Optimal Single-Class Classification Strategies

Ran El-Yaniv
Department of Computer Science
Technion- Israel Institute of Technology
Technion, Israel 32000
rani@cs.technion.ac.il

Mordechai Nisenson
Department of Computer Science
Technion - Israel Institute of Technology
Technion, Israel 32000
motin@cs.technion.ac.il

Abstract

We consider single-class classification (SCC) as a two-person game between the learner and an adversary. In this game the target distribution is completely known to the learner and the learner's goal is to construct a classifier capable of guaranteeing a given tolerance for the false-positive error while minimizing the false negative error. We identify both "hard" and "soft" optimal classification strategies for different types of games and demonstrate that soft classification can provide a significant advantage. Our optimal strategies and bounds provide worst-case *lower bounds* for standard, finite-sample SCC and also motivate new approaches to solving SCC.

1 Introduction

In *Single-Class Classification (SCC)* the learner observes a training set of examples sampled from one *target class*. The goal is to create a classifier that can distinguish the target class from other classes, unknown to the learner during training. This problem is the essence of a great many applications such as intrusion, fault and novelty detection. SCC has been receiving much research attention in the machine learning and pattern recognition communities (for example, the survey papers [7, 8, 4] cite, altogether, over 100 papers). The extensive body of work on SCC, which encompasses mainly empirical studies of heuristic approaches, suffers from a lack of theoretical contributions and few principled (empirical) comparative studies of the proposed solutions. Thus, despite the extent of the existing literature, some of the very basic questions have remained unresolved.

Let $P(x)$ be the underlying distribution of the target class, defined over some space Ω. We call P the *target distribution*. Let $0 < \delta < 1$ be a given tolerance parameter. The learner observes a training set sampled from P and should then construct a classifier capable of distinguishing the target class. We view the SCC problem as a game between the learner and an adversary. The adversary selects another distribution Q over Ω and then a new element of Ω is drawn from $\gamma P + (1-\gamma)Q$, where γ is a switching parameter (unknown to the learner). The goal of the learner is to minimize the false negative error, while guaranteeing that the false positive error will be at most δ.

The main consideration in previous SCC studies has been *statistical*: how can we guarantee a prescribed false positive rate (δ) given a finite sample from P? This question led to many solutions, almost all revolving around the idea of *low-density rejection*. The proposed approaches are typically *generative* or *discriminative*. Generative solutions range from full density estimation [2], to partial density estimation such as quantile estimation [5], level set estimation [1, 9] or local density estimation [3]. In discriminative methods one attempts to generate a decision boundary appropriately enclosing the high density regions of the training set [11].

In this paper we abstract away the statistical estimation component of the problem and model a setting where the learner has a very large sample from the target class. In fact, we assume that the learner knows the target distribution P precisely. While this assumption would render almost the

entire body of SCC literature superfluous, it turns out that a significant, *decision-theoretic* component of the SCC problem remains – one that has so far been overlooked. In any case, the results we obtain here immediately apply to other SCC instances as *lower bounds*.

The fundamental question arising in our setting is: What are optimal strategies for the learner? In particular, is the popular low-density rejection strategy optimal? While most or all SCC papers adopted this strategy, nowhere in the literature could we find a formal justification.

The partially good news is that low-density rejection is worst-case optimal, but only if the learner is confined to "hard" decision strategies. In general, the worst-case optimal learner strategy should be "soft"; that is, the learner should play a randomized strategy, which could result in a very significant gain. We first identify a monotonicity property of optimal SCC strategies and use it to establish the optimality of low-density rejection in the "hard" case. We then show an equivalence between low-density rejection and a constrained two-class classification problem where the other class is the uniform distribution over Ω. This equivalence motivates a new approach to solving SCC problems.

We next turn our attention to the power of the adversary, an issue that has been overlooked in the literature but has crucial impact on the relevancy of SCC solutions in applications. For example, when considering an intrusion detection application (see, e.g., [6]), it is necessary to assume that the "attacking distribution" has some worst-case characteristics and it is important to quantify precisely what the adversary knows or can do. The simple observation in this setting is that an *omniscient and unlimited adversary*, who knows all parameters of the game including the learner's strategy, would completely demolish the learner who uses hard strategies. By using a soft strategy, however, the learner can achieve on average the biased coin false negative rate of $1 - \delta$.

We then analyze the case of an omniscient but limited adversary, who must select a sufficiently distant Q satisfying $D_{KL}(Q||P) \geq \Lambda$, for some known parameter Λ. One of our main contributions is a complete analysis of this game, including identification of the optimal strategy for the learner and the adversary, as well as the best achievable false negative rate. The optimal learner strategy and best achievable rate are obtained via a solution of a linear program specified in terms of the problem parameters. These results are immediately applicable as *lower bounds* for standard (finite-sample) SCC problems, but may also be used to inspire new types of algorithms for standard SCC. While we do not have a closed form expression for the best achievable false-negative rate, we provide a few numerical examples demonstrating and comparing the optimal "hard" and "soft" performance.

2 Problem Formulation

The *single-class classification (SCC)* problem is defined as a game between the *learner* and an *adversary*. The learner receives a training sample of examples from a *target distribution* P defined over some space Ω. On the basis of this training sample, the learner should select a rejection function $r : \Omega \to [0, 1]$, where for each $\omega \in \Omega$, $r_\omega = r(\omega)$ is the probability with which the learner will reject ω. On the basis of any knowledge of P and/or $r(\cdot)$, the adversary selects selects an *attacking distribution* Q, defined over Ω. Then, a new example is drawn from $\gamma P+(1-\gamma)Q$, where $0 < \gamma < 1$, is a *switching probability* unknown to the learner. The *rejection rate* of the learner, using a rejection function r, with respect to any distribution D (over Ω), is $\rho(D) = \rho(r, D) \triangleq \mathbf{E}_D\{r(\omega)\}$. For notational convenience whenever we decorate r (e.g., r',r^*), the corresponding ρ will be decorated accordingly (e.g., ρ',ρ^*). The two main quantities of interest here are the *false positive rate* (type I error) $\rho(P)$, and the *false negative rate* (type II error) $1 - \rho(Q)$.

Before the start of the game, the learner receives a tolerance parameter $0 < \delta < 1$, giving the maximally allowed false positive rate. A rejection function $r(\cdot)$ is *valid* if its false positive rate $\rho(P) \leq \delta$. A valid rejection function (strategy) is optimal if it guarantees the smallest false negative rate amongst all valid strategies.

We consider a model where the learner knows the target distribution P exactly, thus focusing on the decision-theoretic component in SCC. Clearly, our model approximates a setting where the learner has a very large training set, but the results we obtain immediately apply, in any case, as *lower bounds* to other SCC instances.

This SCC game is a two-person zero sum game where the payoff to the learner is $\rho(Q)$. The set $\mathcal{R}_\delta(P) \triangleq \{r : \rho(P) \leq \delta\}$ of valid rejection functions is the learner's strategy space. Let $\mathcal{Q}$ be the

strategy space of the adversary, consisting of all allowable distributions Q that can be selected by the adversary. We are concerned with optimal learner strategies for game variants distinguished by the adversary's knowledge of the learner's strategy, P and/or of δ and by other limitations on $\mathcal{Q}$.

We distinguish a special type of this game, which we call the *hard setting*, where the learner must deterministically reject or accept new events; that is, $r : \Omega \to \{0, 1\}$, and such rejection functions are termed "hard." The more general game defined above (with "soft" functions) is called the *soft setting*. As far as we know, only the hard setting has been considered in the SCC literature thus far.

In the soft setting, given any rejection function, the learner can reduce the type II error by rejecting more (i.e., by increasing $r(\cdot)$). Therefore, for an optimal $r(\cdot)$ we have $\rho(P) = \delta$ (rather than $\rho(P) \leq \delta$). It follows that the switching parameter γ is immaterial to the selection of an optimal strategy. Specifically, the combined error of an optimal strategy is $\gamma\rho(P) + (1-\gamma)(1-\rho(Q)) = \gamma\delta + (1-\gamma)(1-\rho(Q))$, which is minimized by minimizing the type II error, $1-\rho(Q)$.

We assume throughout this paper a finite support of size N; that is, $\Omega = \{1, \ldots, N\}$ and $P \triangleq \{p_1, \ldots, p_N\}$ and $Q \triangleq \{q_1, \ldots, q_N\}$ are probability mass functions. Additionally, a "probability distribution" refers to a distribution over the fixed support set Ω. Note that this assumption still leaves us with an infinite game because the learner's pure strategy space, $\mathcal{R}_\delta(P)$, is infinite.[1]

3 Characterizing Monotone Rejection Functions

In this section we characterize the structure of optimal learner strategies. Intuitively, it seems plausible that the learner should not assign higher rejection values to higher probability events under P. That is, one may expect that a reasonable rejection function $r(\cdot)$ would be monotonically decreasing with probability values (i.e., if $p_j \leq p_k$ then $r_j \geq r_k$). Such monotonicity is a key justification for a very large body of SCC work, which is based on low density rejection strategies. Surprisingly, optimal monotone strategies are not always guaranteed as shown in the following example.

Example 3.1 (Non-Monotone Optimality) *In the hard setting, take* $N = 3$, $P = (0.06, 0.09, 0.85)$ *and* $\delta = 0.1$. *The two* δ*-valid hard rejection functions are* $r' = (1, 0, 0)$ *and* $r'' = (0, 1, 0)$. *Let* $\mathcal{Q} = \{Q = (0.01, 0.02, 0.97)\}$. *Clearly* $\rho'(Q) = 0.01$ *and* $\rho''(Q) = 0.02$ *and therefore,* $r''(\cdot)$ *is optimal despite breaking monotonicity. More generally, this example holds if* $\mathcal{Q} = \{Q : q_2 - q_1 \geq \varepsilon\}$ *for any* $0 < \varepsilon \leq 1$.

In the soft setting, let $N = 2$, $P = (0.2, 0.8)$, *and* $\delta = 0.1$. *We note that* $\mathcal{R}_\delta(P) = \{r^\varepsilon = (0.1 + 4\varepsilon, 0.1 - \varepsilon)\}$, *for* $\varepsilon \in [-0.025, 0.1]$. *We take* $\mathcal{Q} = \{Q = (0.1, 0.9)\}$. *Then* $\rho^\epsilon(Q) = 0.1 + 0.4\varepsilon - 0.9\varepsilon = 0.1 - 0.5\varepsilon$. *This is clearly maximized when we minimize* ε *by taking* $\varepsilon = -0.025$, *and then the optimal rejection function is* $(0, 0.125)$, *which clearly breaks monotonicity. This example also holds for* $\mathcal{Q} = \{Q : q_2 \geq cq_1\}$ *for any* $c > 4$.

Fix P and δ. For any adversary strategy space, $\mathcal{Q}$, let $\mathcal{R}^*_\delta(P)$ be the set of optimal valid rejection functions, $\mathcal{R}^*_\delta \triangleq \{r \in \mathcal{R}_\delta(P) \ : \min_{Q\in\mathcal{Q}} \rho(Q) = \max_{r'\in\mathcal{R}_\delta(P)} \min_{Q\in\mathcal{Q}} \rho'(Q)\}$.[2] We note that $\mathcal{R}^*_\delta$ is never empty in the cases we consider. A simple observation is that for any $r \in \mathcal{R}^*_\delta$ there exists $r' \in \mathcal{R}^*_\delta$ such that $r'(i) = r(i)$ for all i such that $p_i > 0$ and for zero probabilities, $p_j = 0$, $r'(j) = 1$.

The following property ensures that $\mathcal{R}^*_\delta$ will include a monotone (optimal) hard strategy, which means that the search space for the learner can be conveniently confined to monotone strategies. While the set of all distributions satisfies this property, later on we will consider limited strategic adversary spaces where this property still holds.[3]

[1] The game is conveniently described in extensive form (i.e., game tree) where in the first move the learner selects a rejection function, followed by a chance move to determine the source (either P or Q) of the test example (with probability γ). In the case where Q is selected, the adversary chooses (randomly using Q) the test example. In this game the choice of Q depends on knowledge of P and $r(\cdot)$.

[2] For certain strategy spaces, $\mathcal{Q}$, it may be necessary to consider the infimum rather than the minimum. In such cases it may be necessary to replace '$Q \in \mathcal{Q}$' (in definitions, theorems, etc.) with '$Q \in cl(\mathcal{Q})$', where $cl(\mathcal{Q})$ is the closure of $\mathcal{Q}$.

[3] All properties defined in this paper could be made weaker for the purposes of the proofs, but this would needlessly complicate them. Indeed, the way they are currently defined is sufficient for most "reasonable" $\mathcal{Q}$.

Definition 3.2 (Property A) *Let P be a distribution. A set of distributions $\mathcal{Q}$ has Property A w.r.t. P if for all j,k and $Q \in \mathcal{Q}$ such that $p_j < p_k$ and $q_j < q_k$, there exists $Q' \in \mathcal{Q}$ such that $q'_k \leq q_j$, $q'_j \geq q_k$ and for all $i \neq j,k$, we have $q'_i = q_i$.*

Theorem 3.3 (Monotone Hard Decisions) *When the learner is restricted to hard-decisions and $\mathcal{Q}$ satisfies Property A w.r.t. P, then $\exists r \in \mathcal{R}^*_\delta$ such that $p_j < p_k \Rightarrow r(j) \geq r(k)$.*[4]

Proof: Let us assume by contradiction that no such rejection function exists in $\mathcal{R}^*_\delta$. Let $r \in \mathcal{R}^*_\delta$. Let j be such that $p_j = \min_{\omega:r(\omega)=0} p_\omega$. Then, there must exist k, such that $p_j < p_k$ and $r(k) = 1$ (otherwise r is monotone). Define r^* to be r with the values of j and k swapped; that is, $r^*(j) = 1, r^*(k) = 0$ and for all other i, $r^*(i) = r(i)$. We note that $\rho^*(P) = \rho(P) + p_j - p_k < \rho(P) \leq \delta$. Let $Q^* \in \mathcal{Q}$ be such that $\min_Q \rho^*(Q) = \rho^*(Q^*) = \rho(Q^*) + q^*_j - q^*_k$. Thus, if $q^*_j \geq q^*_k$, $\rho^*(Q^*) \geq \rho(Q^*)$. Otherwise, there exists $Q^{*\prime}$ as in Property A and in particular, $q^{*\prime}_k \leq q^*_j$. As a result, $\rho^*(Q^*) = \rho(Q^{*\prime}) + q^*_j - q^{*\prime}_k \geq \rho(Q^{*\prime})$. Therefore, there always exists $Q \in \mathcal{Q}$ such that $\rho^*(Q^*) \geq \rho(Q)$ (either $Q = Q^*$ or $Q = Q^{*\prime}$). Consequently, $\min_Q \rho^*(Q) \geq \min_Q \rho(Q)$, and thus, $r^* \in \mathcal{R}^*_\delta$. As long as there are more j,k pairs which need to have their rejection levels fixed, we label $r = r^*$ and repeat the above procedure. Since the only changes are made to $r^*(j)$ and $r^*(k)$, and since j is the non-rejected event with minimal probability, the procedure will be repeated at most N times. The final r^* is in $\mathcal{R}^*_\delta$ and satisfies $p_j < p_k \Rightarrow r(j) \geq r(k)$. Contradiction. □

Theorem 3.3 provides a formal justification for the *low-density rejection strategy (LDRS)*, popular in the SCC literature. Specifically, assume w.l.o.g. $p_1 \leq p_2 \leq \cdots \leq p_N$. The corresponding δ-valid low density rejection function places $r_j = 1$ iff $\sum_{i=1}^{j} p_i \leq \delta$.

Our discussion on soft decisions is facilitated by Property B and Theorem 3.5 that follow.

Definition 3.4 (Property B) *Let P be a distribution. A set of distributions $\mathcal{Q}$ has Property B w.r.t. P if for all j,k and $Q \in \mathcal{Q}$ such that $0 < p_j \leq p_k$ and $\frac{q_j}{p_j} < \frac{q_k}{p_k}$, there exists $Q' \in \mathcal{Q}$ such that $\frac{q'_j}{p_j} \geq \frac{q'_k}{p_k}$ and for all $i \neq j,k$, $q'_i = q_i$.*

The rather technical proof of the following theorem is omitted for lack of space (and appears in the adjoining, supplementary appendix).

Theorem 3.5 (Monotone Soft Decisions) *If $\mathcal{Q}$ satisfies Property B w.r.t. P, then $\exists r \in \mathcal{R}^*_\delta$ such that: (i)$p_i = 0 \Rightarrow r(i) = 1$; (ii) $p_j < p_k \Rightarrow r(j) \geq r(k)$; and (iii) $p_j = p_k \Rightarrow r(j) = r(k)$.*

4 Low-Density Rejection and Two-Class Classification

In this section we focus on the hard setting. We show that the low-density rejection strategy (LDRS - defined in Section 3) is optimal. Moreover we show that the optimal hard performance can be obtained by solving a constrained two-class classification problem where the other class is the uniform distribution over Ω. The results here consider families $\mathcal{Q}$ that satisfy the following property.

Definition 4.1 (Property C) *Let P be a distribution. A set of distributions $\mathcal{Q}$ has Property C w.r.t. P if for all j,k and $Q \in \mathcal{Q}$ such that $p_j = p_k$ there exists $Q' \in \mathcal{Q}$ such that $q'_k = q_j$, $q'_j = q_k$ and for all $i \neq j,k$, $q'_i = q_i$.*

We state without proof the following lemma (the proof can be found in the appendix).

Lemma 4.2 *Let r^* be a δ-valid low-density rejection function (LDRS). Let r be any monotone δ-valid rejection function. Then $\min_{Q\in\mathcal{Q}} \rho^*(Q) \geq \min_{Q\in\mathcal{Q}} \rho(Q)$ for any $\mathcal{Q}$ satisfying Property C.*

Example 4.3 (Violation of Property C) *We illustrate here that violating Property C may result in a violation of Lemma 4.2. Let $N = 5$, $P = (0.02, 0.03, 0.05, 0.05, 0.85)$, and $\delta = 0.1$. Then the two δ-valid LDRS rejection functions are $r = (1,1,1,0,0)$ and $r' = (1,1,0,1,0)$. Let $\mathcal{Q} = \{Q : q_3 - q_4 > \varepsilon\}$ for some $0 < \varepsilon < 1$. Then, for any $Q \in \mathcal{Q}$, $\rho(Q) - \rho'(Q) = q_3 - q_4 > \varepsilon$, and therefore, for the LDRS, r', there exists a monotone r such that $\min_{Q\in\mathcal{Q}} \rho'(Q) < \min_{Q\in\mathcal{Q}} \rho(Q)$.*

[4] Here we must consider a weaker notion of monotonicity for hard strategies to be both valid and optimal.

When $\mathcal{Q}$ satisfies Property A, then by Theorem 3.3 there exists a monotone *optimal* rejection function. Therefore, the following corollary of Lemma 4.2 establishes the optimality of any LDRS.

Corollary 4.4 *Any δ-valid LDRS is optimal if $\mathcal{Q}$ satisfies both Property A and Property C.*

Thus, any LDRS strategy is indeed worst-case optimal when the learner is willing to be confined to hard rejection functions and when the adversary's space satisfies Property A and Property C. We now show that an (optimal) LDRS solution is equivalent to an optimal solution of the following *constrained* Bayesian two-class decision problem. Let the first class c_1 have distribution $P(x)$ and the second class, c_2, have the uniform distribution $U(x) = 1/N$. Let $0 < c < 1$ and $0 < \epsilon < (N\delta c + 1 - c)/N\delta c$. The classes have priors $\Pr\{c_1\} = c$ and $\Pr\{c_2\} = 1 - c$. The loss function λ_{ij}, giving the cost of deciding c_i instead of c_j $(i, j = 1, 2)$, is $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = (Nc + 1 - c)/(1 - c)$ and $\lambda_{21} = \epsilon$. The goal is to construct a classifier $C(x) \in \{c_1, c_2)$ that minimizes the total Bayesian risk under the constraint that, for a given δ, $\sum_{x:C(x)=c_2} P(x) \leq \delta$. We term this problem "the Bayesian binary problem."

Theorem 4.5 *An optimal binary classifier for the Bayesian binary problem induces an optimal (hard) solution to the SCC problem (an LDRS) when $\mathcal{Q}$ satisfies properties A and C.*

Proof Sketch: Let $C^*(\cdot)$ be an optimal classifier for the Bayesian binary problem. Any classifier $C(\cdot)$ induces a hard rejection function $r(\cdot)$ by taking $r(x) = 1 \Leftrightarrow C(x) = c_2$. Therefore, the set of feasible classifiers (satisfying the constraint) clearly induces $\mathcal{R}_\delta(P)$. Let $M_i(C) \triangleq \{x : C(x) = i\}$. Note that the constraint is equivalent to $\sum_{x \in M_2(C)} P(x) \leq \delta$. The Bayes risk for classifying x as i is $R_i(x) \triangleq \lambda_{ii} \Pr\{c_i|x\} + \lambda_{i(3-i)} \Pr\{c_{3-i}|x\} = \lambda_{i(3-i)} \Pr\{c_{3-i}|x\}$. The total Bayes risk is $R(C) \triangleq \sum_{x \in M_1(C)} R_1(x) + \sum_{x \in M_2(C)} R_2(x)$, which is minimized at $C^*(\cdot)$. It is not difficult to show that $R_1(\cdot)$ and $R_2(\cdot)$ are monotonically decreasing and increasing, respectively. It therefore follows that $x \in M_1(C^*)$, $y \in M_2(C^*) \Rightarrow P(x) \geq P(y)$ (otherwise, by swapping $C^*(x)$ and $C^*(y)$, the constraint can be maintained and $R(C^*)$ decreased). It is also not difficult to show that $R_1(x) \geq 1 > R_2(x)$ for any x. Thus, it follows that $\sum_{y \in M_2(C^*)} P(y) + \min_{x \in M_1(C^*)} P(x) > \delta$ (otherwise, some x could be transferred from $M_1(C^*)$ to $M_2(C^*)$, reducing $R(C^*)$). Together, these two properties immediately imply that $C^*(\cdot)$ induces a δ-valid LDRS. □

Theorem 4.5 motivates a different approach to SCC in which we sample from the uniform distribution over Ω and then attempt to approximate the optimal Bayes solution to the constrained binary problem. It also justifies certain heuristics found in the literature [10, 11].

5 The Omniscient Adversary: Games, Strategies and Bounds

5.1 Unrestricted Adversary

In the first game we analyze an adversary who is completely unrestricted. This means that $\mathcal{Q}$ is the set of all distributions. Unsurprisingly, this game leaves little opportunity for the learner. For any rejection function $r(\cdot)$, define $r_{min} \triangleq \min_i r(i)$ and $I_{min}(r) \triangleq \{i : r(i) = r_{min}\}$. For any distribution D, $\rho(D) = \sum_{i=1}^{N} d_i r(i) \geq \sum_{i=1}^{N} d_i r_{min} = r_{min}$, in particular, $\delta = \rho(P) \geq r_{min}$ and $\min_Q \rho(Q) \geq r_{min}$. By choosing Q such that $q_i = 1$ for some $i \in I_{min}(r)$, the adversary can achieve $\rho(Q) = r_{min}$ (the same rejection rate is achieved by taking any Q with $q_i = 0$ for all $i \notin I_{min}(r)$). In the soft setting, $\min_Q \rho(Q)$ is maximized by the rejection function $r^\delta(i) \triangleq \delta$ for all $p_i > 0$ ($r^\delta(i) \triangleq 1$ for all $p_i = 0$) This is equivalent to flipping a δ-biased coin for non-null events (under P). The best achievable Type II Error is $1 - \delta$. In the hard setting, clearly $r_{min} = 0$ (otherwise $1 > \delta \geq 1$), and the best achievable Type II Error is precisely 1. That is, absolutely nothing can be achieved.

This simple analysis shows the futility of the SCC game when the adversary is too powerful. In order to consider SCC problems at all one must consider reasonable restrictions on the adversary that lead to more useful games. One type of restriction would be to limit the adversary's knowledge of $r(\cdot)$, P and/or of δ. Another type would be to directly limit the strategic choices available to the adversary. In the next section we focus on the latter type.

5.2 A Constrained Adversary

In seeking a quantifiable constraint on $\mathcal{Q}$ it is helpful to recall that the essence of the SCC problem is to try to distinguish between two probability distributions (albeit one of them unknown). A natural constraint is a lower bound on the "distance" between these distributions. Following similar results in hypothesis testing, we would like to consider games in which the adversary must select Q such that $D(P||Q) \geq \Lambda$, for some constant $\Lambda > 0$, where $D(\cdot||\cdot)$ is the KL-divergence. Unfortunately, this constraint is vacuous since $D(P||Q)$ explodes when $q_i \ll p_i$ (for any i). In this case the adversary can optimally play the same strategy as in the unrestricted game while meeting the KL-divergence constraint. Fortunately, by taking $D(Q||P) \geq \Lambda$, we can effectively constrain the adversary.

We note, as usual, that the learner can (and should) reject with probability 1 any null events under P. Thus, an adversary would be foolish to choose a distribution Q that has any probability for these events. Therefore, we henceforth assume w.l.o.g. that $\Omega = \Omega(P) \triangleq \{\omega : p_\omega > 0\}$. Taking $D(Q||P) \triangleq \sum_{i=1}^{N} q_i \log(q_i/p_i)$, we then define $\mathcal{Q} = \mathcal{Q}_\Lambda \triangleq \{Q : D(Q||P) \geq \Lambda\}$. We note that $\mathcal{Q}_\Lambda$ possesses properties A, B and C w.r.t. P,[5] and by Theorems 3.3 and 3.5 there exists a monotone $r \in \mathcal{R}^*_\delta$ (in both the hard and soft settings) and by Corollary 4.4, any δ-valid LDRS is hard-optimal.

If $\max_i p_i \leq 2^{-\Lambda}$, then any Q which is concentrated on a single event meets the constraint $D(Q||P) \geq \Lambda$. Then, the adversary can play the same strategy as in the unrestricted game, and the learner should select r^δ as before. For the game to be non-trivial it is thus required that $\Lambda > \log(1/\max_i p_i)$. Similarly, if the optimal r is such that there exists $j \in I_{min}(r)$ (that is $r(j) = r_{min}$) and $p_j \leq 2^{-\Lambda}$, then a distribution Q that is completely concentrated on j has $D(Q||P) \geq \Lambda$ and achieves $\rho(Q) = r_{min}$ as in the unrestricted game. Therefore, $r = r^\delta$, and so maximizes r_{min}. We thus assume that the optimal r has no such j.

We begin our analysis of the game by identifying some useful characteristics of optimal adversary strategies in Lemma 5.1. Then Theorem 5.2 shows that the effective support of an optimal Q has a size of two at most. Based on these properties, we provide in Theorem 5.3 a linear program that computes the optimal rejection function. The following lemma is stated without its (technical) proof.

Lemma 5.1 *If Q minimizes $\rho(Q)$ and meets the constraint $D(Q||P) \geq \Lambda$ then: (i) $D(Q||P) = \Lambda$; (ii) $p_j < p_k$ and $q_k > 0 \Rightarrow r(j) > r(k)$; (iii) $p_j < p_k$ and $q_j > 0 \Rightarrow q_j \log \frac{q_j}{p_j} + q_k \log \frac{q_k}{p_k} > (q_j + q_k) \log \frac{q_j+q_k}{p_k}$; (iv) $p_j < p_k$ and $q_j > 0 \Rightarrow \frac{q_j}{p_j} > \frac{q_k}{p_k}$; and (v) $q_j, q_k > 0 \Rightarrow p_j \neq p_k$.*

Theorem 5.2 *Any optimal adversarial strategy Q has an effective support of size at most two.*

Proof Sketch: Assume by contradiction that an optimal Q^* has an effective support of size $J \geq 3$. W.l.o.g. we rename events such that the first J events are the effective support of Q^* (i.e., $q^*_i > 0$, $i = 1, \ldots, J$). From part (i) of Lemma 5.1, Q^* is a global minimizer of $\rho(Q)$ subject to the constraints $\sum_{i=1}^{J} q_i \log \frac{q_i}{p_i} = \Lambda$, $q_i > 0$ ($i = 1, \ldots, J$) and $\sum_{i=1}^{J} q_i = 1$. The Lagrangian of this problem is

$$L(Q, \lambda) = \sum_{i=1}^{J} r(i) q_i + \lambda_1 \left(\sum_{i=1}^{J} q_i \log \frac{q_i}{p_i} - \Lambda \right) + \lambda_2 \left(\sum_{i=1}^{J} q_i - 1 \right). \tag{1}$$

It is not hard to show, using parts (iv) and (v) of Lemma 5.1, that Q^* is an extremum point of (1). Taking the partial derivatives of (1) we have: $\frac{\partial L(Q^*,\lambda)}{\partial q_i} = r(i) + \lambda_1 \left(\log \frac{q^*_i}{p_i} + 1 \right) + \lambda_2 = 0$. Solving $\frac{\partial L(Q^*,\lambda)}{\partial q_1} = \frac{\partial L(Q^*,\lambda)}{\partial q_2}$ for λ_1, we get $\lambda_1 = (r(2) - r(1))/(\log \frac{q^*_1}{p_1} - \log \frac{q^*_2}{p_2})$. If we assume (w.l.o.g.) that $p_1 < p_2$, then, from parts (ii) and (iv) of Lemma 5.1, $r(2) < r(1)$ and $q^*_1/p_1 > q^*_2/p_2$. Thus $\lambda_1 < 0$. Therefore, for all i, $\frac{\partial^2 L(Q,\lambda)}{\partial q_i^2} = \frac{\lambda_1}{q_i} < 0$, and (1) is strictly concave. Therefore, since Q^* is an extremum of the (strictly concave) Lagrangian function, it is the unique global maximum.

By part (iv) of Lemma 5.1, the smooth function $f_{P,\Lambda}(q_1, q_2, \ldots, q_{J-1}) \triangleq D(Q||P) - \Lambda$ has a root at Q^* where no partial derivative is zero. Therefore, it has an infinite number of roots in any convex

[5] For any pair j, k such that $p_j \leq p_k$, $D(Q||P)$ does not decrease by transferring all the probability from k to j in Q: $q_j \log \frac{q_j}{p_j} + q_k \log \frac{q_k}{p_k} \leq (q_j + q_k) \log \frac{q_j+q_k}{p_j}$.

domain where Q^* is an internal point. Thus, there exists another distribution, $\tilde{Q} \neq Q^*$, where $\tilde{q}_i > 0$ for $i = 1, \ldots, J$, which meets the equality criteria of the Lagrangian. Since Q^* is the unique global maximum of $L(Q, \lambda)$: $\rho(\tilde{Q}) = L(\tilde{Q}, \lambda) < L(Q^*, \lambda) = \rho(Q^*)$. Contradiction. □

We now turn our attention to the learner's selection of $r(\cdot)$. As already noted, it is sufficient for the learner to consider only monotone rejection functions. Since for these functions $p_j = p_k \Rightarrow r(j) = r(k)$, the learner can partition Ω into $K = K(P)$ event subsets, which correspond, by probability, to "level sets", $S_1, S_2, \ldots, S_K$ (all events in a level set S have probability P_S). We re-index these subsets such that $0 < P_{S_1} < P_{S_2} < \cdots < P_{S_K}$. Define K variables $r_1, r_2, \ldots, r_K$, representing the rejection rate assigned to each of the K level sets ($\forall \omega \in S_i, r(\omega) = r_i$). We group our level sets by probability: $L = \{S : P_S < 2^{-\Lambda}\}$, $M = \{S : P_S = 2^{-\Lambda}\}$, and $H = \{S : P_S > 2^{-\Lambda}\}$.

By Theorem 5.2, the optimal Q which the adversary selects will have an effective support of size 2 at most. If it has an effective support of size 1, then the event ω for which $q_\omega = 1$ cannot be from a level set in L or H (otherwise, part (i) of Lemma 5.1 would be violated). Therefore it must belong to the single level set in M. Thus, if $M = \{S_m\}$ (for some index m), then there are feasible solutions Q such that $q_\omega = 1$ (for $\omega \in S_m$), all of which have $\rho(Q) = r_m$. If, on the other hand, Q has an effective support of size 2, then it is not hard to show that one of the two events must be from a level set $S_l \in L$, and the other, from a level set $S_h \in H$ (since all other combinations result in a violation of either part (i) or part (iii) of Lemma 5.1). Then, there is a single solution to $q_l \log \frac{q_l}{P_{S_l}} + (1 - q_l) \log \frac{1-q_l}{P_{S_h}} = \Lambda$, where q_l and $1 - q_l$ are the probabilities that Q assigns to the events from S_l and S_h, respectively. For such a distribution, $\rho(Q) = q_l r_l + (1 - q_l) r_h$.

Therefore, the adversary's choice of an optimal distribution, Q, must have one of $|L||H| + |M| \leq \lceil \frac{K^2}{4} \rceil$ (possibly different) rejection rates. Each of these rates, $\rho_1, \rho_2, \ldots, \rho_{|L||H|+|M|}$, is a linear combination of at most two variables, r_i and r_j. We introduce an additional variable, z, to represent the max-min rejection rate. We thus have:

Theorem 5.3 *An optimal soft rejection function and the lower-bound on the Type II Error, $1 - z$, is obtained by solving the following linear program:*[6] *maximize$_{r_1, r_2, \ldots, r_K, z}$ z, subject to:*

$$\sum_{i=1}^{K} r_i |S_i| P_{S_i} = \delta, \;\; 1 \geq r_1 \geq r_2 \geq \cdots \geq r_K \geq 0, \;\; \rho_i \geq z, \; i \in \{1, 2, \ldots, |L||H| + |M|\}.$$

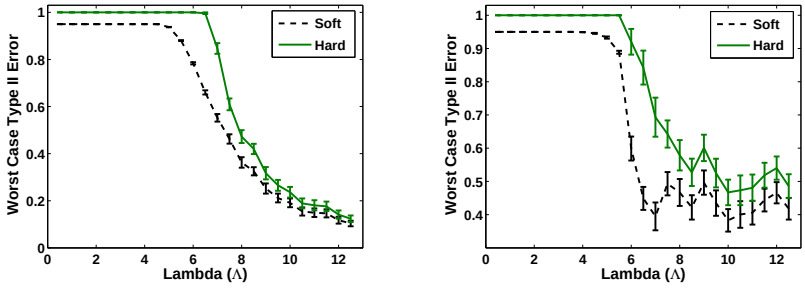
5.2.1 Numerical Examples

We now compare the performance of hard and soft rejection strategies for this constrained game ($D(Q||P) \geq \Lambda$) for various values of Λ, and two different families of target distributions, P over support $N = 50$. The families are arbitrary probability mass functions over N events and discretized Gaussians (over N bins). For each Λ we generated 50 random distributions P for each of the families.[7] For each such P we solved the optimal hard and soft strategies and computed the corresponding worst-case optimal type II error, $1 - \rho(Q)$.

The results for $\delta = 0.05$ are shown in Figure 5.2.1. Other results (not presented) for a wide variety of the problem parameters (e.g., N, δ) are qualitatively the same. It is evident that both the soft and hard strategies are ineffective for small Λ. Clearly, the soft approach has significantly lower error than the hard approach (until Λ becomes "sufficiently large").

[6] Let r^* be the solution to the linear program. Our derivation of the linear program is dependent on the assumption that there is no event $j \in I_{min}(r^*)$ such that $p_j \leq 2^{-\Lambda}$ (see discussion preceding Lemma 5.1). If r^* contradicts this assumption then, as discussed, the optimal strategy is r^δ, which is optimal. It is not hard to prove that in this case $r^* = r^\delta$ anyway, and thus the solution to the linear program is always optimal.

[7] Since $\max_Q D(Q||P) = \log(1/\min_i p_i)$, it is necessary that $\min_i p_i \leq 2^{-\Lambda}$ when generating P (to ensure that a Λ-distant Q exists). Distributions in the first family of arbitrarily random distributions (a) are generated by sampling a point (p_1) uniformly in $(0, 2^{-\Lambda}]$. The other $N - 1$ points are drawn i.i.d. $\sim U(0, 1]$, and then normalized so that their sum is $1 - p_1$. The second family (b) are Gaussians centered at 0 and discretized over N evenly spaced bins in the range $[-10, 10]$. A (discretized) random Gaussian $N(0, \sigma)$ is selected by choosing σ uniformly in some range $[\sigma_{min}, \sigma_{max}]$. σ_{min} is set to the minimum σ ensuring that the first/last bin will not have "zero" probability (due to limited precision). σ_{max} was set so that the cumulative probability in the first/last bin will be $2^{-\Lambda}$, if possible (otherwise σ_{max} is arbitrarily set to $10 * \sigma_{min}$).



(a) Arbitrary (b) Gaussians

Figure 1: Type II Error vs. Λ, for $N = 50$ and $\delta = 0.05$. 50 distributions were generated for each value of Λ ($\Lambda = 0.5, 0.1, \cdots, 12.5$). Error bars depict standard error of the mean (SEM).

6 Concluding Remarks

We have introduced a game-theoretic approach to the SCC problem. This approach lends itself well to analysis, allowing us to prove under what conditions low-density rejection is hard-optimal and if an optimal monotone rejection function is guaranteed to exist. Our analysis introduces soft decision strategies, which allow for significantly better performance. Observing the learner's futility when facing an omniscient and unlimited adversary, we considered restricted adversaries and provided full analysis of an interesting family of constrained games. This work opens up many new avenues for future research. We believe that our results could be useful for inspiring new algorithms for finite-sample SCC problems. For example, the equivalence of low-density rejection to the Bayesian binary problem as shown in Section 3.3 obviously motivates a new approach. Clearly, the utilization of randomized strategies should be carried over to the finite sample case as well. Our approach can be extended and developed in several ways. A very interesting setting to consider is one in which the adversary has partial knowledge of the problem parameters and the learner's strategy. For example, the adversary may only know that P is in some subspace. Additionally, it is desirable to extend our analysis to infinite and continuous event spaces. Finally, it would be very nice to determine an explicit expression for the lower bound obtained by the linear program of Theorem 5.3.

References

[1] S. Ben-David and M. Lindenbaum. Learning distributions by their density-levels - a paradigm for learning without a teacher. In *EuroCOLT*, pages 53–68, 1995.

[2] C.M. Bishop. Novelty detection and neural network validation. *IEE Proceedings - Vision, Image, and Signal Processing*, 141(4):217–222, 1994.

[3] M.M. Breunig, H.P. Kriegel, R.T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *SIGMOD Conference*, pages 93–104, 2000.

[4] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[5] G.R.G. Lanckriet, L. El Ghaoui, and M.I. Jordan. Robust novelty detection with single-class mpm. In *NIPS*, pages 905–912, 2002.

[6] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *SDM*, 2003.

[7] M. Markou and S. Singh. Novelty detection: a review – part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[8] M. Markou and S. Singh. Novelty detection: a review – part 2: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[9] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6, 2005.

[10] David M. J. Tax and Robert P. W. Duin. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2:155–173, 2002.

[11] H. Yu. Single-class classification with mapping convergence. *Machine Learning*, 61(1-3):49–69, 2005.