Journal of Machine Learning Research 11 (2010) 1605-1641 Submitted 10/09; Revised 4/10; Published 5/10

On the Foundations of Noise-free Selective Classification

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

Editor: Gabor Lugosi

Abstract

We consider *selective classification*, a term we adopt here to refer to 'classification with a reject option.' The essence in selective classification is to trade-off classifier coverage for higher accuracy. We term this trade-off the *risk-coverage (RC) trade-off*. Our main objective is to characterize this trade-off and to construct algorithms that can optimally or near optimally achieve the best possible trade-offs in a controlled manner. For noise-free models we present in this paper a thorough analysis of selective classification including characterizations of RC trade-offs in various interesting settings.

Keywords: classification with a reject option, selective classification, perfect learning, high performance classification, risk-coverage trade-off

1. Introduction

In this paper we study the trade-off between coverage and accuracy of classifiers with a reject option, a trade-off we refer to as the *risk-coverage (RC) trade-off*. Our main goal is to characterize this trade-off and to construct algorithms that can optimally or near optimally control it. Throughout the paper we use the term *selective classification* to refer to 'classification with a reject option.' Selective classification was introduced a number of decades ago and among the earliest studies are papers authored by Chow (1957, 1970), focusing on Bayesian solutions for the case where the underlying distributions are fully known. Through the years, selective classification continued to draw attention and numerous papers have been published. The attraction of effective selective classification is rather obvious in applications where one is not concerned with, or can afford partial coverage of the domain, and/or in cases where extremely low risk is a must but is not achievable in standard classification frameworks. Classification problems in medical diagnosis and in bioinformatics are often instances of such applications (Meltzer et al., 2001; Hanczar and Dougherty, 2008).

Despite the relatively large number of research publications on selective classification, the vast majority of these works have been concerned with implementing a reject option within specific learning schemes, by endowing a learning scheme (e.g., neural networks, SVMs) with a reject mechanism. Most of the reject mechanisms were based on "ambiguity" or (lack of) "confidence" principles: "when confused or when in doubt, refuse to classify." While there are many convincing accounts for the potential effectiveness of selective classification in reducing the risk, we are not familiar with a thorough or conclusive discussions on the relative power of the numerous rejection mechanisms that have been considered so far. The very few theoretical works that considered selective classification (see Section 10) do provide some risk or coverage bounds for specific schemes (e.g., ensemble methods) or learning principles (e.g., ERMs), but altogether characterizations of achievable (or non-achievable) RC trade-offs are absent in the current literature. In particular, the work done so far has not facilitated formal discussions of RC trade-off *optimality*.

A thorough understanding and effective use of selective classification requires characterization of the theoretical and practical boundaries of RC trade-offs, which are essential elements in any discussion of *optimality* in selective classification. These missing elements in the current literature are critical when constructing and exploring selective classification schemes and selective classification algorithms that aim at achieving optimality in controlling the RC trade-off.

One of our longer term goals is to provide such characterizations and introduce a notion of optimality for selective classification in the most general agnostic model. As a first step, however, we focus in this work on noiseless settings whereby a perfect hypothesis for the problem at hand exists (the so called "realizable case"). Moreover, we place special emphasis on the extreme case where zero risk has to be guaranteed. For this extreme case, which we call "perfect learning," we provide a thorough analysis that includes tight positive and negative results for the most general types of realizable settings (distribution independent, infinite hypothesis spaces). We also discuss some specific settings (linear classifiers, specific distribution families) and show an efficient algorithm for linear classifiers that achieves "perfect learning" with guaranteed coverage. Our results on "perfect learning" are instrumental in exploring entire RC trade-offs. Recalling known results on optimal standard realizable learning (no rejection is allowed), we show how to "interpolate" bounds and strategies for these two extreme cases (perfect learning and standard learning) so as to reveal upper and lower envelopes of optimal RC trade-offs.

2. Selective Classification: Preliminary Definitions

Let $\mathcal{X}$ be some feature space, for example, d-dimensional vectors in $\mathbb{R}^d$. In standard binary classification, the goal is to learn a binary classifier $f : \mathcal{X} \to \{\pm 1\}$, 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 $\mathcal{X} \times \{\pm 1\}$. We assume that the classifier is to be selected from a hypothesis space $\mathcal{F}$ and focus on the *realizable* setting where the labels are determined by some *unknown target hypothesis* $f^* \in \mathcal{F}$. Thus, it is assumed that P satisfies $\Pr_P(Y = f^*(X)|X) = 1$.

In *selective classification* the learner should output a binary *selective classifier* defined to be a pair (f,g), with f being a standard binary classifier, and $g : \mathcal{X} \to [0,1]$ a *selection function* whose meaning is as follows. When applying the selective classifier to a sample x, its output is:

$$(f,g)(x) \triangleq \begin{cases} reject, & \text{w.p. } 1 - g(x); \\ f(x), & \text{w.p. } g(x). \end{cases} \tag{1}$$

Thus, in its most general form, the selective classifier is *randomized*. Whenever the selection function is a zero-one rule, $g : \mathcal{X} \to \{0,1\}$, we say that the selective classifier is deterministic. Note that "standard learning" (i.e., no rejection is allowed) is the special case of selective classification where $g(x)$ selects all points (i.e., $g(x) \equiv 1$).

The two main characteristics of a selective classifier are its *coverage* and its *risk* (or "true error").

Definition 1 (coverage) *The* coverage *of a selective classifier (f,g) is the mean value of the selection function $g(X)$ taken over the underlying distribution P,*

$$\Phi(f,g) \triangleq \mathbf{E}\left[g(X)\right].$$

Definition 2 (risk) *For a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, we define the risk of a selective classifier (f,g) as the average loss on the accepted samples,*

$$R(f,g) \triangleq \frac{\mathbf{E}\left[\ell(f(X),Y) \cdot g(X)\right]}{\Phi(f,g)}.$$

This risk definition clearly reduces to the standard definition of risk if $g(x) \equiv 1$. Note that (at the outset) both the coverage and risk are *unknown quantities* because they are defined in terms of the unknown underlying distribution P.

We define a learning algorithm ALG to be a (random) function that, given a sample S_m, chooses a selective classifier (f,g). We evaluate learners with respect to their coverage and risk and derive both positive and negative results on achievable risk and coverage. Our model is a slight extension of the standard minimax model for standard statistical learning as described, for example, by Antos and Lugosi (1998). Thus, we consider the following game between the learner and an adversary. The parameters of the game are a domain $\mathcal{X}$ and an hypothesis class $\mathcal{F}$.

1. A tolerance level δ and a training sample size m are given.
2. The learner chooses a learning algorithm ALG.
3. With full knowledge of the learner's choice, the adversary chooses a distribution $P(X)$ over $\mathcal{X}$, and a target hypothesis $f^* \in \mathcal{F}$ (or a distribution over $\mathcal{F}$ according to which f^* is selected).
4. A training sample S_m is drawn i.i.d. according to P and f^*.
5. ALG is applied on S_m and outputs a selective classifier (f,g).

The result of the game is evaluated in terms of the risk and coverage obtained by the chosen selective classifier and clearly, these are random quantities that trade-off each other. A *positive result* in this model is a pair of bounds, $B_R = B_R(\mathcal{F},\delta,m)$ and $B_\Phi = B_\Phi(\mathcal{F},\delta,m)$, for risk and coverage, respectively, that for any δ and m, hold with high probability, of at least $1-\delta$ for *any* distribution P; namely,

$$\Pr\{R(f,g) \leq B_R \ \wedge \ \Phi(f,g) \geq B_\Phi\} \geq 1-\delta.$$

The probability is taken w.r.t. the random choice of training samples S_m, as well as w.r.t. all other random choices introduced, such as a random choice of f^* by the adversary (if applicable), a random choice of (f,g) by ALG (if applicable), and the randomized selection function (Equation (1)).

A *negative result* is a probabilistic statement on the impossibility of any positive result. Thus, in its most general form a negative result is a pair of bounds B_R and B_Φ that, for any δ, satisfy

$$\Pr\{R(f,g) \geq B_R \ \vee \ \Phi(f,g) \leq B_\Phi\} \geq \delta,$$

for *some* probability P. Here again, probability is taken w.r.t. the random choice of the training samples S_m, as well as w.r.t. all other random choices.

For a selective classifier (f,g) with coverage $\Phi(f,g)$ we can specify a Risk-Coverage (RC) trade-off as a bound on the risk $R(f,g)$, expressed in terms of $\Phi(f,g)$. Thus, a *positive result on the RC trade-off* is a probabilistic statement of the following form

$$\Pr\{R(f,g) \leq B(\Phi(f,g),\delta,m)\} \geq 1-\delta.$$

Similarly, a *negative result on the RC trade-off* is a statement of the form,

$$\Pr\{R(f,g) \geq B(\Phi(f,g),\delta,m)\} \geq \delta.$$

Clearly, all results (positive and negative) are qualified by the model parameters, namely the domain $\mathcal{X}$ and the hypothesis space $\mathcal{F}$, and the quality/generality of a result should be assessed w.r.t. generality of these parameters. An additional major consideration is, of course, the computational complexity of the learning algorithm.

Finally, in the sequel we rely on the following standard definition of the version space (Mitchell, 1977).

Definition 3 (version space) *Given an hypothesis class $\mathcal{F}$ and a training sample S_m, the version space $VS_{\mathcal{F},S_m}$ is the set of all hypotheses in $\mathcal{F}$ that classify S_m correctly.*

3. Contributions

The purpose of this section is to provide a high level technical overview of our contributions. Using a training sample S_m, the goal in selective classification is to output a selective classifier (f,g) that has sufficiently low risk with sufficiently high coverage. Obviously, these two quantities trade-off each other. We call the trade-off between risk and coverage the *risk-coverage (RC) trade-off*. The best way to benefit from selective classification is to *control* the creation of the classifier so as to meet a prescribed error/coverage specification along the RC trade-off. For example, it might be desirable to devise a learning system that will receive as input an error constraint (say, 2% error) and, based on a finite (and small) training sample, will be capable of generating a classifier whose ensured test error (w.h.p.) is not larger than 2%, while having the maximum possible coverage of the domain. If the RC trade-off is revealed, it is possible to know if the 2% error constraint can be met and what would be the corresponding coverage.

In Figure 1 we schematically depict elements of the RC trade-off. The x-axis measures risk (error in the case of the $0/1$ loss) and the y-axis is coverage. The entire region depicted, called the *RC plane*, consisting of all (r,c) points in the rectangle of interest, where r is a risk (error) coordinate and c is a coverage coordinate. Assume a fixed problem setting (including an unknown underlying distribution P, m training examples drawn i.i.d. from P, an hypothesis space $\mathcal{F}$ and a tolerance parameter δ). To fully characterize the RC trade-off we need to determine for each point (r,c) on the RC plane if it is (efficiently) “achievable.” We say that (r,c) is (efficiently) *achievable* if there is an (efficient) learning algorithm that will output a selective classifier (f,g) such that with probability of at least $1-\delta$, its coverage is at least c and its risk is at most r.

Notice that point r^* (the coordinate $(r^*,1)$) where the coverage is 1 represents “standard learning.” At this point we require full coverage with certainty and the achievable risk represents the lowest possible risk in our fixed setting (which should be achievable with probability of at least $1-\delta$). Point r^* represent one extreme of the RC trade-off. The other extreme of the RC trade-off is point c^*, where we require zero risk *with certainty*. The coverage at c^* is the optimal (highest possible) in our setting when zero error is required. We call point c^* *perfect learning* because achievable perfect learning means that we can generate a classifier that never errs with certainty for the problem at hand. Note that at the outset, it is not at all clear if non-trivial perfect learning (with guaranteed positive coverage) can be accomplished.

The full RC trade-off is some (unknown) curve connecting points c^* and r^*. This curve passes somewhere in the zone labeled with a question mark and represents optimal selective classification.

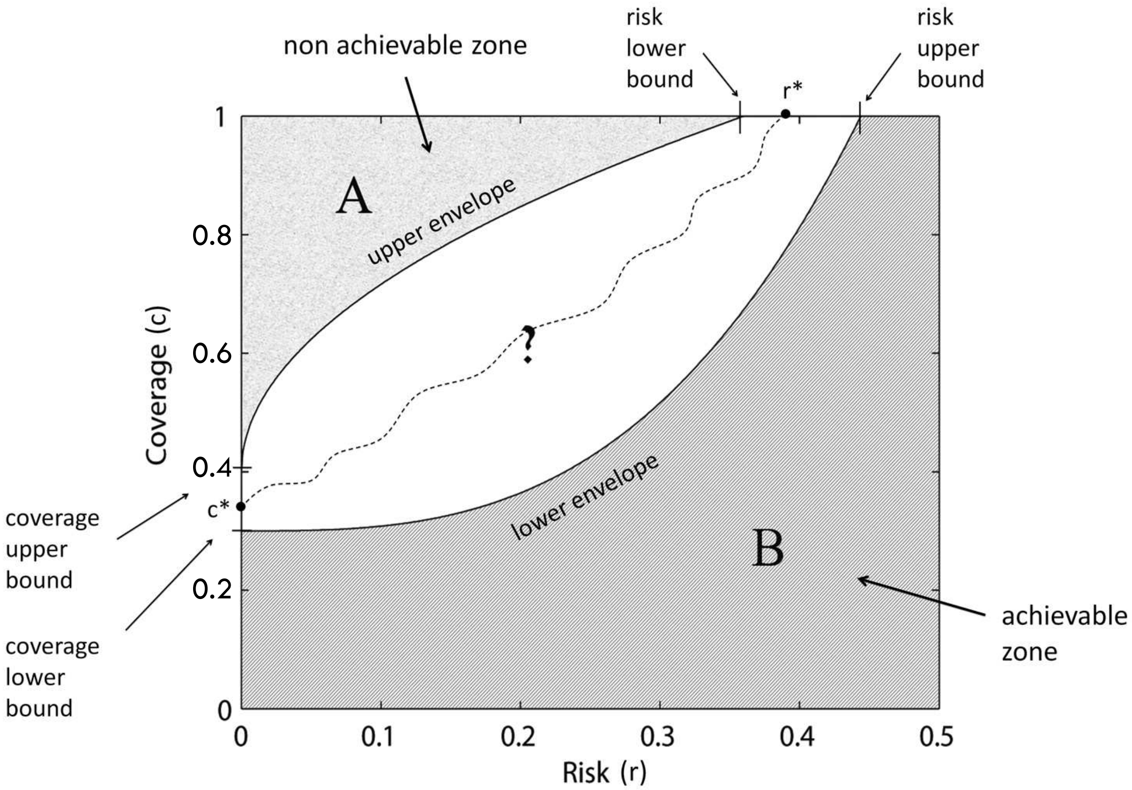

Figure 1: The RC plane and RC trade-off

Points above this curve (e.g., at zone A) are not achievable. Points below this curve (e.g., at zone B) are achievable. One of the main goals of this paper is to study the RC curve and provide as tight as possible boundaries for it. To this end we characterize upper and lower envelopes of the RC curve as schematically depicted in Figure 1. The upper envelop is a boundary of a "non-achievable zone" (zone A) and therefore we consider any upper envelop as a "negative result." The lower envelop is a boundary of an "achievable zone" (zone B) and is therefore considered as a "positive result." Note that upper and lower envelopes, as depicted in the figure, represent two different things, which are formally defined in Section 2 as probabilistic statements on possibility and impossibility.

Point r^* on the RC curve ("standard learning") was extensively studied in the literature. Perfect learning (point c^*) was never considered. For the most part, the existing work on selective classification exhibited (either empirically or theoretically) specific but anecdotal points or curves in the achievable zone (B) but, to the best of our knowledge no systematic attempts were ever made to characterize the RC-curve, which corresponds to *optimal selective classification*. In particular, there are currently no "negative" results attempting to characterize non achievable zones in the RC plane.

Our technical exposition begins by focusing on perfect learning (point c^* in the RC plane). Given the training set S_m, we are required to generate a "perfect" selective classifier (f,g) for which it is known *with certainty* that $R(f,g)=0$.[1] Obviously, zero risk is trivially achieved by taking g that rejects the entire input space $\mathcal{X}$. But is it possible to achieve perfect learning on a guaranteed fraction of the effective volume of $\mathcal{X}$?

Our first observation is Theorem 8, stating that for any finite hypothesis class $\mathcal{F}$, perfect learning with guaranteed coverage is achievable by a particular selective classification strategy. For any tolerance δ, with probability of at least $1-\delta$, it is guaranteed that the coverage achieved by this strategy will be at least

$$1-\frac{1}{m}O(|\mathcal{F}|+\ln(1/\delta)). \tag{2}$$

The learning strategy that achieves this performance is simple and natural and can be termed *consistent selective strategy (CSS)*: take f to be any hypothesis from the version space (with respect to S_m), and construct a g that deterministically rejects any point that is not classified unanimously by all version space hypotheses. This CSS strategy is optimal for perfect learning. We show in Theorem 7 that any other strategy that achieves perfect learning cannot have larger coverage than CSS. It is interesting to note that the optimal selection function is not obtained by thresholding soft classification values, which is the commonly used heuristic.

It is easy to see why the classifier (f,g) selected by CSS has zero risk with certainty. Since f^* is assumed to be in the version space, and since g rejects all instances that are not classified unanimously by all the hypotheses in the version space, any selection of f from the version space will have identical classification to f^*. Nonetheless, it is surprising at the outset that the selection function g doesn't reject a lot and in fact, its rejection rate can be very small for sufficiently large m as it decreases at rate $1/m$.

This distribution-free coverage guarantee (2) is proven to be nearly tight for CSS and therefore, it is the best possible bound for any selective learner. Specifically, as shown in Theorem 11, there exist a particular finite hypothesis class and a particular underlying distribution for which a matching negative result (up to multiplicative constants) holds for any *consistent selective* learner. This result is readily extended to any selective learner by the CSS coverage optimality of Theorem 7.

What about infinite hypothesis spaces? We show in Theorem 14 that it is impossible to provide any coverage guarantees for perfect learning, in the general case. Specifically, for linear classifiers, we show a bad distribution for which any selective learner ensuring zero risk will be forced to reject the entire volume of $\mathcal{X}$, thus failing to guarantee more than zero coverage. Thus, in the general case, point c^* is simply the coordinate $(0,0)$ on the RC plane. The implication of this result is that when aiming at very small risks, the rejection rate might in general be very high (very small coverage), which may be unacceptable in many applications.

So the bad news is that perfect learning with guaranteed coverage cannot in general be achieved if the hypothesis space is infinite. Fortunately, however, this observation does not preclude non-trivial perfect learning in less adverse situations. What can be accomplished are both data-dependent and distribution-dependent guarantees. For any selective hypothesis (f,g), that is consistent with a sample S_m, Theorem 21 ensures perfect learning with a high probability coverage guarantee of the following form:

$$\Phi(f,g)\geq 1-\frac{1}{m}O\left(\gamma(\mathcal{F},\hat{n})\ln\frac{m}{\gamma(\mathcal{F},\hat{n})}+\ln\frac{m}{\delta}\right), \tag{3}$$

[1] The requirement that in perfect learning the risk is zero *with certainty* is dual to the requirement that the coverage is 100% *with certainty* in standard learning.

where $\hat{n}$ is a new empirical quantity measuring the version space compression set size (see Definition 15), and $\gamma(\mathcal{F},k)$ is a new complexity measure of the set of all possible version spaces generated by training samples of size k. We call $\gamma(\mathcal{F},k)$ the "order-k characterizing set complexity" of $\mathcal{F}$, and it is derived using VC-dimension arguments (see Definition 18).

This general data-dependent bound is then applied to linear classifiers. Relying on a classical result in combinatorial geometry that bounds the number of facets of polytopes, we derive in Theorem 27 the following upper bound on the order-k characterizing set complexity of linear classifiers in $\mathbb{R}^d$,

$$\gamma(\mathcal{F},k) \leq O(d^3(k/d)^{d/2}\log k).$$

Plugging this bound to (3) results in a data-dependent compression bound for linear classifiers in terms of $\hat{n}$, the size of the compression set of the version space.

We then consider the evaluation of the compression set size $\hat{n}$ for specific distributions. Using a classical result in geometric probability theory on the average number of maximal random vectors, we show in Lemma 32 that if the underlying distribution is any (unknown) finite mixture of arbitrary multi-dimensional Gaussians in $\mathbb{R}^d$, then the compression set size of the version space obtained using m labeled examples satisfies, with probability of at least $1-\delta$,

$$\hat{n} = O\left((\log m)^d/\delta\right).$$

This bound immediately yields a coverage guarantee for perfect learning of linear classifiers, as stated in Corollary 33. This is a powerful result providing strong indication on the potential effectiveness of perfect learning with guaranteed coverage in a variety of applications.

In Section 7 we derive upper and lower envelopes for the RC curve. Our results on perfect learning described above play a major role in the derivation of these envelopes. We generalize the CSS strategy and define a "controllable selective strategy" (Definition 34). This strategy is parameterized by a number $\alpha \in [0,1]$ which controls the rejection rate by interpolating perfect learning and optimal standard learning. In particular, this strategy, applied with $\alpha = 0$ is perfect learning, and with $\alpha = 1$ it is optimal standard learning (full coverage), which in the realizable case is known to be achieved by any consistent learner. For any finite hypothesis space, the lower envelop we present in Theorem 36 is

$$R_\alpha(f,g) \leq \left(\frac{1-\Phi_0/\Phi_\alpha(f,g)}{1-\Phi_0}\right) \cdot \frac{1}{m}\left(\ln|\mathcal{F}| + \ln\frac{2}{\delta}\right),$$

where Φ_0 is the coverage guarantee of perfect learning in Equation (2), $R_\alpha(f,g)$ is the risk of the "controllable selective strategy" with control parameter α, and $\Phi_\alpha(f,g)$ is the matching coverage.

The upper envelop on the RC curve is then derived in Theorem 37 for any selective classifier (f,g) by constructing a particular bad distribution for which

$$R(f,g) \geq \frac{1}{4\Phi} \cdot \min\left(2\Phi - 1, 2\Phi - 2 + \frac{1}{4m} \cdot \left[VCdim(\mathcal{F}) - \frac{16}{3}\ln\frac{1}{1-2\delta}\right]\right).$$

An exact implementation of the CSS strategy appears as if it should be computationally difficult. Given a particular training set, CSS must reject a point iff it is not classified the same by all hypotheses in the current version space. In Section 8 we show an efficient algorithm that implements CSS of linear classifiers. The main idea leading to this construction is the following observation. Given a test point x we examine if the inclusion of x with either positive or negative labels in the training sets results in linearly separable sets. Clearly, CSS must reject x iff both these augmented sets are linearly separable. Thus, the construction of the CSS selection function can be reduced to two tests of linear separability, which can be efficiently accomplished using known techniques for testing linear separability. Note that we do not construct the selection function explicitly during the training process, a task that may indeed require intensive computation. Rather, we benefit from a "lazy learning" approach whereby the selection function is constructed at test time per each example (as in nearest neighbor algorithms).

4. Perfect Learning with Finite Hypothesis Spaces

In this section we consider the simplest case of realizable learning with a finite hypothesis space $\mathcal{F}$. We show that perfect selective classification with guaranteed coverage is achievable (from a learning-theoretic perspective) by a learning strategy termed *consistent selective strategy (CSS)*. Moreover, CSS is shown to be optimal in its coverage rate, which is fully characterized by providing lower and upper bounds that match in their asymptotic behavior in the sample size m. We start by defining a region in $\mathcal{X}$, which is termed the "maximal agreement set." Any hypothesis that is consistent with the sample S_m is guaranteed to be consistent with the target hypothesis f^* on this entire region.

Definition 4 (agreement set) *Let $\mathcal{G} \subseteq \mathcal{F}$. A subset $\mathcal{X}' \subseteq \mathcal{X}$ is an* agreement set *with respect to $\mathcal{G}$ if all hypotheses in $\mathcal{G}$ agree on every instance in $\mathcal{X}'$, namely,*

$$\forall \; g_1, g_2 \in \mathcal{G}, \;\; x \in \mathcal{X}', \quad g_1(x) = g_2(x).$$

Definition 5 (maximal agreement set) *Let $\mathcal{G} \subseteq \mathcal{F}$. The* maximal agreement set *with respect to $\mathcal{G}$ is the union of all agreement sets with respect to $\mathcal{G}$.*

Recall that the version space $VS_{\mathcal{F},S_m} \subseteq \mathcal{F}$ is the set of all hypotheses that classify S_m correctly (Definition 3).

Definition 6 (consistent selective strategy (CSS)) *Given S_m, a* consistent selective strategy (CSS) *is a selective classification strategy that takes f to be any hypothesis in $VS_{\mathcal{F},S_m}$ (i.e., a consistent learner), and takes a (deterministic) selection function g that equals one for all points in the maximal agreement set with respect to $VS_{\mathcal{F},S_m}$, and zero otherwise.*

Recall that the (unknown) labeling hypothesis f^* is in $VS_{\mathcal{F},S_m}$. Thus, CSS simply rejects all points that might incur an error with respect to f^*. An immediate consequence is that any CSS selective hypothesis (f,g) always satisfies $R(f,g) = 0$. The main concern, however, is whether its coverage $\Phi(f,g)$ can be bounded from below and whether any other strategy that achieves perfect learning with certainty can achieve better coverage. The following theorem proves that CSS has the largest possible coverage among all strategies.

Theorem 7 (CSS coverage optimality) *Given S_m, let (f,g) be a selective classifier chosen by any strategy that ensures zero risk with certainty for* any *unknown distribution P and* any *target concept $f^* \in \mathcal{F}$. Let (f_c, g_c) be a selective classifier selected by CSS using S_m. Then, $\Phi(f,g) \leq \Phi(f_c, g_c)$.*

Proof For the sake of simplicity we limit the discussion to deterministic strategies. The extension to stochastic strategies is omitted but is straightforward. Given a hypothetical sample $\tilde{S}_m$ of size m, let $(\tilde{f}_c, \tilde{g}_c)$ be the selective classifier chosen by CSS and let $(\tilde{f}, \tilde{g})$ be the selective classifier chosen by any competing strategy. Assume that there exists $x_0 \in \mathcal{X}$ ($x_0 \notin \tilde{S}_m$) such that $\tilde{g}(x_0) = 1$ and $\tilde{g}_c(x_0) = 0$. According to the CSS construction of $\tilde{g}_c$, since $\tilde{g}_c(x_0) = 0$, there are at least two hypotheses $h_1, h_2 \in VS_{\mathcal{F},\tilde{S}_m}$ such that $h_1(x_0) \neq h_2(x_0)$. Assume, without loss of generality, that $h_1(x_0) = \tilde{f}(x_0)$. We will now construct a new "imaginary" classification problem and show that, under the above assumption, the competing strategy fails to guarantee zero risk with certainty. Let the imaginary target concept f'^* be h_2 and the imaginary underlying distribution P' be

$$P'(x) = \begin{cases} (1-\varepsilon)/m, & \text{if } x \in \tilde{S}_m; \\ \varepsilon, & \text{if } x = x_0; \\ 0, & \text{otherwise.} \end{cases}$$

Imagine a random sample S'_m drawn i.i.d from P'. There is a positive (perhaps small) probability that S'_m will equal $\tilde{S}_m$, in which case $(f', g') = (\tilde{f}, \tilde{g})$. Since $g'(x_0) = \tilde{g}(x_0) = 1$ and $f^*(x_0) \neq f'(x_0)$, with positive probability $R(f', g') = \varepsilon > 0$. Contradiction to the assumption that the competing strategy achieves perfect learning with certainty. It follows that for any sample $\tilde{S}_m$ and for any $x \in \mathcal{X}$, if $\tilde{g}(x) = 1$ then $\tilde{g}_c(x) = 1$. Consequently, for any unknown distribution P, $\Phi(\tilde{f}, \tilde{g}) \leq \Phi(\tilde{f}_c, \tilde{g}_c)$. ■

The next result establishes the existence of perfect learning with guaranteed coverage in the finite case.

Theorem 8 (guaranteed coverage) *Assume a finite $\mathcal{F}$ and let (f,g) be a selective classifier selected by CSS. Then, $R(f,g) = 0$ and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(f,g) \geq 1 - \frac{1}{m}\left((\ln 2)\min\{|\mathcal{F}|, |\mathcal{X}|\} + \ln\frac{1}{\delta}\right). \tag{4}$$

Proof For any ε, let $G_1, G_2, \ldots, G_k$, be all the hypothesis subsets of $\mathcal{F}$ with corresponding maximal agreement sets, $\lambda_1, \lambda_2, \ldots, \lambda_k$, such that each λ_i has volume of at most $1-\varepsilon$ with respect to P. For any $1 \leq i \leq k$, the probability that a single point will be randomly drawn from λ_i is thus at most $1-\varepsilon$. The probability that all training points will be drawn from λ_i is therefore at most $(1-\varepsilon)^m$. If a training point x is in $\mathcal{X} \setminus \lambda_i$, then there are at least two hypotheses $f_1, f_2 \in G_i$ that do not agree on x. Hence,

$$\Pr_P(G_i \subseteq VS_{\mathcal{F},S_m}) \leq (1-\varepsilon)^m.$$

We note that

$$k \leq 2^{\min\{|\mathcal{F}|,|\mathcal{X}|\}},$$

and by the union bound,

$$\Pr_P(\exists G_i \quad G_i \subseteq VS_{\mathcal{F},S_m}) \leq k \cdot (1-\varepsilon)^m \leq 2^{\min\{|\mathcal{F}|,|\mathcal{X}|\}} \cdot (1-\varepsilon)^m.$$

Therefore, with probability of at least $1 - 2^{\min\{|\mathcal{F}|,|\mathcal{X}|\}} \cdot (1-\varepsilon)^m$, the version space $VS_{\mathcal{F},S_m}$ differs from any subset G_i, and hence it has a maximal agreement set with volume of *at least* $1-\varepsilon$. Using the inequality $1-\varepsilon \leq \exp(-\varepsilon)$, we have

$$2^{\min\{|\mathcal{F}|,|\mathcal{X}|\}} \cdot (1-\varepsilon)^m \leq 2^{\min\{|\mathcal{F}|,|\mathcal{X}|\}} \cdot \exp(-m\varepsilon).$$

Equating the right-hand side to δ and solving for ε completes the proof. ■

A leading term in the coverage guarantee (4) is $|\mathcal{F}|$. In corresponding results in standard consistent learning (Haussler, 1988) the corresponding term is $\log|\mathcal{F}|$. This may raise a concern on the tightness of (4). However, as shown in Corollary 13, this bound is tight (up to multiplicative constants). To prove the Corollary we will require the following two definitions.

Definition 9 (binomial tail distribution) *Let $Z_1, Z_2, \ldots Z_m$ be m independent Bernoulli random variables each with a success probability p. Then for any $0 \le k \le m$ we define*

$$Bin(m,k,p) \triangleq \Pr\left(\sum_{i=1}^{m} Z_i \le k\right).$$

Definition 10 (binomial tail inversion, Langford, 2005) *For any $0 \le \delta \le 1$ we define*

$$\overline{Bin}(m,k,\delta) \triangleq \max_p \{p : Bin(m,k,p) \ge \delta\}.$$

Theorem 11 (non-achievable coverage, implicit bound) *Let $0 \le \delta \le \frac{1}{2}$, m, and $n > 1$ be given. There exist a distribution P, that depends on m and n, and a finite hypothesis class $\mathcal{F}$ of size n, such that for any selective classifier (f,g), chosen from $\mathcal{F}$ by CSS (so $R(f,g) = 0$) using a training sample S_m drawn i.i.d. according to P, with probability of at least δ,*

$$\Phi(f,g) \le 1 - \frac{1}{2} \cdot \overline{Bin}\left(m, \frac{|\mathcal{F}|}{2}, 2\delta\right).$$

Proof Let $\mathcal{X} \triangleq \{e_1, e_2, \ldots e_{n+1}\}$ be the standard (vector) basis of $\mathbb{R}^{n+1}$, $\mathcal{X}' \triangleq \mathcal{X} \setminus \{e_{n+1}\}$ and P be the source distribution over $\mathcal{X}$ satisfying

$$P(e_i) \triangleq \begin{cases} \overline{Bin}\left(m, \frac{n}{2}, 2\delta\right)/n, & \text{if } i \le n; \\ 1 - \overline{Bin}\left(m, \frac{n}{2}, 2\delta\right), & \text{otherwise;} \end{cases}$$

where $\overline{Bin}(m,k,\delta)$ is the binomial tail inversion (Definition 10). Since

$$\overline{Bin}\left(m, \frac{n}{2}, 2\delta\right) \triangleq \max_p \left\{p : Bin\left(m, \frac{n}{2}, p\right) \ge 2\delta\right\},$$

and S_m is drawn i.i.d. according to P, we get that with probability of at least 2δ,

$$\left|\left\{x \in S_m : x \in \mathcal{X}'\right\}\right| \le \frac{n}{2}.$$

Let $\mathcal{F}$ be the class of singletons such that

$$f_i(e_j) \triangleq \begin{cases} 1, & \text{if } i = j; \\ -1, & \text{otherwise.} \end{cases}$$

Taking $f^* \triangleq f_{i^*}$, for some $1 \leq i^* \leq n$, we have,

$$\begin{aligned}
& \Pr\left(e_{i^*} \notin S_m, \left|\left\{x \in S_m : x \in \mathcal{X}'\right\}\right| \leq \frac{n}{2}\right) \\
= \ & \Pr\left(e_{i^*} \notin S_m \,\middle|\, \left|\left\{x \in S_m : x \in \mathcal{X}'\right\}\right| \leq \frac{n}{2}\right) \cdot \Pr\left(\left|\left\{x \in S_m : x \in \mathcal{X}'\right\}\right| \leq \frac{n}{2}\right) \\
\geq \ & \left(1 - \frac{1}{n}\right)^{\frac{n}{2}} \cdot 2\delta \geq \delta.
\end{aligned}$$

If $e_{i^*} \notin S_m$ then all samples in S_m are negative, so each sample in $\mathcal{X}'$ can reduce the version space $VS_{\mathcal{F},S_m}$ by at most one hypothesis. Hence, with probability of at least δ,

$$|VS_{\mathcal{F},S_m}| \geq |\mathcal{F}| - \frac{n}{2} = \frac{n}{2}.$$

Since the coverage $\Phi(f,g)$ is the volume of the maximal agreement set with respect to the version space $VS_{\mathcal{F},S_m}$, it follows that

$$\Phi(f,g) \quad = \quad 1 - |VS_{\mathcal{F},S_m}| \cdot \frac{\overline{Bin}\left(m, \frac{n}{2}, 2\delta\right)}{n} \leq 1 - \frac{1}{2} \cdot \overline{Bin}\left(m, \frac{|\mathcal{F}|}{2}, 2\delta\right).$$

■

Remark 12 *The result of Theorem 11 is based on the use of the class of singletons. Augmenting this class by the empty set and choosing a uniform distribution over $\mathcal{X}$ results in a tighter bound. However, the bound will be significantly less general as it will hold only for a single hypothesis in $\mathcal{F}$ and not for any hypothesis in $\mathcal{F}$.*

Corollary 13 (non-achievable coverage, explicit bound) *Let $0 \leq \delta \leq \frac{1}{4}$, m, and $n > 1$ be given. There exist a distribution P, that depends on m and n, and a finite hypothesis class $\mathcal{F}$ of size n, such that for any selective classifier (f,g), chosen from $\mathcal{F}$ by CSS (so $R(f,g) = 0$) using a training sample S_m drawn i.i.d. according to P, with probability of at least δ,*

$$\Phi(f,g) \leq \max\left\{0, 1 - \frac{1}{8m}\left(|\mathcal{F}| - \frac{16}{3}\ln\frac{1}{1-2\delta}\right)\right\}.$$

Proof Applying Lemma 43 we get

$$\overline{Bin}\left(m, \frac{|\mathcal{F}|}{2}, 2\delta\right) \geq \min\left\{1, \frac{|\mathcal{F}|}{4m} - \frac{4}{3m}\ln\frac{1}{1-2\delta}\right\}.$$

Applying Theorem 11 completes the proof. ■

5. Consistent Selective Classification Over Infinite Hypothesis Spaces

In this section we consider an infinite hypothesis space $\mathcal{F}$. We show that in the general case, perfect selective classification with guaranteed (non-zero) coverage is not achievable even when $\mathcal{F}$ has a finite VC-dimension. We then derive a meaningful coverage guarantee using posterior information on the source distribution (data-dependent bound).

We start this section with a negative result that precludes non-trivial perfect learning when $\mathcal{F}$ is the set of linear classifiers. The result is obtained by constructing a particularly bad distribution.

Theorem 14 (non-achievable coverage) *Let m and $d > 2$ be given. There exist a distribution P, an infinite hypothesis class $\mathcal{F}$ with a finite VC-dimension d, and a target hypothesis in $\mathcal{F}$, such that $\Phi(f,g) = 0$ for any selective classifier (f,g), chosen from $\mathcal{F}$ by CSS using a training sample S_m drawn i.i.d. according to P.*

Proof Let $\mathcal{F}$ be the class of all linear classifiers in $\mathbb{R}^2$ and let P be a uniform distribution over the arcs,

$$(x-2)^2 + y^2 = 2, \quad x < 1,$$

and

$$(x+2)^2 + y^2 = 2, \quad x > -1.$$

Figure 2 depicts this construction. The training set S_m consists of points on these arcs, labeled by any linear classifier that passes between the arcs. The maximal agreement set, A, with respect to the version space $VS_{\mathcal{F},S_m}$ is partitioned into two subsets A^+ and A^- according to the labels obtained by hypotheses in the version space. Clearly, A^+ is confined by a polygon whose vertices lie on the right-hand side arc. Since P is concentrated on the arc, the probability volume of A^+ is exactly zero for any finite m. The same analysis holds for A^-, and therefore the coverage is forced to be zero. The VC-dimension of the class of all linear classifiers in $\mathbb{R}^2$ is 3. Embedding the distribution P in a higher dimensional space $\mathbb{R}^d$ and using the class of all linear classifiers in $\mathbb{R}^d$ completes the proof. ■

A direct corollary of Theorem 14 is that, in the general case, perfect selective classification with distribution-free guaranteed coverage is not achievable for infinite hypothesis spaces. However, this is certainly not the end of the story for perfect learning. In the remainder of this paper we derive meaningful coverage guarantees using posterior or prior information on the source distribution (data- and distribution-dependent bounds).

In order to guarantee meaningful coverage we first need to study the complexity of the selection function $g(x)$ chosen by CSS. The complexity of the classification function $f(x)$ is determined only by the hypothesis class $\mathcal{F}$ and it is independent of the sample size itself. However, the complexity of $g(x)$ (the maximal agreement set) chosen by CSS generally depends on the sample size. Therefore, increasing the training sample size does not necessarily guarantee non-trivial coverage. Our main task is to find the complexity class of the family of maximal agreement sets from which $g(x)$ is chosen. Let us define the family of all maximal agreement sets as $\mathcal{H} = \bigcup \mathcal{H}_n$ such that $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \subset \ldots$. We can now exploit the fact that CSS chooses a maximal agreement set that belongs to a specific subclass $\mathcal{H}_n$ with a complexity measured in terms of the VC dimension of $\mathcal{H}_n$. We term this approach *Structural Coverage Maximization (SCM)* following the analogous and familiar *Structural*

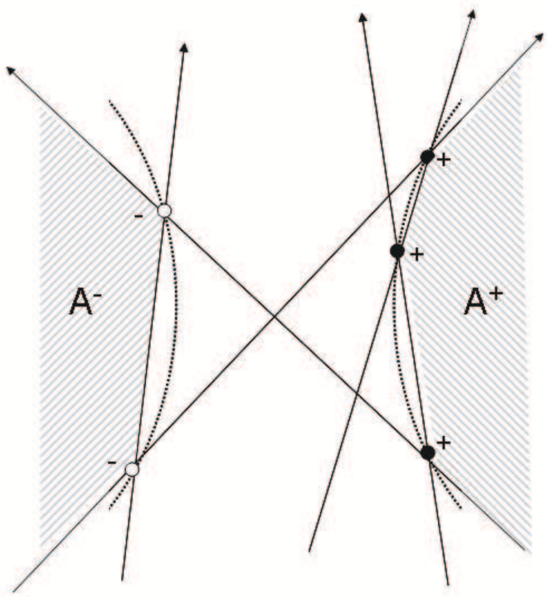

Figure 2: A worst-case distribution for linear classifiers: points are drawn uniformly at random on the two arcs and labeled by a linear classifier that passes between these arcs. The probability volume of the maximal agreement set is zero.

Risk Minimization (SRM) approach (Vapnik, 1998). A useful way to parameterize $\mathcal{H}$ is to use the size of the "version space compression set" (Definition 15).

Definition 15 (version space compression set) *Let S_m be a labeled sample of m points and let $VS_{\mathcal{F},S_m}$ be the induced version space. The* version space compression set, *$S_{\hat{n}} \subseteq S_m$ is a smallest subset of S_m satisfying $VS_{\mathcal{F},S_m} = VS_{\mathcal{F},S_{\hat{n}}}$. Note that for any given $\mathcal{F}$ and S_m, the size of the version space compression set, denoted $\hat{n} = \hat{n}(\mathcal{F},S_m)$, is unique.*

Since a maximal agreement set is a region in $\mathcal{X}$, rather than an hypothesis, we formally define the dual hypothesis that matches every maximal agreement set.

Definition 16 (characterizing hypothesis) *Let $\mathcal{G} \subseteq \mathcal{F}$ and let $A_{\mathcal{G}}$ be the maximal agreement set with respect to $\mathcal{G}$. The* characterizing hypothesis *of $\mathcal{G}$, $f_{\mathcal{G}}(x)$ is a binary hypothesis over $\mathcal{X}$ obtaining positive values over $A_{\mathcal{G}}$ and zero otherwise.*

We are now ready to formally define $\mathcal{H}_n$, a class we term "order-*n* characterizing set."

Definition 17 (order-*n* characterizing set) *For each n, let $\mathcal{S}_n$ be the set of all possible labeled samples of size n (all n-subsets, each with all possible labelings). The order-n* characterizing set *of $\mathcal{F}$, denoted $\mathcal{H}_n$, is the set of all characterizing hypotheses $f_{\mathcal{G}}(x)$, where $\mathcal{G} \subseteq \mathcal{F}$ is a version space induced by some member of $\mathcal{S}_n$.*

Definition 18 (characterizing set complexity) *Let $\mathcal{H}_n$ be the order-n characterizing set of $\mathcal{F}$. The order-n* characterizing set complexity *of $\mathcal{F}$, denoted $\gamma(\mathcal{F},n)$, is the VC-dimension of $\mathcal{H}_n$.*

Lemma 19 *The characterizing hypothesis $f_{VS_{\mathcal{F},S_m}}(x)$ belongs to the order-$\hat{n}$ characterizing set of $\mathcal{F}$, where $\hat{n} = \hat{n}(\mathcal{F}, S_m)$ is the size of the version space compression set.*

Proof According to Definition 15, there exists a subset $S_{\hat{n}} \subset S_m$ of size $\hat{n}$ such that $VS_{\mathcal{F},S_m} = VS_{\mathcal{F},S_{\hat{n}}}$. The rest of the proof follows immediately from Definition 17. ■

Before stating the main result of this section, we state a classical result that will be used later.

Theorem 20 (Vapnik and Chervonenkis, 1971; Anthony and Bartlett, 1999, p.53) *Let $\mathcal{F}$ be a hypothesis space with VC-dimension h. For any probability distribution P on $\mathcal{X} \times \{\pm 1\}$, with probability of at least $1-\delta$ over the choice of S_m from P^m, any hypothesis $f \in \mathcal{F}$ consistent with S_m satisfies*

$$R(f) \leq \varepsilon(h,m,\delta) = \frac{2}{m}\left[h \ln \frac{2em}{h} + \ln \frac{2}{\delta}\right], \tag{5}$$

where $R(f) \triangleq \mathbf{E}\left[\mathbb{I}(f(x) \neq f^*(x))\right]$ is the risk of f.

We note that inequality (5) actually holds only for $h \leq m$. For any $h > m$ it is clear that no meaningful upper bound on the risk can be achieved. It is easy to fix the inequality for the general case by replacing $\ln\left(\frac{2em}{h}\right)$ by $\ln_+\left(\frac{2em}{h}\right)$, where $ln_+(x) \triangleq \max(\ln(x), 1)$.

Theorem 21 (data-dependent coverage guarantee) *For any m, let $a_1, a_2, \ldots, a_m \in \mathbb{R}$ be given, such that $a_i \geq 0$ and $\sum_{i=1}^m a_i \leq 1$. Let (f,g) be a selective CSS classifier. Then, $R(f,g) = 0$, and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(f,g) \geq 1 - \frac{2}{m}\left[\gamma(\mathcal{F},\hat{n}) \ln_+\left(\frac{2em}{\gamma(\mathcal{F},\hat{n})}\right) + \ln \frac{2}{a_{\hat{n}}\delta}\right],$$

where $\hat{n}$ is the size of the version space compression set, $\gamma(\mathcal{F},\hat{n})$ is the order-$\hat{n}$ characterizing set complexity of $\mathcal{F}$.

Proof Given our sample $S_m = \{(x_i, f^*(x_i))\}_{i=1}^m$ (labeled by the unknown target function f^*), we define the "synthetic" sample $S'_m = \{(x_i, 1)\}_{i=1}^m$. S'_m can be assumed to have been sampled i.i.d from the marginal distribution of X with positive labels (P').

Theorem 20 can now be applied on the synthetic problem with the training sample S'_m, the distribution P', and the hypothesis space taken to be $\mathcal{H}_i$, the order-i characterizing set of $\mathcal{F}$. It follows that for all $f \in VS_{\mathcal{H}_i,S'_m}$, with probability of at least $1 - a_i\delta$ over choices of S'_m from $(P')^m$,

$$\Pr_{P'}(f(x) \neq 1) \quad \leq \quad \frac{2}{m}\left[h_i \ln\left(\frac{2em}{h_i}\right) + \ln \frac{2}{a_i\delta}\right], \tag{6}$$

where h_i is the VC-dimension of $\mathcal{H}_i$. Then, applying the union bound yields, with probability of at least $1-\delta$, that inequality (6) holds simultaneously for all $1 \leq i \leq m$.

All hypotheses in the version space $VS_{\mathcal{F},S_m}$ agree on all samples in S_m. Hence, the characterizing hypothesis $f_{VS_{\mathcal{F},S_m}}(x) = 1$ for any point $x \in S_m$. Let $\hat{n}$ be the size of the version space compression set. According to Lemma 19, $f_{VS_{\mathcal{F},S_m}}(x) \in \mathcal{H}_{\hat{n}}$. Noting that $f_{VS_{\mathcal{F},S_m}}(x) = 1$ for any $x \in S'_m$, we learn that $f_{VS_{\mathcal{F},S_m}}(x) \in VS_{\mathcal{H}_{\hat{n}},S'_m}$. Therefore, with probability of at least $1-\delta$ over choices of S_m,

$$\Pr_P(f_{VS_{\mathcal{F},S_m}}(x) \neq 1) \leq \frac{2}{m}\left[h_{\hat{n}} \ln\left(\frac{2em}{h_{\hat{n}}}\right) + \ln \frac{2}{a_{\hat{n}}\delta}\right].$$

Since $\Phi(f,g) = \Pr_P(f_{VS_{\mathcal{F},S_m}}(x) = 1)$, and $h_{\hat{n}}$ is the order-$\hat{n}$ characterizing set complexity of $\mathcal{F}$, the proof is complete. ■

6. Consistent Selective Classification With Linear Classifiers

The data dependent bound in Theorem 21 is stated in terms of a new complexity measure (the "characterizing set complexity" of Definition 18). Can this measure be explicitly evaluated or bounded for some interesting hypothesis classes? In this section we consider the class of linear classifiers in $\mathbb{R}^d$. Relying on a classical result from combinatorial geometry, we infer an explicit upper bound on the characterizing set complexity for linear classifiers. Combining this bound with Theorem 21, we immediately obtain a data-dependent compression coverage guarantee, as stated in Corollary 28. We then show that if the unknown underlying distribution is a finite mixture of Gaussians, then CSS will ensure perfect learning with guaranteed coverage. This powerful result, which is stated in Corollary 33, indicates that consistent selective classification might be relevant in various applications of interest.

Fix any positive integer d, and let $\mathcal{F} \triangleq \{f_{\bar{w},\phi}(\bar{x})\}$ be the class of all linear binary classifiers in $\mathbb{R}^d$, where $\bar{w}$ are d-dimensional real vectors, ϕ are scalars, and

$$f_{\bar{w},\phi}(\bar{x}) = \begin{cases} +1, & \bar{w}^T\bar{x} - \phi \geq 0; \\ -1, & \bar{w}^T\bar{x} - \phi < 0. \end{cases}$$

Given a binary labeled training sample S_m, define $R^+ \triangleq R^+(S_m) \subseteq \mathbb{R}^d$ to be the subset of the maximal agreement set with respect to the version space $VS_{\mathcal{F},S_m}$, consisting of all points with positive labels. R^+ is called the 'maximal positive agreement set.' The 'maximal negative agreement set', $R^- \triangleq R^-(S_m)$, is defined similarly. Before continuing, we define a new symmetric hypothesis class $\tilde{\mathcal{F}}$ that allows for a simpler analysis. Let $\tilde{\mathcal{F}} \triangleq \{f_{\bar{w},\phi}(\bar{x})\}$ be the function class

$$\tilde{f}_{\bar{w},\phi}(\bar{x}) = \begin{cases} +1, & \text{if } \bar{w}^T\bar{x} - \phi > 0; \\ 0, & \text{if } \bar{w}^T\bar{x} - \phi = 0; \\ -1, & \text{if } \bar{w}^T\bar{x} - \phi < 0, \end{cases}$$

where we interpret 0 as a classification that agrees with both $+1$ and -1. Given a sample S_m, we define $\tilde{R}^+ \subseteq \mathbb{R}^d$ to be the region in $\mathbb{R}^d$ for which any hypothesis in the version space[2] $VS_{\tilde{\mathcal{F}},S_m}$ classifies either $+1$ or 0 (i.e., this is the maximal positive agreement set). We define $\tilde{R}^-$ analogously with respect to negative or zero classifications. While $\mathcal{F}$ and $\tilde{\mathcal{F}}$ are not identical, the maximal agreement sets they induce are identical. This is stated in the following technical lemma whose proof appears in the appendix.

Lemma 22 (maximal agreement set equivalence) *For any linearly separable sample S_m, $R^+ = \tilde{R}^+$ and $R^- = \tilde{R}^-$.*

The next technical lemma, whose proof also appears in the appendix, provides useful information on the geometry of the maximal agreement set for the class of linear classifiers.

[2] Any hypothesis in $\tilde{\mathcal{F}}$ that classifies every sample in S_m correctly or as 0 belongs to the version space.

Lemma 23 (maximal agreement set geometry I) *Let S_m be a linearly separable labeled sample that is a spanning set of $\mathbb{R}^d$. Then the regions R^+ and R^- are each an intersection of a finite number of half-spaces, with at least d samples on the boundary of each half-space.*

Our goal is to bound the characterizing set complexity of $\mathcal{F}$. As we show below, this complexity measure is directly related to the number of facets of the convex hull of n points in $\mathbb{R}^d$. The following classical combinatorial geometry theorem by Klee (see Preparata and Shamos, 1990, page 98) is thus particularly useful. The statement of Klee's theorem provided here is readily obtained from the original by using the Stirling approximation of the binomial coefficient.

Theorem 24 (Klee, 1966) *The number of facets of a d-polytope with n vertices is at most*

$$2\cdot\left(\frac{en}{\lfloor d/2\rfloor}\right)^{\lfloor d/2\rfloor}. \tag{7}$$

An immediate conclusion is that (7) upper bounds the number of facets of the convex hull of n points in $\mathbb{R}^d$ (which is of course a d-polytope).

Lemma 25 (maximal agreement set geometry II) *Let S_n be a linearly separable sample consisting of $n \geq d+1$ labeled points. Then the regions $R^+(S_n)$ and $R^-(S_n)$ are each an intersection of at most*

$$2(d+1)\cdot\left(\frac{2en}{d}\right)^{\left\lfloor\frac{d+1}{2}\right\rfloor}$$

half-spaces in $\mathbb{R}^d$.

Proof For the sake of clarity, we limit the analysis to a sample S_n in general position; that is, we assume that no more than d points lie on a $(d-1)$-dimensional plane. Handling a sample S_n in arbitrary position can be straightforwardly treated by including an appropriate infinitesimal displacement of the points (the technical proof is omitted).

By Lemma 22, we can limit our discussion to the hypothesis space $\tilde{\mathcal{F}}$ (rather than $\mathcal{F}$). Since S_n includes more than d samples in general position it is a spanning set of $\mathbb{R}^d$. According to Lemma 23, R^+ is an intersection of a finite number of half-spaces, with at least d samples on the boundary of each half-space (and *exactly* d in the general position). Let $S^+ \subseteq S_n$ be the subset of all positive samples in S_n, and $S^- \subseteq S_n$, the negative ones. Let $\tilde{f}_{\bar{w},\phi}$ be one of the half-spaces defining R^+. Then,

$$\forall \bar{x} \in S_n \quad \begin{cases} \bar{w}^T\bar{x}-\phi \geq 0, & \text{if } \bar{x}\in S^+; \\ \bar{w}^T\bar{x}-\phi \leq 0, & \text{if } \bar{x}\in S^-. \end{cases}$$

Also, exactly d samples, $\bar{x}$, satisfy $\overline{wx}-\phi = 0$.

We now embed the samples in $\mathbb{R}^{d+1}$ using the following transformation, $\bar{x} \to \bar{x}'$:

$$\bar{x}' \triangleq \begin{cases} (0,\bar{x}), & \text{if } \bar{x}\in S^+; \\ (1,-\bar{x}), & \text{if } \bar{x}\in S^-. \end{cases}$$

For each half-space $(\bar{w},\phi)$ in $\mathbb{R}^d$ we define a unique half-space, $(\bar{w}',\phi')$, in $\mathbb{R}^{d+1}$,

$$\bar{w}' \triangleq (2\phi,\bar{w}), \quad \phi' \triangleq \phi.$$

We observe that

$$\overline{w'}^T\overline{x'} - \phi' = \begin{cases} \bar{w}^T\bar{x} - \phi \geq 0, & \text{if } \bar{x} \in S^+; \\ 2\phi - \bar{w}^T\bar{x} - \phi = -(\bar{w}^T\bar{x} - \phi) \geq 0, & \text{if } \bar{x} \in S^-, \end{cases}$$

and for exactly d samples we have

$$\overline{w'}^T\overline{x'} - \phi' = \begin{cases} \bar{w}^T\bar{x} - \phi = 0, & \text{if } \bar{x} \in S^+; \\ 2\phi - \bar{w}^T\bar{x} - \phi = -(\bar{w}^T\bar{x} - \phi) = 0, & \text{if } \bar{x} \in S^-. \end{cases}$$

Let $\bar{v}$ be any orthogonal vector to the d samples on the boundary of the half-space. Defining

$$\overline{w''} \triangleq \overline{w'} + \alpha\bar{v}, \quad \phi'' \triangleq \phi',$$

with an appropriate choice of α we have,

$$\forall \overline{x'} \in S_n \quad \overline{w''}^T\overline{x'} - \phi'' = \overline{w'}^T\overline{x'} - \phi' + \alpha\overline{v'}^T\overline{x'} \geq 0,$$

and for exactly $d+1$ samples (including the original d samples),

$$\overline{w''}\overline{x'} - \phi'' = 0.$$

We observe that $\tilde{f}_{\overline{w''},\phi''}$ is a facet of the convex hull of the samples in $\mathbb{R}^{d+1}$. Up to $d+1$ different half-spaces in $\mathbb{R}^d$ can be transformed into a single half-space in $\mathbb{R}^{d+1}$ (the number of combinations of choosing d samples out of $d+1$ samples on the boundary). Using Theorem 24, we bound the number $F(d)$ of facets of the convex hull of the points in $\mathbb{R}^{d+1}$ as follows:

$$F(d) \leq 2 \cdot \left(\frac{en}{\lfloor\frac{d+1}{2}\rfloor}\right)^{\lfloor\frac{d+1}{2}\rfloor} \leq 2 \cdot \left(\frac{2en}{d}\right)^{\lfloor\frac{d+1}{2}\rfloor}.$$

Since up to $d+1$ half-spaces in $\mathbb{R}^d$ can be mapped onto a single facet of the convex hull in $\mathbb{R}^{d+1}$, we can bound the number of half-spaces in $\mathbb{R}^d$ by

$$(d+1) \cdot F(d) \leq 2(d+1) \cdot \left(\frac{2en}{d}\right)^{\lfloor\frac{d+1}{2}\rfloor}.$$

■

Lemma 26 (*Blumer et al., 1989, Lemma 3.2.3*) *Let $\mathcal{F}$ be a binary hypothesis class of finite VC dimension $h \geq 1$. For all $k \geq 1$, define the k-fold intersection,*

$$\mathcal{F}_{k\cap} \triangleq \left\{\cap_{i=1}^k f_i : f_i \in \mathcal{F}, 1 \leq i \leq k\right\},$$

and the k-fold union,

$$\mathcal{F}_{k\cup} \triangleq \left\{\cup_{i=1}^k f_i : f_i \in \mathcal{F}, 1 \leq i \leq k\right\}.$$

Then, for all $k \geq 1$,

$$VC(\mathcal{F}_{k\cap}), VC(\mathcal{F}_{k\cap}) \leq 2hk\log(3k).$$

Lemma 27 (characterizing set complexity) *Fix $d \geq 2$ and $n > d$. Let $\mathcal{F}$ be the class of all linear binary classifiers in $\mathbb{R}^d$. Then, the order-n characterizing set complexity of $\mathcal{F}$ satisfies*

$$\gamma(\mathcal{F}, n) \leq 83 \cdot (d+1)^3 \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \log n.$$

Proof Let $\mathcal{G} = \mathcal{F}_{k\cap}$ be the class of k-fold intersections of half-spaces in $\mathbb{R}^d$. Since the VC dimension of the class of all half-spaces in $\mathbb{R}^d$ is $d+1$, we obtain, using Lemma 26, that the VC dimension of $\mathcal{G}$ satisfies

$$VC(\mathcal{G}) \leq 2k \log(3k)(d+1).$$

Let $\mathcal{H}_n$ be the order-n characterizing set of $\mathcal{F}$. From Lemma 25 we know that any hypothesis $f \in \mathcal{H}_n$ is a union of two regions, where each region is an intersection of no more than

$$k = 2(d+1) \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor}$$

half-spaces in $\mathbb{R}^d$. Therefore, $\mathcal{H}_n \subset \mathcal{G}_{2\cup}$. Using Lemma 26, we get

$$\begin{aligned} VC(\mathcal{H}_n) &\leq VC(\mathcal{G}_{2\cup}) \leq 4\log(6) \cdot VC(\mathcal{G}) \leq 8k \log(6) \log(3k)(d+1) \\ &\leq 16(d+1)^2 \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \log(6) \cdot \log\left(6(d+1) \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor}\right). \end{aligned}$$

For $n > d \geq 2$ we get

$$\begin{aligned} \log\left(6(d+1) \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor}\right) &\leq \log(6n) + \left\lfloor \frac{d+1}{2} \right\rfloor \cdot \log \frac{2en}{d} \\ \leq \quad 3 \cdot \log n + \left\lfloor \frac{d+1}{2} \right\rfloor \cdot \log n^2 &\leq (d+4) \cdot \log n \leq 2 \cdot (d+1) \cdot \log n. \end{aligned}$$

Therefore,

$$VC(\mathcal{H}_n) \quad \leq \quad 83 \cdot (d+1)^3 \cdot \left(\frac{2en}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \log n$$

■

Corollary 28 (data-dependent coverage guarantee) *Let $\mathcal{F}$ be the class of linear binary classifiers in $\mathbb{R}^d$ and assume that the conditions of Theorem 21 hold. Then, $R(f,g) = 0$, and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(f,g) \geq 1 - \frac{2}{m}\left[83(d+1)^3 \Lambda_{\hat{n},d} \ln_+\left(\frac{2em}{\Lambda_{\hat{n},d}}\right) + \ln\frac{2}{a_{\hat{n}}\delta}\right],$$

where $\hat{n}$ is the size of the empirical version space compression set, and

$$\Lambda_{\hat{n},d} = \left(\frac{2e\hat{n}}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \log \hat{n}.$$

Proof Define

$$\Psi(\gamma(\mathcal{F},n)) \quad \triangleq \quad 1-\frac{2}{m}\left[\gamma(\mathcal{F},n)\ln_+\left(\frac{2em}{\gamma(\mathcal{F},n)}\right)+\ln\frac{2}{a_n\delta}\right].$$

We note that $\Psi(\gamma(\mathcal{F},n))$ is a continues function. For any $\gamma(\mathcal{F},n) < 2m$

$$\frac{\partial\Psi(\gamma(\mathcal{F},n))}{\partial\gamma(\mathcal{F},n)} = -\frac{2}{m}\ln\frac{2em}{\gamma(\mathcal{F},n)}+\frac{2}{m} < 0,$$

and for any $\gamma(\mathcal{F},n) > 2m$

$$\frac{\partial\Psi(\gamma(\mathcal{F},n))}{\partial\gamma(\mathcal{F},n)} = -\frac{2}{m} < 0.$$

Thus, $\Psi(\gamma(\mathcal{F},n))$ is monotonically decreasing. Noting that $\ln_+(x)$ is monotonically increasing, by applying Theorem 21 together with Lemma 27 the proof is complete. ■

As long as the empirical version space compression set size $\hat{n}$ is sufficiently small compared to m, Corollary 28 provides a meaningful coverage guarantee. Since $\hat{n}$ might depend on m, it is hard to analyze the effective rate of the bound. To further explore this guarantee, we now bound $\hat{n}$ in terms of m for a specific family of source distributions and derive a distribution-dependent coverage guarantee.

Theorem 29 (Bentley, Kung, Schkolnick, and Thompson, 1978) *If m points in d dimensions have their components chosen independently from any set of continuous distributions (possibly different for each component), then the expected number of convex hull vertices v is*

$$\mathbf{E}[v] = O\left((\log m)^{d-1}\right).$$

Definition 30 (sliced multivariate Gaussian distribution) *A* sliced multivariate Gaussian *distribution, $\mathcal{N}(\Sigma,\mu,w,\phi)$, is a multivariate Gaussian distribution restricted by a half space in $\mathbb{R}^d$. Thus, if Σ is a non-singular covariance matrix, the pdf of the sliced Gaussian is*

$$\frac{1}{C}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\cdot\mathbb{I}(w^Tx-\phi\geq 0),$$

where $\mu = (\mu_1,\ldots,\mu_d)^T$, $\mathbb{I}$ is the indicator function and C is an appropriate normalization factor.

Lemma 31 *Let P be a sliced multivariate Gaussian distribution. If m points are chosen independently from P, then the expected number of convex hull vertices is $O\left((\log m)^{d-1}\right)$.*

Proof Let $X\sim\mathcal{N}(\Sigma,\mu,w,\phi)$ and $Y\sim\mathcal{N}(\Sigma,\mu)$. There is a random vector Z, whose components are independent standard normal random variables, a vector μ, and a matrix A such that $Y = AZ+\mu$. Since

$$w^Ty-\phi = w^T(Az+\mu)-\phi = w^TAz+w^T\mu-\phi,$$

we get that $X = AZ_0+\mu$, where $Z_0\sim\mathcal{N}(I,0,w^TA,\phi-w^T\mu)$. Due to the spherical symmetry of Z, we can choose the half-space $(w^TA,\phi-w^T\mu)$ to be axis-aligned by rotating the axes. We note that the d components of Z are chosen independently and that the axis-aligned half-space enforces restriction only on one of the axes. Therefore, the components of Z_0 are chosen independently as well. Applying Theorem 29, we get that if m points are chosen independently from Z_0, then the expected number of convex hull vertices is $O\left((\log m)^{d-1}\right)$. The proof is complete by noting that the number of convex hull vertices is preserved under affine transformations. ■

Lemma 32 (version space compression set size) *Let $\mathcal{F}$ be the class of all linear binary classifiers in $\mathbb{R}^d$. Assume that the underlying distribution P is a mixture of a fixed number of Gaussians. Then, for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$, the empirical version space compression set size is*

$$\hat{n} = O\left(\frac{(\log m)^{d-1}}{\delta}\right).$$

Proof Let S_n be a version space compression set. Consider $\bar{x_0} \in S_n$. Since S_n is a compression set there is a half-space, $(\bar{w}, \phi)$, such that $f_{\bar{w},\phi} \in VS_{\mathcal{F},S_n \setminus \{\bar{x_0}\}}$ and $f_{\bar{w},\phi} \notin VS_{\mathcal{F},S_n}$. W.l.o.g. assume that $\bar{x_0} \in S_n$ is positive; thus $\bar{w}^T \bar{x_0} - \phi < 0$, and for any other positive point $\bar{x} \in S_n$, $\bar{w}^T \bar{x} - \phi \geq 0$. For an appropriate $\phi' < \phi$, there exists a half-space $(\bar{w}, \phi')$ such that $\bar{w}^T \bar{x_0} - \phi' = 0$, and for any other positive point $\bar{x} \in S_n$, $\bar{w}^T \bar{x} - \phi' > 0$. Therefore, $\bar{x_0}$ is a convex hull vertex. It follows that we can bound the number of positive samples in S_n by the number of vertices of the convex hull of all the positive points. Defining v as the number of convex hull vertices and using Markov's inequality, we get that for any $\varepsilon > 0$,

$$\Pr(v \geq \varepsilon) \leq \frac{\mathbf{E}[v]}{\varepsilon}.$$

Since f^* is a linear classifier, the underlying distribution of the positive points is a mixture of sliced multivariate Gaussians. Using Lemmas 31 and 44, we get that with probability of at least $1-\delta$,

$$v \leq \frac{\mathbf{E}[v]}{\delta} = O\left(\frac{(\log m)^{d-1}}{\delta}\right).$$

Repeating the same arguments for the negative points completes the proof. ■

Corollary 33 (distribution-dependent coverage guarantee) *Let $\mathcal{F}$ be the class of all linear binary classifiers in $\mathbb{R}^d$, and let P be a mixture of a fixed number of Gaussians. Then, $R(f,g) = 0$, and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(f,g) \geq 1 - O\left(\frac{(\log m)^{d^2}}{m} \cdot \frac{1}{\delta^{(d+3)/2}}\right).$$

Proof

$$\Lambda_{\hat{n},d} = \left(\frac{2e\hat{n}}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \log \hat{n} \leq \left(\frac{2e}{d}\right)^{\left\lfloor \frac{d+1}{2} \right\rfloor} \cdot \hat{n}^{\frac{d+3}{2}}.$$

Applying Lemma 32,

$$\Lambda_{\hat{n},d} = O\left(\frac{(\log m)^{d^2}}{\delta^{(d+3)/2}}\right).$$

The proof is complete by noting that $\Lambda_{\hat{n},d} \geq 1$ and using Corollary 28 with $a_i = 2^{-i}$. ■

7. Risk-coverage Trade-off Envelopes

In previous sections we have shown that by compromising the coverage we can achieve zero risk. This is in contrast to the classical setting, where we compromise risk to achieve full coverage. Is it possible to learn a selective classifier with full control over this trade-off? What are the performance limitations of this trade-off control?

In this section we present some answers to these questions thus deriving lower and upper envelopes for the risk-coverage (RC) trade-off. These results heavily rely on the previous results on perfect learning and on classical results on standard learning without rejection. The envelopes are obtained by interpolating bounds on these two extreme types of learning. We begin this section by deriving a lower envelop; that is, we introduce a strategy that can control the RC trade-off.

7.1 Lower Envelop: Controlling the Coverage-risk Trade-off

Our lower RC envelop is facilitated by the following strategy, which is a generalization of the consistent selective classification strategy (CSS) of Definition 6.

Definition 34 (controllable selective strategy) *Given a* mixing *parameter* $0 \leq \alpha \leq 1$*, the* controllable selective strategy *chooses a selective classifier* (f,g) *such that* f *is in the version space* $VS_{\mathcal{F},S_m}$ *(as in CSS), and* g *is defined as follows:* $g(x) = 1$ *for any* x *in the maximal agreement set,* A*, with respect to* $VS_{\mathcal{F},S_m}$*, and* $g(x) = \alpha$ *for any* $x \in \mathcal{X} \setminus A$*.*

Clearly, CSS is a special case of the controllable selective strategy obtained with $\alpha = 0$. Standard consistent learning (in the classical setting) is the special case obtained with $\alpha = 1$. We now state a well known (and elementary) upper bound for classical realizable learning.

Theorem 35 (Haussler, 1988) *Let* $\mathcal{F}$ *be any finite hypothesis class. Let* $f \in VS_{\mathcal{F},S_m}$ *be a classifier chosen by any consistent learner. Then, for any* $0 \leq \delta \leq 1$*, with probability of at least* $1-\delta$*,*

$$R(f) \leq \frac{1}{m}\left(\ln|\mathcal{F}| + \ln\frac{1}{\delta}\right),$$

where $R(f)$ *is standard risk (true error) of the classifier* f*.*

The following result provides a distribution independent upper bound on the risk of the controllable selective strategy as a function of its coverage.

Theorem 36 (lower envelop) *Let* $\mathcal{F}$ *be any finite hypothesis class. Let* (f,g) *be a selective classifier chosen by a controllable selective learner after observing a training sample* S_m*. Then, for any* $0 \leq \delta \leq 1$*, with probability of at least* $1-\delta$*,*

$$R(f,g) \leq \left(\frac{1-\Phi_0/\Phi(f,g)}{1-\Phi_0}\right) \cdot \frac{1}{m}\left(\ln|\mathcal{F}| + \ln\frac{2}{\delta}\right),$$

where

$$\Phi_0 \triangleq 1 - \frac{1}{m}\left((\ln 2)|\mathcal{F}| + \ln\frac{2}{\delta}\right).$$

Proof For any controllable selective learner with a mixing parameter α we have,

$$\Phi(f,g) \quad = \quad \mathbf{E}\left[g(X)\right] = \mathbf{E}\left[\mathbb{I}(g(X)=1)\right] + \alpha\mathbf{E}\left[\mathbb{I}(g(X)\neq 1)\right].$$

By Theorem 8, with probability of at least $1-\frac{\delta}{2}$,

$$\mathbf{E}\left[\mathbb{I}(g(X)=1)\right] \geq 1-\frac{1}{m}\left((\ln 2)|\mathcal{F}|+\ln\frac{2}{\delta}\right) \triangleq \Phi_0.$$

Therefore, since $\Phi(f,g)\leq 1$,

$$\alpha = \frac{\Phi(f,g)-\mathbf{E}\left[\mathbb{I}(g(X)=1)\right]}{1-\mathbf{E}\left[\mathbb{I}(g(X)=1)\right]} \leq \frac{\Phi(f,g)-\Phi_0}{1-\Phi_0}.$$

Using the law of total expectation we get

$$\begin{aligned}
\mathbf{E}\left[\ell(f(X),Y)\cdot g(X)\right] &= \overbrace{\mathbf{E}\left[\ell(f(X),Y)\cdot g(X) \mid g(x)=1\right]}^{0}\cdot\Pr\left(g(X)=1\right)\\
&+ \mathbf{E}\left[\ell(f(X),Y)\cdot g(X) \mid g(x)=\alpha\right]\cdot\Pr\left(g(X)=\alpha\right)\\
&= \alpha\cdot\mathbf{E}\left[\ell(f(X),Y) \mid g(x)=\alpha\right]\cdot\Pr\left(g(X)=\alpha\right)\\
&= \alpha\cdot\mathbf{E}\left[\ell(f(X),Y)\right].
\end{aligned}$$

According to Definition 2.2

$$R(f,g) \quad = \quad \frac{\mathbf{E}\left[\ell(f(X),Y)\cdot g(X)\right]}{\Phi(f,g)} = \frac{\alpha\cdot\mathbf{E}\left[\ell(f(X),Y)\right]}{\Phi(f,g)} = \frac{\alpha\cdot R(f,g)}{\Phi(f,g)}.$$

Applying Theorem 35 together with the union bound completes the proof. ■

7.2 Upper Envelop: Trade-off Control Limitation

We now present a negative result which identifies a region of non-achievable coverage-risk trade-off on the RC plane. The statement is a probabilistic lower bound on the risk of *any* selective classifier expressed as a function of the coverage. It negates any high probability upper bound on the risk of the classifier (where the probability is over choice of S_m and the target hypothesis).

Theorem 37 (non-achievable coverage-risk trade-off) *Let $\mathcal{F}$ be any hypothesis class and let $0\leq\delta\leq\frac{1}{4}$ and m be given. There exists a distribution P (that depends on $\mathcal{F}$), such that for any selective classifier (f,g), chosen using a training sample S_m drawn i.i.d. according to P, with probability of at least δ,*

$$R(f,g)\geq\min\left(\frac{1}{2}-\frac{1}{4\Phi(f,g)},\frac{1}{2}-\frac{1}{2\Phi(f,g)}+\frac{1}{16m\cdot\Phi(f,g)}\cdot\left[VCdim(\mathcal{F})-\frac{16}{3}\ln\frac{1}{1-2\delta}\right]\right)$$

Proof If η is the VC-dimension of hypothesis class $\mathcal{F}$, there exists a set of data points $\mathcal{X}'=\{e_1,e_2,\ldots e_\eta\}$ shattered by $\mathcal{F}$. Let $\mathcal{X}\triangleq\mathcal{X}'\cup\{e_{\eta+1}\}$. The bad distribution is constructed as follows. Define $\overline{Bin}(m,k,\delta)$, the binomial tail inversion,

$$\overline{Bin}\left(m,\frac{\eta}{2},2\delta\right)\triangleq\max_p\left\{p: Bin\left(m,\frac{\eta}{2},p\right)\geq 2\delta\right\},$$

where $Bin(m,k,p)$ is the binomial tail. Define P to be the source distribution over $\mathcal{X}$ satisfying

$$P(e_i) \triangleq \begin{cases} \overline{Bin}\left(m, \frac{\eta}{2}, 2\delta\right)/\eta, & \text{if } i \leq \eta; \\ 1 - \overline{Bin}\left(m, \frac{\eta}{2}, 2\delta\right), & \text{otherwise,} \end{cases}$$

Assuming that the training sample is selected i.i.d. from P, it follows that with probability of at least 2δ,

$$\left|\left\{x \in S_m : x \in \mathcal{X}'\right\}\right| \leq \frac{\eta}{2}.$$

$\mathcal{F}$ shatters $\mathcal{X}'$ thus inducing all dichotomies over $\mathcal{X}'$. Every sample from $\mathcal{X}'$ can reduce the version space by half, so with probability of at least 2δ, the version space $VS_{\mathcal{F},S_m}$ includes all dichotomies over at least $\frac{\eta}{2}$ instances. Therefore, over these instances (referred to as $x_1, x_2, \ldots, x_{\eta/2}$), with probability of $1/2$ the error is at least $\frac{1}{2}$.[3]

$$\begin{aligned} \Phi(f,g) &= \sum_{i=1}^{\eta+1} \{P(e_i) \cdot g(e_i)\} = P(e_1) \cdot \sum_{i=1}^{\eta} g(e_i) + P(e_{\eta+1}) \cdot g(e_{\eta+1}) \\ &\leq P(e_1) \cdot \sum_{i=1}^{\frac{\eta}{2}} g(x_i) + \frac{\eta}{2} \cdot P(e_1) + P(e_{\eta+1}) = P(e_1) \cdot \sum_{i=1}^{\frac{\eta}{2}} g(x_i) + 1 - \frac{\eta}{2} \cdot P(e_1) \\ &\Longrightarrow P(e_1) \cdot \sum_{i=1}^{\frac{\eta}{2}} g(x_i) \geq \Phi(f,g) + \frac{\eta}{2} \cdot P(e_1) - 1. \end{aligned}$$

$$\begin{aligned} \Phi(f,g) \cdot R(f,g) &= \sum_{i=1}^{\eta+1} \{P(e_i) \cdot g(e_i) \cdot \mathbb{I}(f(e_i) \neq f^*(e_i))\} \geq \sum_{i=1}^{\frac{\eta}{2}} \left\{P(x_i) \cdot g(x_i) \cdot \frac{1}{2}\right\} \\ &\geq \frac{\Phi(f,g) - 1}{2} + \frac{\eta}{4} \cdot P(e_1) = \frac{\Phi(f,g) - 1}{2} + \frac{1}{4} \cdot \overline{Bin}\left(m, \frac{\eta}{2}, 2\delta\right). \end{aligned}$$

Applying Lemma 43 we get

$$R(f,g) \geq \min\left(\frac{1}{2} - \frac{1}{4\Phi(f,g)}, \frac{1}{2} - \frac{1}{2\Phi(f,g)} + \frac{1}{16m \cdot \Phi(f,g)} \cdot \left[VCdim(\mathcal{F}) - \frac{16}{3} \ln \frac{1}{1-2\delta}\right]\right)$$

■

Corollary 38 *Let $0 \leq \delta \leq \frac{1}{4}$, m, and $n > 1$ be given. There exist a distribution P, that depends on m and n, and a finite hypothesis class $\mathcal{F}$ of size n, such that for any selective classifier (f,g), chosen using a training sample S_m drawn i.i.d. according to P, with probability of at least δ, if*

$$\Phi(f,g) \geq max\left\{\frac{3}{4}, 1 - \frac{1}{16m} \cdot \left[VCdim(\mathcal{F}) - \frac{16}{3} \ln \frac{1}{1-2\delta}\right]\right\}$$

then

$$R(f,g) \geq \frac{1}{16m} \cdot \left[VCdim(\mathcal{F}) - \frac{16}{3} \ln \frac{1}{1-2\delta}\right].$$

[3] According to the game theoretic setting the adversary can choose a distribution over $\mathcal{F}$. In this case the expectation in the risk is averaged over random instances and random labels. Therefore, the error over the instances $x_1, x_2, \ldots, x_{\eta/2}$ is exactly $1/2$ and we can replace the term 2δ with δ.

Proof Assuming

$$\Phi(f,g) \geq max\left\{\frac{3}{4}, 1 - \frac{1}{16m} \cdot \left[VCdim(\mathcal{F}) - \frac{16}{3}\ln\frac{1}{1-2\delta}\right]\right\},$$

we apply Theorem 37 to complete the proof. ■

8. CSS Implementation: Lazy CSS

In previous sections we analyzed the performance of CSS and proved that (in the realizable case) it can achieve sharp coverage rates under reasonable assumptions on the source distribution while guaranteeing zero error on the accepted samples. However, it remains unclear whether an efficient implementation of CSS is at reach. In this section we propose an algorithm for CSS and show that it can be efficiently implemented for linear classifiers.

The following method, which we term *lazy CSS*, is very similar to the implicit selective sampling algorithm of Cohn et al. (1994). Instead of explicitly constructing the CSS selection function g during training (which indeed can be a very complex task), we adapt a "lazy learning" approach that can potentially facilitate an efficient CSS implementation during test time. In particular, we propose to evaluate $g(x)$ at any given test point x during the classification process. For the training set S_m and a test point x we define the following two sets:

$$S_{m,x}^{+} \triangleq S_m \cup \{(x,+1)\}, \quad S_{m,x}^{-} \triangleq S_m \cup \{(x,-1)\};$$

that is, $S_{m,x}^{+}$ is the (labeled) training set S_m augmented by the test point x labeled positively, and $S_{m,x}^{-}$ is S_m augmented by x labeled negatively. The selection value $g(x)$ is determined as follows: $g(x) = 0$ (i.e., x is rejected) iff there exist hypotheses $f^{+}, f^{-} \in \mathcal{F}$ that are consistent with $S_{m,x}^{+}$ and $S_{m,x}^{-}$, respectively.

The following lemma states that the selection function $g(x)$ constructed by lazy CSS is a precise implementation of CSS.

Lemma 39 *Let $\mathcal{F}$ be any hypothesis class, S_m a labeled training set, and x, a test point. Then x belongs to the maximal agreement set of $VS_{\mathcal{F},S_m}$ iff there is no hypothesis $f \in \mathcal{F}$ that is consistent with either $S_{m,x}^{+}$ or $S_{m,x}^{-}$.*

Proof If there exist hypotheses $f^{+}, f^{-} \in \mathcal{F}$ that are consistent with $S_{m,x}^{+}$ and $S_{m,x}^{-}$, then there exist two hypotheses in $\mathcal{F}$ that correctly classify S_m (therefore they belong to $VS_{\mathcal{F},S_m}$) but disagree on x. Hence, x does not belong to the maximal agreement set of $VS_{\mathcal{F},S_m}$. Conversely, if x does not belong to the maximal agreement set of $VS_{\mathcal{F},S_m}$, then there are two hypotheses, f_1 and f_2, which correctly classify S_m but disagree on x. Let's assume, without loss of generality, that f_1 classifies x positively. Then, f_1 is consistent with $S_{m,x}^{+}$ and f_2 is consistent with $S_{m,x}^{-}$. Thus there exist hypotheses $f^{+}, f^{-} \in \mathcal{F}$ that are consistent with $S_{m,x}^{+}$ and $S_{m,x}^{-}$. ■

For the case of linear classifiers it follows that computing the lazy CSS selection function for any test point is reduced to two applications of a linear separability test. Yogananda et al. (2007) recently presented a fast linear separability test with a worst case time complexity of $O(mr^3)$ and space complexity of $O(md)$, where m is the number of points, d is the dimension and $r \leq \min(m, d+1)$.

Remark 40 *For the realizable case we can modify* any *rejection mechanism by restricting rejection* only *to the region chosen for rejection by CSS. Since CSS accepts only samples that are guaranteed to have zero test error, the overall performance of the modified rejection mechanism is guaranteed to be at least as good as the original mechanism. Using this technique we were able to improve the performance (RC curve) of the most commonly used rejection mechanism for linear classifiers, which rejects samples according to a simple symmetric distance from the decision boundary (a "margin").*

9. Which Rejection Model?

In classical classification and Bayes decision theory the goal is to minimize a cost function (or a loss function), where the cost is specified by a $K \times K$ cost matrix ($K = 2$ for the binary case). Given the cost matrix, the objective is to select a classifier that minimizes the average weighted cost (over unobserved instances) as specified by this matrix. When introducing rejection it is necessary to introduce a suitable optimization criterion (which is referred here also as a 'rejection model'). Obviously, the desired criterion should take into account both the risk of the classifier and its coverage. The question we discuss in this section is: what would be an appropriate optimization criterion for selective classification?

A very common rejection model in the literature is the *cost model*, whereby a specific cost d is associated with rejection (see, e.g., Tortorella, 2001) and the objective is to minimize the generalized *rejective risk function*,

$$\ell_c(f,g) \quad \triangleq \quad d \cdot \mathbf{E}\left[1 - g(X)\right] + \mathbf{E}\left[\mathbb{I}(f(X) \neq Y) \cdot g(X)\right]. \tag{8}$$

Given our definitions of risk and coverage, the function (8) can be easily expressed as a function over the RC plane of Figure 1,

$$\ell_c(R,\Phi) = d\,(1 - \Phi(f,g)) + R(f,g)\Phi(f,g). \tag{9}$$

For any fixed d, Equation (9) defines level sets (or elevation contour lines) over the RC plane. For example, Figure 3(a) depicts elevation contour lines induced by (9) with a rejection cost $d = 0.3$. The thick line in this figure represent our knowledge of the optimal RC trade-off. Thus, an optimal classifier, according to this cost model, has a risk-coverage profile that minimizes the cost (9) with respect to all choices on the RC trade-off curve. This optimal choice is depicted in Figure 3(a) by the black dot. This popular cost model was refined to accommodate differentiation between the cost of false positive and false negative as well as different costs for rejection of positive and negative samples (Herbei and Wegkamp, 2006; Pietraszek, 2005; Tortorella, 2001; Santos-Pereira and Pires, 2005). Such extensions or refinements are appealing because they allow for additional control and more flexibility in modeling the problem at hand. Nevertheless, these cost models are often criticized for lack of usability in applications where it is impossible or hard to precisely quantify the cost of rejection. It is interesting to note that for an ideal Bayesian setting, where the underlying distribution is completely known, Chow showed (Chow, 1970) that the cost d upper bounds the probability of misclassification. In this case one can control the classification error by specifying a matching rejection cost.

In Pietraszek (2005) two additional optimization models are introduced. The first, *bounded-improvement* model, is depicted as contour elevation lines over the RC plane in Figure 3(c). In this model, given a constraint on the misclassification cost, the classifier should reject as few samples as possible. The constraint is specified by the ∞ symbol in the RC plane, which is the cost defined for the entire rectangle containing all risk-coverage profiles having risk larger than the constraint (0.3 in this example). In the second, *bounded-abstention* model (depicted in Figure 3(b)), given a constraint on the coverage (0.5 in this example), the classifier should have the lowest misclassification cost. It is argued in Pietraszek (2005) that these models are more suitable than the above cost model in many applications, for instance, when a classifier with limited classification throughput (e.g., a human expert) should handle the rejected instances, and in a medical and quality assurance applications, where the goal is to reduce the misclassification cost to a user-defined value.

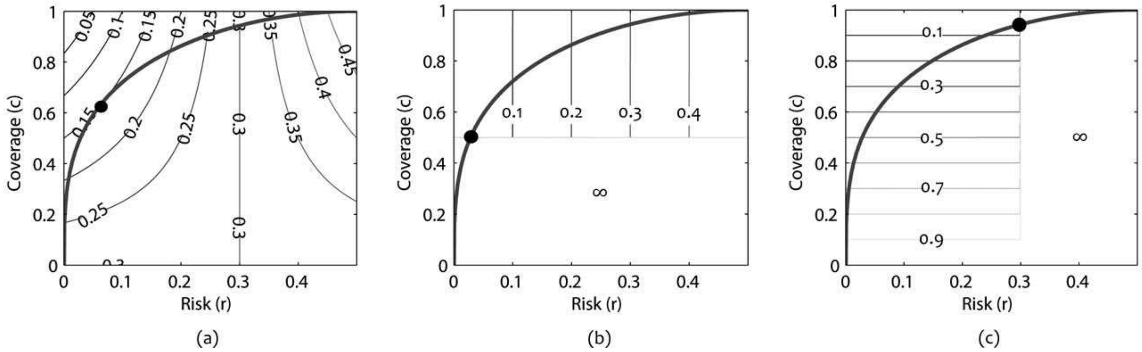

Figure 3: Rejection models: (a) cost (b) bounded-abstention (c) bounded-improvement

Which cost model among the above three is the right model? This question is, obviously, ill-defined and the answer depends on the application. Thus, when deriving bounds for a specific generalized rejective risk function the results are limited to only one specific model. Instead, one can handle *any* rejective risk function over the RC plane by identifying the RC trade-off. Specifically, by bounding the coverage and the risk separately (as we do in this paper) we can in principle optimize *any* generalized rejective risk function according to any desired rejection model including the cost, the bounded-improvement and bounded-abstention models.

10. Related Work

The idea of classification with a reject option dates back to Chow's seminal papers (Chow, 1957, 1970). These papers analyzed both the Bayes-optimal reject decision and the reject-rate vs. error trade-off. This is done under the 0-1 loss function, assuming that the underlying distribution is completely known. The Bayes-optimal rejection policy is based, as in standard classification, on maximum a posteriori probabilities. Instances should be rejected whenever none of the posteriori probabilities are sufficiently distinct. This type of rejection can be termed *ambiguity-based* rejection. Referring to the diagram in Figure 1, one of Chow's main results (for the case of complete probabilistic knowledge), is that the optimal RC trade-off (depicted by the dotted line) is monotonically increasing.

While the optimal decision can be identified in the case of complete probabilistic knowledge, it was argued (Fumera et al., 2000) that when the a posteriori probabilities are estimated with errors, Chow's rule (Chow, 1970) does not provide the optimal error-reject trade-off. Tortorella (2001) and Santos-Pereira and Pires (2005) discussed Bayesian-optimal decisions in the case of arbitrary cost matrices. In these papers the optimal reject rule was chosen based on the ROC curve evaluated on a subset of the training data. As in most papers on the subject emerging from the engineering community (see, e.g., Fumera et al., 2000; Fumera and Roli, 2002; Pietraszek, 2005; Bounsiar et al., 2006; Landgrebe et al., 2006) no probabilistic or other guarantees are provided for the misclassification error.

Very few studies have focused on error bounds for classifiers with a reject option. Hellman (1970) proposed and analyzed two rejection rules for the nearest neighbor algorithm. Extending Cover and Hart's classic result for the 1-nearest neighbor algorithm (Cover and Hart, 1967), Hellman showed that the test error (over non-rejected points) of a nearest neighbor algorithm with a reject option can be bounded *asymptotically* (as the sample size approaches infinity) by some factor of the Bayes error (with reject). To the best of our knowledge, this excess risk bound is the first that has been introduced in the context of classification with a reject option.

Herbei and Wegkamp (2006) developed excess risk bounds for the classification with a reject option setting where the loss function is the 0-1 loss, extended such that the cost of each reject point is $0 \leq d \leq 1/2$ (cost model; see Section 9). This result generalizes the excess risk bounds of Tsybakov (2004) for standard binary classification without reject (which is equivalent to the case $d = 1/2$). The bound applies to any empirical error minimization technique. This result is further extended in Bartlett and Wegkamp (2007) and Wegkamp (2007) in various ways, including the use of the hinge loss function for efficient optimization. The main results of Herbei and Wegkamp (both for plug-in rules and empirical risk minimization) degenerate, in the realizable case, to a meaningless bound, where classification with a reject option is not guaranteed to be any better than classification without reject. These results are also limited only to the cost model (see discussion on Section 9). Saying that, we must also note that comparing bounds that were derived for the agnostic setting with our results can be misleading or "unfair" since the agnostic setting is much more difficult. The only purpose of this comparison is to clarify that the results here are not special cases of any of the currently known agnostic bounds.

Freund et al. (2004) studied a simple ensemble method for binary classification. Given an hypothesis class $\mathcal{F}$, the method outputs a weighted average of all the hypotheses in $\mathcal{F}$ such that 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 inconclusive (i.e., sufficiently close to zero). Two regret bounds for this algorithm were derived. The first bounds the probability of error when the classifier decides not to reject. If ε is the error of the best hypothesis in $\mathcal{F}$, the error of the aggregating algorithm is bounded above (w.h.p.) by $2\varepsilon + O(\frac{1}{m^{1/2-\theta}})$, where $0 < \theta < 1/2$ is an hyperparameter. The authors also proved that for a sufficiently large training sample size, $m = \Omega((\sqrt{ln(1/\delta)}ln(|\mathcal{F}|))^{1/\theta})$, the probability that the algorithms will abstain from prediction is bounded above by $5\varepsilon + O(\frac{\ln|\mathcal{F}|}{\sqrt{m^{1/2-\theta}}})$. To the best of our knowledge, these bounds are the first to provide some guarantee for both the error of the classifier and the coverage. Therefore, these results are related to the bounded-improvement and bounded-abstention models (see Section 9). As was rightfully stated by the authors, the final aggregated hypothesis can significantly outperform the best base-hypothesis in $\mathcal{F}$ in some favorable situations. Unfortunately, the regret bound provided does not exploit these situations, as it is bounded by twice the generalization error of the best hypothesis. Referring to the diagram in Figure 1, The results of Freund et al. can be depicted as a curve in region B (thus characterizing some achievable zone). For the realizable case, the bounds of Freund et al. achieve much slower rates than those we derive in this paper.

Selective classification is related to selective sampling (Atlas et al., 1990). In selective sampling the learner sequentially processes unlabeled examples, and for each one determines whether or not to request a label. One of the earliest active learning algorithms for the realizable case (termed "mellow active learner") was proposed by Cohn et al. (1994). Their well motivated approach is to request labels only for samples that belong to the region of disagreement (the complement of our maximal agreement set). As mentioned in Section 8, this is very similar to our CSS. Hanneke studied the rate for which the region of disagreement collapses as the algorithm processes examples (Hanneke, 2007, 2009). He introduced the notion of *disagreement coefficient* and derived upper bounds on the label complexity in active learning expressed in terms of this coefficient. While his results capture the convergence rate of the region of disagreement as a function of the number of *label requests*, in selective classification we are interested in convergence rates as a function of the size of the *training set* (in selective sampling the number of labels does not necessarily match the number of samples). Specific disagreement coefficient values were recently derived for some interesting hypothesis classes including homogeneous linear classifiers in $\mathbb{R}^d$ under uniform data distribution (Hanneke, 2007) and linear classifiers in $\mathbb{R}^d$ under smooth data density bounded away from zero (Friedman, 2009). While coverage bounds and label complexity bounds cannot be directly compared, we conjecture that formal connections between these two settings exist because the disagreement region plays a key role in both. The precise relation between these two settings is yet to be discovered.

11. Concluding Remarks

Selective classification is well recognized as a very attractive technique for improving classification accuracy. In fact, it is among very few methods that can help in practical applications where sufficiently low error cannot be achieved in the standard model. Nevertheless, not enough is known about selective classification in order to harness its power in a controlled, optimal way, or to avoid its use in cases where it cannot sufficiently help.

In this work we made a first step toward a rigorous analysis of selective classification by revealing properties of the risk-coverage trade-off, which represents optimal selective classification. By focusing on the extreme case of perfect learning we were able to derive initial results for entire risk-coverage trade-offs.

Many interesting questions are left open. Among the most important open questions are the following. What would be an analogous concept to perfect learning in the fully agnostic (non-realizable) setting? What is the precise relation between selective classification and selective sampling? Is it possible to implement efficiently the CSS strategy and prove useful bounds for other natural hypothesis classes? Can selective classification be rigorously analyzed in transductive, semi-supervised or active settings? With respect to agnostic extensions, while it doesn't make much sense to talk about "perfect learning" in a noisy setting, it is meaningful and interesting to consider the analogous concept to regret (or excess risk) bounds. Here we could employ a selective strategy aiming at achieving the error rate of the best hypothesis in the class precisely (and perhaps with certainty).

Acknowledgments

We thank Charles Elkan for useful discussions and the anonymous referees for excellent comments.

Appendix A.

Lemma 41 *For $u > \sqrt{2}v > 0$,*

$$\frac{u-v}{u+v} \geq 1 - 4 \cdot \frac{v}{u}.$$

Proof

$$\frac{u-v}{u+v} \quad = \quad 1 - \frac{2v}{u+v} = 1 - 2v\frac{u-v}{u^2-v^2} \geq 1 - 2v\frac{u}{u^2-v^2}.$$

Since $u > \sqrt{2}v$, we have

$$u^2 - v^2 > \frac{u^2}{2}.$$

Applying to the previous inequality completes the proof. ■

Lemma 42 (Bernstein's inequality Hoeffding, 1963) *Let $X_1, \ldots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then for all positive t,*

$$\Pr\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left\{-\frac{t^2/2}{\sum \mathbf{E}\left[X_j^2\right] + Mt/3}\right\}.$$

Lemma 43 (binomial tail inversion lower bound) *For $k > 0$ and $\delta \leq \frac{1}{2}$,*

$$\overline{Bin}(m,k,\delta) \geq \min\left(1, \frac{k}{2m} - \frac{4}{3m}\ln\frac{1}{1-\delta}\right).$$

Proof Let $Z_1, \ldots Z_m$ be independent Bernoulli random variables each with a success probability $0 \leq p \leq 1$. Setting $W_i \triangleq Z_i - p$,

$$\begin{aligned} Bin(m,k,p) \quad &= \quad \Pr_{Z_1,\ldots,Z_m \sim B(p)^m}\left(\sum_{i=1}^{m} Z_i \leq k\right) = 1 - \Pr\left(\sum_{i=1}^{m} Z_i > k\right) \\ &= \quad 1 - \Pr\left(\sum_{i=1}^{m} W_i > k - mp\right). \end{aligned}$$

Clearly, $\mathbf{E}[W_i] = 0$, $|W_i| \leq 1$, and $\mathbf{E}\left[W_i^2\right] = p \cdot (1-p)^2 + (1-p) \cdot p^2 = p \cdot (1-p)$. Using Lemma 42 (Bernstein's inequality) we thus obtain,

$$Bin(m,k,p) \geq 1 - \exp\left(-\frac{(k-mp)^2/2}{mp(1-p) + (k-mp)/3}\right).$$

Since $(1-p) \leq 1$,

$$\begin{aligned}\frac{(k-mp)^2/2}{mp(1-p)+(k-mp)/3} &\geq \frac{(k-mp)^2}{2mp+\frac{2}{3}\cdot(k-mp)} = \frac{(k-mp)^2}{\frac{4}{3}mp+\frac{2}{3}\cdot k} \\ &\geq \frac{k^2-2mpk}{\frac{4}{3}mp+\frac{2}{3}\cdot k}.\end{aligned}$$

Therefore,

$$Bin(m,k,p) \geq 1 - e^{-\frac{k^2-2mpk}{\frac{4}{3}mp+\frac{2}{3}\cdot k}}.$$

Equating the right-hand side to δ and solving for p, we have

$$p \leq \frac{k}{2m} \cdot \frac{k-\frac{2}{3}t}{k+\frac{2}{3}\cdot t},$$

where $t \triangleq \ln\frac{1}{(1-\delta)}$. Choosing

$$p = \min\left\{1, \frac{1}{2m}\left(k - \frac{8}{3}\ln\frac{1}{1-\delta}\right)\right\} = \min\left\{1, \frac{k}{2m}\cdot\left(1-4\cdot\frac{\frac{2}{3}t}{k}\right)\right\},$$

using the fact that $k \geq 1 > \frac{2\sqrt{2}}{3}\ln\frac{1}{1/2} \geq \frac{2\sqrt{2}}{3}t$, and applying Lemma 41, we get

$$p \leq \frac{k}{2m} \cdot \frac{k-\frac{2}{3}t}{k+\frac{2}{3}\cdot t}.$$

Therefore, $Bin(m,k,p) \geq \delta$. Since $\overline{Bin}(m,k,\delta) = \max_p\{p : Bin(m,k,p) \geq \delta\}$, we conclude that

$$\overline{Bin}(m,k,\delta) \geq p = \min\left(1, \frac{k}{2m} - \frac{4}{3m}\ln\frac{1}{1-\delta}\right).$$

■

Lemma 44 *Let S_1 and S_2 be two sets in $\mathbb{R}^d$. Then,*

$$H(S_1 \cup S_2) \leq H(S_1) + H(S_2),$$

where $H(S)$ is the number of convex hull vertices of S.

Proof Assume $x \in S_1 \cup S_2$ is a convex hull vertex of $S_1 \cup S_2$. Then, there is a half-space (w,ϕ) such that, $w \cdot x - \phi = 0$, and any other $y \in S_1 \cup S_2$ satisfies $w \cdot y - \phi > 0$. Assume w.l.o.g. that $x \in S_1$. Then it is clear that any $y \in S_1$ satisfies $w \cdot y - \phi > 0$. Therefore, x is a convex hull vertex of S_1. ■

Proof of Lemma 22 Let $S^+ \subseteq S_m$ be the set of all positive samples in S_m, and $S^- \subseteq S_m$ be the set of all negative samples. Let $\bar{x}_0 \in R^+$. There exists a hypothesis $f_{\bar{w},\phi}(\bar{x})$ such that

$$\begin{aligned}\forall \quad \bar{x} \in S^+, \quad \bar{w}^T\bar{x} - \phi \geq 0;\\ \forall \quad \bar{x} \in S^-, \quad \bar{w}^T\bar{x} - \phi < 0,\end{aligned}$$

and

$$\bar{w}^T\bar{x}_0 - \phi \geq 0.$$

Let's assume that $\bar{x}_0 \notin \tilde{R}^+$. Then, there exists a hypothesis $\tilde{f}_{\bar{w}',\phi'}(\bar{x})$ such that

$$\begin{array}{lll} \forall & \bar{x} \in S^+, & \bar{w}'^T\bar{x} - \phi' \geq 0; \\ \forall & \bar{x} \in S^-, & \bar{w}'^T\bar{x} - \phi' \leq 0, \end{array}$$

and

$$\bar{w}'^T\bar{x}_0 - \phi' < 0.$$

Defining

$$\bar{w}_0 \triangleq \bar{w} + \alpha\bar{w}', \qquad \phi_0 \triangleq \phi + \alpha\phi',$$

where

$$\alpha > \left| \frac{\bar{w}^T\bar{x}_0 - \phi}{\bar{w}'^T\bar{x}_0 - \phi'} \right|,$$

we deduce that there exists a hypothesis $f_{\bar{w}_0,\phi_0}(\bar{x})$ such that

$$\begin{array}{lll} \forall & \bar{x} \in S^+, & \bar{w}_0^T\bar{x} - \phi_0 \geq 0; \\ \forall & \bar{x} \in S^-, & \bar{w}_0^T\bar{x} - \phi_0 < 0, \end{array}$$

and

$$\begin{array}{rcl} \bar{w}_0^T\bar{x}_0 - \phi_0 & = & \bar{w}^T\bar{x}_0 - \phi + \alpha\left[\bar{w}'^T\bar{x}_0 - \phi'\right] = \bar{w}^T\bar{x}_0 - \phi - \alpha\left|\bar{w}'^T\bar{x}_0 - \phi'\right| \\ & < & \bar{w}^T\bar{x}_0 - \phi - \left|\bar{w}^T\bar{x}_0 - \phi\right| = 0. \end{array}$$

Therefore, $\bar{x}_0 \notin R^+$. Contradiction. Hence, $\bar{x}_0 \in \tilde{R}^+$ and $R^+ \subseteq \tilde{R}^+$. The proof that $R^- \subseteq \tilde{R}^-$ follows the same argument.

To prove that $\tilde{R}^+ \subseteq R^+$, we look at $VS_{\tilde{\mathcal{F}},S_m}$:

$$\forall \tilde{f}_{\bar{w},\phi} \in VS_{\tilde{\mathcal{F}},S_m}, \bar{x} \in \tilde{R}^+ \quad \bar{w}^T\bar{x} - \phi \geq 0.$$

We observe that if $f_{\bar{w},\phi} \in VS_{\mathcal{F},S_m}$, then $\tilde{f}_{\bar{w},\phi} \in VS_{\tilde{\mathcal{F}},S_m}$. Therefore,

$$\forall f_{\bar{w},\phi} \in VS_{\mathcal{F},S_m}, \bar{x} \in \tilde{R}^+ \quad \bar{w}^T\bar{x} - \phi \geq 0.$$

Hence, $\tilde{R}^+ \subseteq R^+$.

It remains to prove that $\tilde{R}^- \subseteq R^-$. Assuming $\bar{x}_0 \notin R^-$ implies that there exists a hypothesis $f_{\bar{w},\phi}(\bar{x})$ such that

$$\begin{array}{lll} \forall & \bar{x} \in S^+, & \bar{w}^T\bar{x} - \phi \geq 0; \\ \forall & \bar{x} \in S^-, & \bar{w}^T\bar{x} - \phi < 0, \end{array}$$

and

$$\bar{w}^T\bar{x}_0 - \phi \geq 0.$$

Defining[4]

$$\bar{w}_0 \triangleq \bar{w}, \quad \phi_0 \triangleq \phi - \left| \max_{\bar{x} \in S^-} \left(\bar{w}^T \bar{x} - \phi \right) \right|,$$

we conclude that there exists a hypothesis $\tilde{f}_{\bar{w}_0,\phi_0}(\bar{x})$ such that

$$\begin{aligned} &\forall \quad \bar{x} \in S^+ \quad \bar{w}_0^T \bar{x} - \phi_0 \geq 0; \\ &\forall \quad \bar{x} \in S^- \quad \bar{w}_0^T \bar{x} - \phi_0 \leq \max_{\bar{x} \in S^-} \left(\bar{w}^T \bar{x} - \phi \right) + \left| \max_{\bar{x} \in S^-} \left(\bar{w}^T \bar{x} - \phi \right) \right| = 0, \end{aligned}$$

and

$$\bar{w}_0^T \bar{x}_0 - \phi_0 > 0.$$

Therefore, $\bar{x}_0 \notin \tilde{R}^-$, so $\tilde{R}^- \subseteq R^-$. ■

Proof of Lemma 23 According to Lemma 22, $R^+ = \tilde{R}^+$ and $R^- = \tilde{R}^-$. Therefore, we can restrict our discussion to the hypothesis class $\tilde{\mathcal{F}}$. Due to the symmetry of the hypothesis class $\tilde{\mathcal{F}}$ we will concentrate only on the positive region R^+. Set $G \triangleq VS_{\tilde{\mathcal{F}},S_m}$. By definition,

$$\tilde{R}^+ = \bigcap_{f'_{\bar{w},\phi} \in G} f'_{\bar{w},\phi},$$

where $f'_{\bar{w},\phi}$ denotes the region in $\mathcal{X}$ for which the linear classifier $f'_{\bar{w},\phi}$ obtains the value one or zero. Let $f_{\bar{w},\phi} \in G$ be a half-space with $k < d$ points on its boundary. We will prove that there exist two half-spaces in G ($f_{\bar{w}_1,\phi_1}$, $f_{\bar{w}_2,\phi_2}$) such that each has at least $k+1$ samples on its boundary and

$$f_{\bar{w},\phi} \bigcap f_{\bar{w}_1,\phi_1} \bigcap f_{\bar{w}_2,\phi_2} = f_{\bar{w}_1,\phi_1} \bigcap f_{\bar{w}_2,\phi_2}.$$

Therefore,

$$\tilde{R}^+ = \bigcap_{f'_{\bar{w},\phi} \in G \setminus \{f_{\bar{w},\phi}\}} f'_{\bar{w},\phi}.$$

Repeating this process recursively with every half-space in G, with less than d points on its boundary, completes the proof.

Before proceeding with the rigorous analysis let's review the main idea behind the proof. If a half-space in $\mathbb{R}^d$ has less than d points on its boundary, it has at least one degree of freedom. Rotating the half-space clockwise or counterclockwise around a specific axis (defined by the points on the boundary) by sufficiently small angles will maintain correct classification over S_m. We will rotate the half-space clockwise and counterclockwise until "touching" the first point in S_m on each direction. This operation will maintain correct classification but will result in having one additional point on the boundary. Then we only have to show that the intersection of the three half-spaces (original and two rotated ones) is the same as the intersection of the two rotated ones.

[4] If S^- is an empty set we can arbitrarily define $\phi_0 \triangleq \phi - 1$.

Let $f_{\bar{w},\phi} \in G$ be a half-space with $k < d$ points on its boundary. Without loss of generality assume that these points are $S_m^0 \triangleq \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k\}$. For the sake of simplicity we will first *translate* the space such that $\bar{x}_1$ will lie on the origin. Since $\bar{x}_1$ is on the boundary of the half-space we get

$$\bar{w}^T\bar{x}_1 - \phi = 0 \quad \Longrightarrow \quad \phi = \bar{w}^T\bar{x}_1.$$

Therefore,

$$\forall \bar{x} \in S_m^0 \quad 0 = \bar{w}^T\bar{x} - \phi = \bar{w}^T\bar{x} - \bar{w}^T\bar{x}_1 = \bar{w}^T(\bar{x} - \bar{x}_1).$$

Hence, the weight vector $\bar{w}$ is orthogonal to all the translated samples $(\bar{x}_1 - \bar{x}_1), \ldots, (\bar{x}_k - \bar{x}_1)$. We now have $k < d$ vectors in $\mathbb{R}^d$ (including the weight vector) so we can always find at least one vector $\bar{v}$ which is orthogonal to all the rest. We now *rotate* the translated samples around the origin so as to align the vector $\bar{w}$ with the first axis, and align the vector $\bar{v}$ with the second axis. From now on all translated and rotated coordinates and vectors will be marked with prime. Define the following rotation matrix in $\mathbb{R}^d$,

$$R_\theta \triangleq \left(\begin{array}{ccccc} \cos\theta & \sin\theta & 0 & 0 & \ldots \\ -\sin\theta & \cos\theta & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \\ \vdots & \vdots & \vdots & & \ddots \end{array}\right).$$

We can now define two new half-spaces in the translated and rotated space, $f_{R_\alpha\bar{w}',0}$ and $f_{R_{-\beta}\bar{w}',0}$, where

$$\alpha = \max_{0<\alpha'\leq\pi}\{\alpha' \quad | \quad \forall \bar{x}' \in S_m \quad (\bar{w}'^T\bar{x}')\cdot(R_{\alpha'}\bar{w}')^T\bar{x}' \geq 0\}, \tag{10}$$

and

$$\beta = \max_{0<\beta'\leq\pi}\{\beta' \quad | \quad \forall \bar{x}' \in S_m \quad (\bar{w}'^T\bar{x}')\cdot(R_{-\beta'}\bar{w}')^T\bar{x}' \geq 0\}. \tag{11}$$

According to Claim 45, both $f_{R_\alpha\bar{w}',0}$ and $f_{R_{-\beta}\bar{w}',0}$ correctly classify S_m and have at least $k+1$ samples on their boundaries.

Now we examine the intersection of $f_{R_\alpha\bar{w}',0}$ and $f_{R_{-\beta}\bar{w}',0}$. According to Claim 46, if $(R_\alpha\bar{w}')^T\bar{x}' \geq 0$ and $(R_{-\beta}\bar{w}')^T\bar{x}' \geq 0$, then $\bar{w}'^T\bar{x}' \geq 0$. The intersection of $f_{R_\alpha\bar{w}',0}$, $f_{R_{-\beta}\bar{w}',0}$ and $f_{\bar{w}',0}$ thus equals the intersection of $f_{R_\alpha\bar{w}',0}$ and $f_{R_{-\beta}\bar{w}',0}$, as required. ■

Claim 45 *Both $f_{R_\alpha\bar{w}',0}$ and $f_{R_{-\beta}\bar{w}',0}$ correctly classify S_m and have at least $k+1$ samples on their boundaries.*

Proof We note that after translation all half-spaces pass through the origin, so $\phi' = 0$. Recall the definitions of α and β as maximums (Equations (10) and (11), respectively). We show that the maximums over α' and β' are well defined. Let $\bar{x}' = (x_1', x_2', \cdots x_d')^T$. Since $\bar{w}' = (1, 0, 0, \cdots)^T$ we get that $R_\alpha\bar{w}' = (\cos\alpha, -\sin\alpha, 0, \ldots)^T$ and

$$(\bar{w}'^T\bar{x}')\cdot(R_{\alpha'}\bar{w}')^T\bar{x}' = {x_1'}^2\cos\alpha - {x'}_1\cdot{x'}_2\cdot\sin\alpha.$$

Since S_m is a spanning set of $\mathbb{R}^d$, at least one sample has a vector with component $x_1' \neq 0$. As all components are finite and ${x_1'}^2 > 0$, we can always find a sufficiently small α' such that ${x_1'}^2\cos\alpha' -$

$x'_1 \cdot x'_2 \cdot \sin\alpha' > 0$. Hence, the maximum exists. Furthermore, for $\alpha' = \pi$ we have ${x'_1}^2 \cos\pi - x'_1 \cdot x'_2 \cdot \sin\pi = -{x'_1}^2 < 0$. Noticing that ${x'_1}^2 \cos\alpha - x'_1 \cdot x'_2 \cdot \sin\alpha$ is continuous in α, and applying the intermediate value theorem, we know that $0 < \alpha < \pi$ and ${x'_1}^2 \cos\alpha - x'_1 \cdot x'_2 \cdot \sin\alpha = 0$. Therefore, there exists a sample in S_m that is not on the boundary of $f_{\overline{w}',0}$ (since $x'_1 \neq 0$) but on the boundary of $f_{R_\alpha \overline{w}',0}$. Recall that all points in S^0_m are orthogonal to $\overline{w}' = (1,0,0,\cdots)^T$ and $\overline{v} = (0,1,0,\cdots)^T$. Therefore,

$$\forall \overline{x}' \in S^0_m \quad (R_\alpha \overline{w}')^T \overline{x}' = x'_1 \cdot \cos\alpha - x'_2 \cdot \sin\alpha = \overline{w}'^T \overline{x} \cdot \cos\alpha - \overline{v}'^T \overline{x} \cdot \sin\alpha = 0,$$

and they reside on the boundary of $f_{R_\alpha \overline{w}',0}$. Overall, $f_{R_\alpha \overline{w}',0}$ correctly classifies S_m and has at least $k+1$ samples on its boundary. The same argument applies for β by symmetry. ■

Claim 46 *Using the notation introduced in the proof of Lemma 23, if* $(R_\alpha \overline{w}')^T \overline{x}' \geq 0$ *and* $(R_{-\beta} \overline{w}')^T \overline{x}' \geq 0$*, then*

$$\overline{w}'^T \overline{x}' \geq 0.$$

Proof If $(R_\alpha \overline{w}')^T \overline{x}' \geq 0$ and $(R_{-\beta} \overline{w}')^T \overline{x}' \geq 0$, then

$$\begin{cases} x'_1 \cos\alpha - x'_2 \cdot \sin\alpha \geq 0, \\ x'_1 \cos\beta + x'_2 \cdot \sin\beta \geq 0. \end{cases}$$

Multiplying the first inequality by $\sin\beta > 0$, the second inequality by $\sin\alpha > 0$, and adding the two we have

$$\sin(\alpha+\beta) \cdot x'_1 \geq 0.$$

According to Claim 47 below, $\sin(\alpha+\beta) \geq 0$. If $\sin(\alpha+\beta) = 0$, then $(\alpha+\beta) = \pi$ and $\cos(\alpha+\beta) = -1$. Using the trigonometric identities

$$\begin{aligned} \cos(\alpha-\beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta; \\ \sin(\alpha-\beta) &= \sin\alpha\cos\beta - \cos\alpha\sin\beta, \end{aligned}$$

we get that

$$\cos\beta = \cos(\beta+\alpha-\alpha) = \cos(\alpha+\beta) \cdot \cos\alpha + \sin(\alpha+\beta) \cdot \sin\alpha = -\cos\alpha,$$

and

$$\sin\beta = \sin(\beta+\alpha-\alpha) = \sin(\alpha+\beta) \cdot \cos\alpha - \cos(\alpha+\beta) \cdot \sin\alpha = \sin\alpha.$$

Therefore, for any $x'_1 \cos\alpha - x'_2 \cdot \sin\alpha > 0$, it holds that $x'_1 \cos\beta + x'_2 \cdot \sin\beta < 0$ and $\tilde{R}^+$ is degenerated. Contradiction to the fact that S_m is a spanning set of the $\mathbb{R}^d$. Therefore, $\sin(\alpha+\beta) > 0$, $x'_1 \geq 0$ and $\overline{w}'^T \overline{x}' \geq 0$. ■

Claim 47 *Using the notation introduced in the proof of Lemma 23,*

$$\sin(\alpha+\beta) \geq 0.$$

Proof By definition we get that for all samples in S_m,

$$\begin{cases} {x'_1}^2 \cos\alpha - x'_1 \cdot x'_2 \cdot \sin\alpha \geq 0, \\ {x'_1}^2 \cos\beta + x'_1 \cdot x'_2 \cdot \sin\beta \geq 0. \end{cases}$$

Multiplying the first inequality by $\sin\beta > 0$ $(0 < \beta < \pi)$, the second inequality by $\sin\alpha > 0$, adding the two, and using the trigonometric identity

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta,$$

we have

$$\sin(\alpha+\beta)\cdot {x'_1}^2 \geq 0.$$

Since there is a sample in S_m with a vector component $x'_1 \neq 0$, we conclude that $\sin(\alpha+\beta) \geq 0$. ■

References

M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations*. Cambridge University Press, 1999.

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30(1): 31–56, 1998.

L. Atlas, D. Cohn, R. Ladner, A.M. El-Sharkawi, and R.J. Marks. Training connectionist networks with queries and selective sampling. In *Neural Information Processing Systems (NIPS)*, pages 566–573, 1990.

P.L. Bartlett and M.H. Wegkamp. Classification with a reject option using a hinge loss. Technical report M980, Department of Statistics, Florida State University, 2007.

J.L. Bentley, H.T. Kung, M. Schkolnick, and C.D. Thompson. On the average number of maxima in a set of vectors and applications. *Journal of the ACM*, 25, 1978.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36, 1989.

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

C.K. Chow. An optimum character recognition system using decision function. *IEEE Transactions on Computers*, 6(4):247–254, 1957.

C.K. Chow. On optimum recognition error and reject trade-off. *IEEE Transactions 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.

T.M. Cover and P. Hart. Neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

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

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

G. Fumera and F. Roli. Support vector machines with embedded reject option. In *Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, pages 811–919, 2002.

G. Fumera, F. Roli, and G. Giacinto. Multiple reject thresholds for improving classification reliability. In *Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 863–871, 2000.

B. Hanczar and E.R. Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2008.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36:177–221, 1988.

M.E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems, Man and Cybernetics*, 6:179–185, 1970.

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

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

J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

P. S. Meltzer, J. Khan, J. S. Wei, M Ringnér, L. H. Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, C. R. Antonescu, and C Peterson. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6), June 2001.

T. Mitchell. Version spaces: a candidate elimination approach to rule learning. In *IJCAI'77: Proceedings of the 5th international joint conference on Artificial Intelligence*, pages 305–310, 1977.

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

F.P. Preparata and M.I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, 1990.

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

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

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1): 135–166, 2004.

V. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

M.H. Wegkamp. Lasso type classifiers with a reject option. *Electronic Journal of Statistics*, 1: 155–168, 2007.

A.P. Yogananda, M.N. Murthy, and G. Lakshmi. A fast linear separability test by projection of positive points on subspaces. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 713–720, 2007.