The Journal of Machine Learning Research*, *98888*, 1469–1587.

Hellman, M. (1970). The nearest neighbor classification rule with a reject option. *IEEE Trans. on Systems Sc. and Cyb.*, *6*, 179–185.

Herbei, R., & Wegkamp, M. (2006). Classification with reject option. *The Canadian Journal of Statistics*, *34*(4), 709–721.

Kakade, S., & Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808.

Koltchinskii, V. (2006). 2004 IMS medallion lecture: Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, *34*, 2593–2656.

Landgrebe, T., Tax, D., Paclík, P., & Duin, R. (2006). The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, *27*(8), 908–917.

Li, L., & Littman, M. L. (2010). Reducing reinforcement learning to kwik online regression. *Annals of Mathematics and Artificial Intelligence*, 217–237.

Li, L., Littman, M., & Walsh, T. (2008). Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 568–575. ACM.

Li, L. (2009). *A unifying framework for computational reinforcement learning theory*. Ph.D. thesis, Rutgers, The State University of New Jersey.

Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Annales de la Faculté des Sciences de Toulouse*, Vol. 9, pp. 245–303. Université Paul Sabatier.

Mendelson, S. (2002). Improving the sample complexity using global data. *Information Theory, IEEE Transactions on*, *48*(7), 1977–1991.

Mukherjee, S. (2003). Chapter 9. classifying microarray data using support vector machines. In *of scientists from the University of Pennsylvania School of Medicine and the School of Engineering and Applied Science*. Kluwer Academic Publishers.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J. P., & Poggio, T. (1998). Support vector machine classification of microarray data. Tech. rep., AI Memo 1677, Massachusetts Institute of Technology.

Orabona, F., & Cesa-Bianchi, N. (2011). Better algorithms for selective sampling. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 433–440.

Pietraszek, T. (2005). Optimizing abstaining classifiers using ROC analysis. In *Proceedings of the Twenty-Second International Conference on Machine Learning(ICML)*, pp. 665–672.

Santos-Pereira, C., & Pires, A. (2005). On optimal reject rules and ROC curves. *Pattern Recognition Letters*, *26*(7), 943–952.

Seung, H., Opper, M., & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning theory (COLT)*, pp. 287–294.

Society, A. C. (2010). Cancer facts & figures 2010..

Sousa, R., Mora, B., & Cardoso, J. (2009). An ordinal data method for the classification with reject option. In *ICMLA*, pp. 746–750. IEEE Computer Society.

Strehl, A. L., & Littman, M. L. (2007). Online linear regression and its application to model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1417–1424.

Tauman Kalai, A., Klivans, A., Mansour, Y., & Servedio, R. (2008). Agnostically learning halfspaces. *SIAM J. Comput.*, *37*(6), 1777–1805.

Tortorella, F. (2001). An optimal reject rule for binary classifiers. *Lecture Notes in Computer Science*, *1876*, 611–620.

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

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

Wang, L. (2011). Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *JMLR*, 2269–2292.

Wegkap, M. (2007). Lasso type classifiers with a reject option. *Electronic Journal of Statistics*, *1*, 155–168.

Wiener, Y. (2013). *Theoretical Foundations of Selective Prediction*. Ph.D. thesis, Technion — Israel Institute of Technology.

Wiener, Y., & El Yaniv, R. (2012). Pointwise tracking the optimal regression function. In *Advances in Neural Information Processing Systems 25*, pp. 2051–2059.

Wiener, Y., Hanneke, S., & El-Yaniv, R. (2014). A compression technique for analyzing disagreement-based active learning. *arXiv preprint arXiv:1404.1504*.

Zhang, T. (2005). Data dependent concentration bounds for sequential prediction algorithms. In *Learning Theory*, pp. 173–187. Springer.