Learning with Rejection

Corinna Cortes[1], Giulia DeSalvo[2], and Mehryar Mohri[2,1]

[1] Google Research, 111 8th Avenue, New York, NY
[2] Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY

Abstract. We introduce a novel framework for classification with a rejection option that consists of simultaneously learning two functions: a classifier along with a rejection function. We present a full theoretical analysis of this framework including new data-dependent learning bounds in terms of the Rademacher complexities of the classifier and rejection families as well as consistency and calibration results. These theoretical guarantees guide us in designing new algorithms that can exploit different kernel-based hypothesis sets for the classifier and rejection functions. We compare and contrast our general framework with the special case of confidence-based rejection for which we devise alternative loss functions and algorithms as well. We report the results of several experiments showing that our kernel-based algorithms can yield a notable improvement over the best existing confidence-based rejection algorithm.

1 Introduction

We consider a flexible binary classification scenario where the learner is given the option to reject an instance instead of predicting its label, thereby incurring some pre-specified cost, typically less than that of a random prediction. While classification with a rejection option has received little attention in the past, it is in fact a scenario of great significance that frequently arises in applications. Incorrect predictions can be costly, especially in applications such as medical diagnosis and bioinformatics. In comparison, the cost of abstaining from prediction, which may be that of additional medical tests, or that of routing a call to a customer representative in a spoken-dialog system, is often more acceptable. From a learning perspective, abstaining from fitting systematic outliers can also result in a more accurate predictor. Accurate algorithms for learning with rejection can further be useful to developing solutions for other learning problems such as active learning [4].

Various problems related to the scenario of learning with a rejection option have been studied in the past. The trade-off between error rate and rejection rate was first studied by Chow [5, 6] who also provided an analysis of the Bayes optimal decision for this setting. Later, several publications studied an optimal rejection rule based on the ROC curve and a subset of the training data [16, 29, 26], while others used rejection options or *punting* to reduce misclassification rate [15, 27, 2, 20, 24], though with no theoretical analysis or guarantee.

More generally, few studies have presented general error bounds in this area, but some have given risk bounds for specific scenarios. Freund et al. [14] studied

an ensemble method and presented an algorithm that predicts with a weighted average of the hypotheses while abstaining on some examples without incurring a cost. Herbei and Wegkamp [18] considered classification with a rejection option that incurs a cost and provided bounds for these ternary functions.

One of the most influential works in this area has been that of Bartlett and Wegkamp [1] who studied a natural discontinuous loss function taking into account the cost of a rejection. They used consistency results to define a convex and continuous *Double Hinge Loss* (DHL) surrogate loss upper-bounding that rejection loss, which they also used to derive an algorithm. A series of follow-up articles further extended this publication, including [33] which used the same convex surrogate while focusing on the l_1 penalty. Grandvalet et al. [17] derived a convex surrogate based on [1] that aims at estimating conditional probabilities only in the vicinity of the threshold points of the optimal decision rule. They also provided some preliminary experimental results comparing the DHL algorithm and their variant with a naive rejection algorithm. Under the same rejection rule, Yuan and Wegkamp [32] studied the infinite sample consistency for classification with a reject option.

Using a different approach based on active learning, El-Yaniv and Wiener [11] studied the trade-off between the coverage and accuracy of classifiers and, in a subsequent paper ([12]) provided a strategy to learn a certain type of selective classification, which they define as *weakly optimal*, that has diminishing rejection rate under some Bernstein-type conditions. Finally, several papers have discussed learning with rejection in the multi-class setting [28, 10, 3], reinforcement learning [22], and in online learning [34].

There are also several learning scenarios tangentially related to the rejection scenario we consider, though they are distinct and hence require a very different approach. Sequential learning with budget constraints is a related framework that admits two stages: first a classifier is learned, next the classifier is fixed and a rejection function is learned [30, 31]. Since it assumes a fixed predictor and only admits the rejection function as an argument, the corresponding loss function is quite different from ours. Another somewhat similar approach is that of cost-sensitive learning where a class-dependent cost can be used [13]. One could think of adopting that framework here to account for the different costs for rejection and incorrect prediction. However, the cost-sensitive framework assumes a distribution over the classes or labels, which, here, would be the set $\{-1, 1, ®\}$, with ® the rejection symbol. But, ® is not a class and there is no natural distribution over that set in our scenario.

In this paper, we introduce a novel framework for classification with a rejection option that consists of simultaneously learning a pair of functions (h, r): a predictor h along with a rejection function r, each selected from a different hypothesis set. This is a more general framework than that the special case of confidence-based rejection studied by Bartlett and Wegkamp [1] and others, where the rejection function is constrained to be a thresholded function of the predictor's scores. Our novel framework opens up a new perspective on the problem of learning with rejection for which we present a full theoretical analy-

sis, including new data-dependent learning bounds in terms of the Rademacher complexities of the classifier and rejection families, as well as consistency and calibration results. We derive convex surrogates for this framework that are realizable $(\mathcal{H}, \mathcal{R})$-consistent. These guarantees in turn guide the design of a variety of algorithms for learning with rejection. We describe in depth two different types of algorithms: the first type uses kernel-based hypothesis classes, the second type confidence-based rejection functions. We report the results of experiments comparing the performance of these algorithms and that of the DHL algorithm.

The paper is organized as follows. Section 2 introduces our novel learning framework and contrasts it with that of Bartlett and Wegkamp [1]. Section 3 provides generalization guarantees for learning with rejection. It also analyzes two convex surrogates of the loss along with consistency results and provides margin-based learning guarantees. In Section 4, we present an algorithm with kernel-based hypothesis sets derived from our learning bounds. In Section 5, we further examine the special case of confidence-based rejection by analyzing various algorithmic alternatives. Lastly, we report the results of several experiments comparing the performance of our algorithms with that of DHL (Section 6).

2 Learning problem

Let $\mathcal{X}$ denote the input space. We assume as in standard supervised learning that training and test points are drawn i.i.d. according to some fixed yet unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$. We present a new general model for learning with rejection, which includes the confidence-based models as a special case.

2.1 General rejection model

The learning scenario we consider is that of binary classification with rejection. Let ® denote the rejection symbol. For any given instance $x \in \mathcal{X}$, the learner has the option of abstaining or *rejecting* that instance and returning the symbol ®, or assigning to it a label $\widehat{y} \in \{-1, +1\}$. If the learner rejects an instance, it incurs some loss $c(x) \in [0, 1]$; if it does not reject but assigns an incorrect label, it incurs a cost of one; otherwise, it suffers no loss. Thus, the learner's output is a pair (h, r) where $h \colon \mathcal{X} \to \mathbb{R}$ is the hypothesis used for predicting a label for points not rejected using $\operatorname{sign}(h)$ and where $r \colon \mathcal{X} \to \mathbb{R}$ is a function determining the points $x \in \mathcal{X}$ to be rejected according to $r(x) \leq 0$.

The problem is distinct from a standard multi-class classification problem since no point is inherently labeled with ®. Its natural loss function L is defined by

$$L(h, r, x, y) = 1_{yh(x) \leq 0} 1_{r(x) > 0} + c(x) 1_{r(x) \leq 0}, \tag{1}$$

for any pair of functions (h, r) and labeled sample $(x, y) \in \mathcal{X} \times \{-1, +1\}$, thus extending the loss function considered by [1]. In what follows, we assume for simplicity that c is a constant function, though part of our analysis is applicable to the general case. Observe that for $c \geq \frac{1}{2}$, on average, there is no incentive for rejection since a random guess can never incur an expected cost of more than $\frac{1}{2}$.

$h^*(x) = \eta(x) - \frac{1}{2}$ and
$r^*(x) = |h^*(x)| - (\frac{1}{2} - c)$.

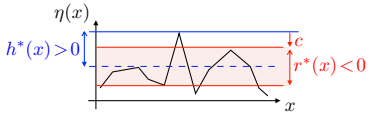

Fig. 1. Mathematical expression and illustration of the optimal classification and rejection function for the Bayes solution. Note, as c increases, the rejection region shrinks.

For biased distributions, one may further limit c to the fraction of the smallest class. For $c = 0$, we obtain a trivial solution by rejecting all points, so we restrict c to the case of $c \in]0, \frac{1}{2}[$.

Let $\mathcal{H}$ and $\mathcal{R}$ denote two families of functions mapping $\mathcal{X}$ to $\mathbb{R}$. The learning problem consists of using a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn i.i.d. from $\mathcal{D}^m$ to determine a pair $(h, r) \in \mathcal{H} \times \mathcal{R}$ with a small expected rejection loss $R(h, r)$

$$R(h, r) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[1_{yh(x)\leq 0} 1_{r(x)>0} + c 1_{r(x)\leq 0}\right]. \tag{2}$$

We denote by $\widehat{R}_S(h, r)$ the empirical loss of a pair $(h, r) \in \mathcal{H} \times \mathcal{R}$ over the sample S and use $(x, y) \sim S$ to denote the draw of (x, y) according to the empirical distribution defined by S: $\widehat{R}_S(h, r) = \mathbb{E}_{(x,y)\sim S}\left[1_{yh(x)\leq 0} 1_{r(x)>0} + c 1_{r(x)\leq 0}\right]$.

2.2 Confidence-based rejection model

Learning with rejection based on two independent yet jointly learned functions h and r introduces a completely novel approach to this subject. However, our new framework encompasses much of the previous work on this problem, e.g. [1], is a special case where rejection is based on the magnitude of the value of the predictor h, that is $x \in \mathcal{X}$ is rejected if $|h(x)| \leq \gamma$ for some $\gamma \geq 0$. Thus, r is implicitly defined in the terms of the predictor h by $r(x) = |h(x)| - \gamma$.

This specific choice of the rejection function r is natural when considering the Bayes solution (h^*, r^*) of the learning problem, that is the one where the distribution $\mathcal{D}$ is known. Indeed, for any $x \in \mathcal{X}$, let $\eta(x)$ be defined by $\eta(x) = \mathbb{P}[Y = +1|x]$. For a standard binary classification problem, it is known that the predictor h^* defined for any $x \in \mathcal{X}$ by $h^*(x) = \eta(x) - \frac{1}{2}$ is optimal since $\text{sign}(h^*(x)) = \max\{\eta(x), 1 - \eta(x)\}$. For any $x \in \mathcal{X}$, the misclassification loss of h^* is $\mathbb{E}[1_{yh(x)\leq 0}|x] = \min\{\eta(x), 1 - \eta(x)\}$. The optimal rejection r^* should therefore be defined such that $r^*(x) \leq 0$, meaning x is rejected, if and only if

$$\begin{aligned}
\min\{\eta(x), 1 - \eta(x)\} \geq c &\Leftrightarrow 1 - \max\{\eta(x), 1 - \eta(x)\} \geq c \\
&\Leftrightarrow \max\{\eta(x), 1 - \eta(x)\} \leq 1 - c \\
&\Leftrightarrow \max\{\eta(x) - \tfrac{1}{2}, \tfrac{1}{2} - \eta(x)\} \leq \tfrac{1}{2} - c \Leftrightarrow |h^*(x)| \leq \tfrac{1}{2} - c.
\end{aligned}$$

Thus, we can choose h^* and r^* as in Figure 1, which also provides an illustration of confidence-based rejection. However, when predictors are selected out of a limited subset $\mathcal{H}$ of all measurable functions over $\mathcal{X}$, requiring the rejection function r to be defined as $r(x) = |h(x)| - \gamma$, for some $h \in \mathcal{H}$, can be too restrictive. Consider, for example, the case where $\mathcal{H}$ is a family of linear functions.

Fig. 2. The best predictor h is defined by the threshold θ: $h(x) = x - \theta$. For $c < \frac{1}{2}$, the region defined by $X \leq \eta$ should be rejected. Note that the corresponding rejection function r defined by $r(x) = x - \eta$ cannot be defined as $|h(x)| \leq \gamma$ for some $\gamma > 0$.

Figure 2 shows a simple case in dimension one where the optimal rejection region cannot be defined simply as a function of the best predictor h. The model for learning with rejection that we describe where a pair (h, r) is selected is more general. In the next section, we study the problem of learning such a pair.

3 Theoretical analysis

We first give a generalization bound for the problem of learning with our rejection loss function as well as consistency results. Next, to devise efficient learning algorithms, we give general convex upper bounds on the rejection loss. For several of these convex surrogate losses, we prove margin-based guarantees that we subsequently use to define our learning algorithms (Section 4).

3.1 Generalization bound

Theorem 1. *Let $\mathcal{H}$ and $\mathcal{R}$ be families of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample S of size m from $\mathcal{D}$, the following holds for all $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

$$R(h, r) \leq \widehat{R}_S(h, r) + \mathfrak{R}_m(\mathcal{H}) + (1 + c)\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof. Let $\mathcal{L}_{\mathcal{H},\mathcal{R}}$ be the family of functions $\mathcal{L}_{\mathcal{H},\mathcal{R}} = \big\{(x, y) \mapsto L(h, r, x, y), h \in \mathcal{H}, r \in \mathcal{R}\big\}$. Since the loss function L takes values in $[0, 1]$, by the general Rademacher complexity bound [19], with probability at least $1-\delta$, the following holds for all $(h, r) \in \mathcal{H} \times \mathcal{R}$: $R(h, r) \leq \widehat{R}_S(h, r) + 2\mathfrak{R}_m(\mathcal{L}_{\mathcal{H},\mathcal{R}}) + \sqrt{\frac{\log 1/\delta}{2m}}$. Now, the Rademacher complexity can be bounded as follows:

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{L}_{\mathcal{H},\mathcal{R}}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{(h,r)\in\mathcal{H}\times\mathcal{R}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{y_i h(x_i)\leq 0} 1_{r(x_i)>0} + \sigma_i c\, 1_{r(x_i)\leq 0}\right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{(h,r)\in\mathcal{H}\times\mathcal{R}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x_i)\neq y_i} 1_{r(x_i)=+1}\right] + c \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{r(x_i)=-1}\right].
\end{aligned}$$

By Lemma 1 (below), the Rademacher complexity of products of indicator functions can be bounded by the sum of the Rademacher complexities of each indicator function class, thus, $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{(h,r)\in\mathcal{H}\times\mathcal{R}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x_i)\neq y_i} 1_{r(x_i)=+1}\right] \leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x_i)\neq y_i}\right] + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{r(x_i)=+1}\right]$. The proof

can be completed by using the known fact that the Rademacher complexity of indicator functions based on a family of functions taking values in $\{-1,+1\}$ is equal to one half the Rademacher complexity of that family. □

To derive an explicit bound in terms of $\mathcal{H}$ and $\mathcal{R}$ in Theorem 1, we make use of the following lemma relating the Rademacher complexity of a product of two (or more) families of functions to the sum of the Rademacher complexity of each family, whose proof can be found in [9].

Lemma 1. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two families of functions mapping $\mathcal{X}$ to $[-1,+1]$. Let $\mathcal{F} = \{f_1 f_2 \colon f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. Then, the empirical Rademacher complexities of $\mathcal{F}$ for any sample S of size m are bounded: $\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq 2\big(\widehat{\mathfrak{R}}_S(\mathcal{F}_1) + \widehat{\mathfrak{R}}_S(\mathcal{F}_2)\big)$.*

The theorem gives generalization guarantees for learning with a family of predictors $\mathcal{H}$ and rejection function $\mathcal{R}$ mapping to $\{-1,+1\}$ that admit Rademacher complexities in $O(1/\sqrt{m})$. For such families, it suggests to select the pair (h,r) to minimize the right-hand side. As with the zero-one loss, minimizing $\widehat{R}_S(h,r)$ is computationally hard for most families of functions. Thus, in the next section, we study convex upper bounds that lead to more efficient optimization problems, while admitting favorable learning guarantees as well as consistency results.

3.2 Convex surrogate losses

We first present general convex upper bounds on the rejection loss. Let $u \mapsto \Phi(-u)$ and $u \mapsto \Psi(-u)$ be convex functions upper-bounding $1_{u \leq 0}$. Since for any $a, b \in \mathbb{R}$, $\max(a,b) = \frac{a+b+|b-a|}{2} \geq \frac{a+b}{2}$, the following inequalities hold with $\alpha > 0$ and $\beta > 0$:

$$
\begin{aligned}
L(h,r,x,y) &= 1_{yh(x)\leq 0} 1_{r(x)>0} + c\, 1_{r(x)\leq 0} = \max\left(1_{yh(x)\leq 0} 1_{-r(x)<0}, c\, 1_{r(x)\leq 0}\right) \\
&\leq \max\left(1_{\max(yh(x),-r(x))\leq 0}, c\, 1_{r(x)\leq 0}\right) \leq \max\left(1_{\frac{yh(x)-r(x)}{2}\leq 0}, c\, 1_{r(x)\leq 0}\right) \\
&\leq \max\left(1_{\alpha\frac{yh(x)-r(x)}{2}\leq 0}, c\, 1_{\beta r(x)\leq 0}\right) \\
&\leq \max\left(\Phi\left(\tfrac{\alpha}{2}(r(x)-yh(x))\right), c\Psi(-\beta r(x))\right) \qquad (3) \\
&\leq \Phi\left(\tfrac{\alpha}{2}(r(x)-yh(x))\right) + c\Psi(-\beta r(x)). \qquad (4)
\end{aligned}
$$

Since Φ and Ψ are convex, their composition with an affine function of h and r is also a convex function of h and r. Since the maximum of two convex functions is convex, the right-hand side of (3) is a convex function of h and r. Similarly, the right-hand side of (4) is a convex function of h and r. In the specific case where the Hinge loss is used for both $u \mapsto \Phi(-u)$ and $u \mapsto \Psi(-u)$, we obtain the following two convex upper bounds, Max Hinge (MH) and Plus Hinge (PH), also illustrated in Figure 3:

$$
\begin{aligned}
L_{\mathrm{MH}}(h,r,x,y) &= \max\left(1 + \tfrac{\alpha}{2}(r(x)-yh(x)), c\,(1-\beta r(x)), 0\right) \\
L_{\mathrm{PH}}(h,r,x,y) &= \max\left(1 + \tfrac{\alpha}{2}(r(x)-yh(x)), 0\right) + \max\left(c\,(1-\beta r(x)), 0\right).
\end{aligned}
$$

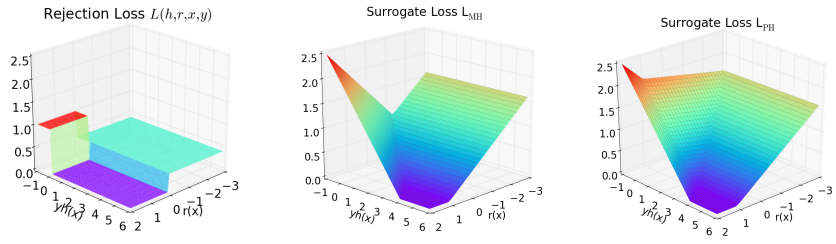

Fig. 3. From the left, the figures show the rejection loss L, the convex surrogate loss L_{MH}, and the convex surrogate loss L_{PH} as a function of $yh(x)$ and $r(x)$, for the cost value $c = 0.4$. The convex surrogates have a steeper left surface reflecting the rejection loss's penalty of incorrectly classifying a point while their gentler right surface of the surrogates reflects the lower cost c of abstaining. Also, the figures clearly show that the surrogate loss L_{PH} is an upper bound on L_{MH}.

3.3 Consistency results

In this section, we present a series of theoretical results related to the consistency of the convex surrogate losses introduced. We first prove the calibration and consistency for specific choices of the parameters α and β. Next, we show that the excess risk with respect to the rejection loss can be bounded by its counterpart defined via our surrogate loss. We further prove a general realizable $(\mathcal{H}, \mathcal{R})$-consistency for our surrogate losses.

Calibration. The constants $\alpha > 0$ and $\beta > 0$ are introduced in order to calibrate the surrogate loss with respect to the Bayes solution. Let $(h^*_{\mathrm{M}}, r^*_{\mathrm{M}})$ be a pair attaining the infimum of the expected surrogate loss $\mathbb{E}_{(x,y)}(L_{\mathrm{MH}}(h, r, x, y))$ over all measurable functions. Recall from Section 2, the Bayes classifier is denoted by (h^*, r^*). The following lemma shows that for $\alpha = 1$ and $\beta = \frac{1}{1-2c}$, the loss L_{MH} is calibrated, that is the sign of $(h^*_{\mathrm{M}}, r^*_{\mathrm{M}})$ matches the sign of (h^*, r^*).

Theorem 2. *Let (h^*_M, r^*_M) denote a pair attaining the infimum of the expected surrogate loss, $\mathbb{E}_{(x,y)}[L_{\mathrm{MH}}(h^*_M, r^*_M, x, y)] = \inf_{(h,r)\in meas} \mathbb{E}_{(x,y)}[L_{\mathrm{MH}}(h, r, x, y)]$. Then, for $\beta = \frac{1}{1-2c}$ and $\alpha = 1$,*

1. *the surrogate loss L_{MH} is calibrated with respect to the Bayes classifier: $\operatorname{sign}(h^*) = \operatorname{sign}(h^*_M)$ and $\operatorname{sign}(r^*) = \operatorname{sign}(r^*_M)$;*
2. *furthermore, the following equality holds for the infima over pairs of measurable functions:*

$$\inf_{(h,r)} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)] = (3 - 2c) \inf_{(h,r)} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L(h, r, x, y)].$$

Proof Sketch. The expected surrogate loss can be written in terms of $\eta(x)$: $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)] = \mathbb{E}_x[\eta(x)\phi(-h(x), r(x)) + (1-\eta(x))\phi(h(x), r(x))]$, with $\phi(-h(x), r(x)) = \max(1 + \frac{1}{2}(r(x) - h(x)), c(1 - \frac{1}{1-2c}r(x)), 0)$. Let the argument of the expectation, $\eta(x)\phi(-h(x), r(x)) + (1-\eta(x))\phi(h(x), r(x))$, be denoted

by $\mathcal{L}_\phi(\eta(x), h(x), r(x))$. Since the infimum is over all measurable functions, to determine $(h^*_{\mathrm{M}}, r^*_{\mathrm{M}})$ it suffices to determine, for any fixed x the minimizer of $(u, v) \mapsto \mathcal{L}_\phi(\eta(x), u, v)$. For a fixed x, minimizing $\mathcal{L}_\phi(\eta(x), u, v)$ with respect to (u, v) is equivalent to minimizing seven LPs. One can check that the optimal points of these LPs are in the set $(u, v) \in \{(0, (2c-2)(1-2c)), (3-2c, 1-2c), (-3+2c, 1-2c)\}$. Evaluating $\mathcal{L}_\phi(\eta(x), u, v)$ at these points, we find that $\mathcal{L}_\phi(\eta(x), 3-2c, 1-2c) = (3-2c)(1-\eta(x))$, $\mathcal{L}_\phi(\eta(x), -3+2c, 1-2c) = (3-2c)(\eta(x))$, and $\mathcal{L}_\phi(\eta(x), 0, (2c-2)(1-2c)) = (3-2c)c$. Thus, we can conclude that the minimum of $\mathcal{L}_\phi(\eta(x), u, v)$ is attained at $(3-2c)\big[\eta(x)1_{\eta(x)<c} + c1_{c\leq\eta(x)\leq 1-c} + \big(1-\eta(x)\big)1_{\eta(x)>1-c}\big]$, which completes the proof. □

Excess risk bound. Here, we show upper bounds on the excess risk in terms of the surrogate loss excess risk. Let R^* denote the Bayes rejection loss, that is $R^* = \inf_{(h,r)} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(h, r, x, y)]$, where the infimum is taken over all measurable functions and similarly let R^*_L denote $\inf_{(h,r)} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)]$.

Theorem 3. *Let $R_L(h, r) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)]$ denote the expected surrogate loss of a pair (h, r). Then, the surrogate excess of (h, r) is upper bounded by its surrogate excess error as follows:*

$$R(h, r) - R^* \leq \tfrac{1}{(1-c)(1-2c)}\big(R_L(h, r) - R^*_L\big).$$

Proof Sketch. Let $\mathcal{L}^*(\eta(x))$ denote the expected loss of the Bayes solution conditioned on x, $\mathcal{L}^*(\eta(x)) = \eta(x)1_{\eta(x)<c} + c1_{c\leq\eta(x)\leq 1-c} + (1-\eta(x))1_{\eta(x)>1-c}$. Then

$$\begin{aligned} R(h, r) - R(h^*, r^*) = \mathop{\mathbb{E}}_{x}\big[&(\eta(x) - \mathcal{L}^*(\eta(x)))1_{h(x)<0, r>(x)0} \\ &+ (1-\eta(x) - \mathcal{L}^*(\eta(x)))1_{h(x)\geq 0, r(x)>0} + (c - \mathcal{L}^*(\eta(x)))1_{r(x)\leq 0}\big]. \end{aligned} \tag{5}$$

Since $\mathcal{L}^*(\eta(x))$ admits three values, we can distinguish three cases and give a proof for each. When $c \leq \eta(x) \leq 1-c$, $\mathcal{L}^*(\eta(x)) = c$, that is $r^* \leq 0$ and $r^*_L \leq 0$, by calibration. In that case, Equation 5 can be written as $R(h, r) - R(h^*, r^*) = \mathbb{E}_x\big((\eta(x)-c)1_{h(x)<0, r(x)>0} + (1-\eta(x)-c)1_{h(x)\geq 0, r(x)>0}\big)$. Note that the indicator functions on the right-hand side are mutually exclusive, thus, it suffices to show that each component is bounded. □

$(\mathcal{H}, \mathcal{R})$-consistency. The standard notion of loss consistency does not take into account the hypothesis set H used since it assumes an optimization carried out over the set of all measurable functions. Long and Servedio [23] proposed instead a notion of *H-consistency* precisely meant to take the hypothesis set used into consideration. They showed empirically that using loss functions that are H-consistent can lead to significantly better performances than using a loss function known to be consistent. Here, we prove that our surrogate losses are *realizable* $(\mathcal{H}, \mathcal{R})$*-consistent*, a hypothesis-set-specific notion of consistency under our framework. The realizable setting in learning with rejection means that there exists a function that never rejects and correctly classifies

all points. A loss l is *realizable* $(\mathcal{H}, \mathcal{R})$*-consistent* if for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and any $\epsilon > 0$, there exists $\delta > 0$ such that if $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l(h, r, x, y)\right] - \inf_{(h,r)\in(\mathcal{H},\mathcal{R})} \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h, r, x, y)]\right| \leq \delta$, then $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(h, r, x, y)] \leq \epsilon$.

Theorem 4. *Let $(u, v) \mapsto \Phi(-u, -v)$ be a non-increasing function upper-bounding $(u, v) \mapsto 1_{u\leq 0}1_{v>0} + c1_{v\leq 0}$ such that for any fixed v, $\lim_{u\to+\infty} \Phi(-u, -v) = 0$ and for any fixed v, $u \mapsto \Phi(-u, -v)$ is bounded over $\mathbb{R}_+$. Let $(\mathcal{H}, \mathcal{R})$ be pair of families of functions mapping $\mathcal{X}$ to $\mathbb{R}$ where $\mathcal{H}$ is closed under multiplication by a positive scalar ($\mathcal{H}$ is a cone). Then, the loss function $(h, r, x, y) \mapsto \Phi(-yh(x), -r(x))$ is realizable $(\mathcal{H}, \mathcal{R})$-consistent.*

Proof. Let $\mathcal{D}$ be a distribution for which $(h^*, r^*) \in (\mathcal{H}, \mathcal{R})$ achieves zero error, thus $yh^*(x) > 0$ and $r^*(x) > 0$ for all x in the support of $\mathcal{D}$. Fix $\epsilon > 0$ and assume that $\left|\mathbb{E}\left[\Phi\big(-yh(x), -r(x)\big)\right] - \inf_{(h,r)\in(\mathcal{H},\mathcal{R})} \mathbb{E}\left[\Phi\big(-yh(x), -r(x)\big)\right]\right| \leq \epsilon$ for some $(h, r) \in (\mathcal{H}, \mathcal{R})$. Then, since $1_{u\leq 0}1_{v>0} + c1_{v\leq 0} \leq \Phi(-u, -v)$ and since μh^* is in $\mathcal{H}$ for any $\mu > 0$, the following holds for any $\mu > 0$:

$$\begin{aligned}\mathbb{E}\left[L(h, r, x, y)\right] &\leq \mathbb{E}\left[\Phi\big(-yh(x), -r(x)\big)\right] \leq \mathbb{E}\left[\Phi\big(-\mu yh^*(x), -r^*(x)\big)\right] + \epsilon \\ &\leq \mathbb{E}\left[\Phi\big(-\mu yh^*(x), -r^*(x)\big)|r^*(x) > 0\right]\mathbb{P}[r^*(x) > 0] + \epsilon.\end{aligned}$$

Now, $u \mapsto \Phi(-\mu yh^*(x), -r^*(x))$ is bounded for $yh^*(x) > 0$ and $r^*(x) > 0$; since $\lim_{\mu\to+\infty} \Phi(-\mu yh^*(x), -r^*(x)) = 0$, by Lebesgue's dominated convergence theorem $\lim_{\mu\to+\infty} \mathbb{E}[\Phi(-\mu yh^*(x), -r^*(x))|r^*(x) > 0] = 0$. Thus, $\mathbb{E}[L(h, r, x, y)] \leq \epsilon$ for all $\epsilon > 0$, which concludes the proof. □

The conditions of the theorem hold in particular for the exponential and the logistic functions as well as hinge-type losses. Thus, the theorem shows that the general convex surrogate losses we defined are realizable $(\mathcal{H}, \mathcal{R})$-consistent when the functions Φ or Ψ are exponential or logistic functions.

3.4 Margin bounds

In this section, we give margin-based learning guarantees for the loss function L_{MH}. Since L_{PH} is a simple upper bound on L_{MH}, its margin-based learning bound can be derived similarly. In fact, the same technique can be used to derive margin-based guarantees for the subsequent convex surrogate loss functions we present.

For any $\rho, \rho' > 0$, the margin-loss associated to L_{MH} is given by $L_{\text{MH}}^{\rho,\rho'}(h, r, x, y) = \max\left(\max\left(1+\frac{\alpha}{2}\left(\frac{r(x)}{\rho'} - \frac{yh(x)}{\rho}\right), 0\right), \max\left(c\left(1-\beta\frac{r(x)}{\rho'}\right), 0\right)\right)$. The theorem enables us to derive margin-based learning guarantees. The proof requires dealing with this max-based surrogate loss, which is a non-standard derivation.

Theorem 5. *Let $\mathcal{H}$ and $\mathcal{R}$ be families of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample S of size m from $\mathcal{D}$, the following holds for all $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

$$R(h, r) \leq \mathop{\mathbb{E}}_{(x,y)\sim S}[L_{\text{MH}}(h, r, x, y)] + \alpha\mathfrak{R}_m(\mathcal{H}) + (2\beta c + \alpha)\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Proof. Let $\mathcal{L}_{\mathrm{MH},\mathcal{H},\mathcal{R}}$ be the family of functions defined by $\mathcal{L}_{\mathcal{H},\mathcal{R}} = \big\{(x,y) \mapsto \min\big(L_{\mathrm{MH}}(h,r,x,y),1\big), h \in \mathcal{H}, r \in \mathcal{R}\big\}$. Since $\min(L_{\mathrm{MH}},1)$ is bounded by one, by the general Rademacher complexity generalization bound [19], with probability at least $1-\delta$ over the draw of a sample S, the following holds:

$$R(h,r) \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\min(L_{\mathrm{MH}}(h,r,x,y),1)] \leq \mathop{\mathbb{E}}_{(x,y)\sim S}[\min(L_{\mathrm{MH}}(h,r,x,y),1)] +$$
$$2\mathfrak{R}_m(\mathcal{L}_{\mathrm{MH},\mathcal{H},\mathcal{R}}) + \sqrt{\frac{\log 1/\delta}{2m}} \leq \mathop{\mathbb{E}}_{(x,y)\sim S}[L_{\mathrm{MH}}(h,r,x,y)] + 2\mathfrak{R}_m(\mathcal{L}_{\mathrm{MH},\mathcal{H},\mathcal{R}}) + \sqrt{\frac{\log 1/\delta}{2m}}.$$

Observe that we can express L_{MH} as follows: $\max\big(\max\big(1+\frac{\alpha}{2}(r(x)-yh(x)),0\big), \max\big(c\,(1-\beta r(x)),0\big)\big)$. Therefore, since for any $a,b \in \mathbb{R}$, $\min\big(\max(a,b),1\big) = \max\big(\min(a,1),\min(b,1)\big)$, we can re-write $\min(L_{\mathrm{MH}},1)$ as:

$$\begin{aligned}&\max\big(\min\big(\max(1+\tfrac{\alpha}{2}(r(x)-yh(x)),0),1\big), \min\big(\max(c\,(1-\beta r(x)),0),1\big)\big)\\ &\leq \min\big(\max(1+\tfrac{\alpha}{2}(r(x)-yh(x)),0),1\big) + \min\big(\max(c\,(1-\beta r(x)),0),1\big).\end{aligned}$$

Since $u \mapsto \min\big(\max(1+\frac{\alpha u}{2},0),1\big)$ is $\frac{\alpha}{2}$-Lipschitz and $u \mapsto \min\big(\max(c\,(1-\beta u),0),1\big)$ is $c\beta$-Lipschitz, by Talagrand's contraction lemma [21],

$$\begin{aligned}\mathfrak{R}_m(L_{\mathrm{MH},\mathcal{H},\mathcal{R}}) &\leq \tfrac{\alpha}{2}\mathfrak{R}_m\Big(\big\{(x,y) \mapsto r(x)-yh(x)\big\}\Big) + \beta c\,\mathfrak{R}_m\big(\big\{(x,y)\mapsto r(x)\big\}\big)\\ &\leq \tfrac{\alpha}{2}\big(\mathfrak{R}_m(\mathcal{R}) + \mathfrak{R}_m(\mathcal{H})\big) + \beta c\,\mathfrak{R}_m(\mathcal{R}) = \frac{\alpha}{2}\mathfrak{R}_m(\mathcal{H}) + \big(\beta c + \tfrac{\alpha}{2}\big)\mathfrak{R}_m(\mathcal{R}),\end{aligned}$$

which completes the proof. □

The following corollary is then a direct consequence of the theorem above.

Corollary 1. *Let $\mathcal{H}$ and $\mathcal{R}$ be families of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Fix $\rho,\rho' > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{D}$, the following holds for all $(h,r) \in \mathcal{H}\times\mathcal{R}$:*

$$R(h,r) \leq \mathop{\mathbb{E}}_{(x,y)\sim S}[L_{\mathrm{MH}}^{\rho,\rho'}(h,r,x,y)] + \frac{\alpha}{\rho}\mathfrak{R}_m(\mathcal{H}) + \frac{2\beta c+\alpha}{\rho'}\mathfrak{R}_m(\mathcal{R}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Then, via [19], the bound of Corollary 1 can be shown to hold uniformly for all $\rho,\rho' \in (0,1)$, at the price of a term in $O\Big(\sqrt{\frac{\log\log 1/\rho}{m}} + \sqrt{\frac{\log\log 1/\rho'}{m}}\Big)$.

4 Algorithms for kernel-based hypotheses

In this section, we devise new algorithms for learning with a rejection option when $\mathcal{H}$ and $\mathcal{R}$ are kernel-based hypotheses. We use Corollary 1 to guide the optimization problems for our algorithms.

Let $\mathcal{H}$ and $\mathcal{R}$ be hypotheses sets defined in terms of PSD kernels K and K' over $\mathcal{X}$:

$$\mathcal{H} = \{x \to \boldsymbol{w}\cdot\boldsymbol{\Phi}(x)\colon \|\boldsymbol{w}\| \leq \Lambda\} \text{ and } \mathcal{R} = \{x \to \boldsymbol{u}\cdot\boldsymbol{\Phi}'(x)\colon \|\boldsymbol{u}\| \leq \Lambda'\},$$

where $\boldsymbol{\Phi}$ is the feature mapping associated to K and $\boldsymbol{\Phi}'$ the feature mapping associated to K' and where $\Lambda, \Lambda' \geq 0$ are hyperparameters. One key advantage of this formulation is that different kernels can be used to define $\mathcal{H}$ and $\mathcal{R}$, thereby providing a greater flexibility for the learning algorithm. In particular, when using a second-degree polynomial for the feature vector Φ', the rejection function corresponds to abstaining on an ellipsoidal region, which covers confidence-based rejection. For example, the Bartlett and Wegkamp [1] solution consists of choosing $\boldsymbol{\Phi}'(x) = \boldsymbol{\Phi}(x)$, $\boldsymbol{u} = \boldsymbol{w}$, and the rejection function, $r(x) = |h(x)| - \gamma$.

Corollary 2. *Let $\mathcal{H}$ and $\mathcal{R}$ be the hypothesis spaces as defined above. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample S of size m from $\mathcal{D}$, the following holds for all $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

$$R(h,r) \leq \mathop{\mathbb{E}}_{(x,y)\sim S}[L_{\mathrm{MH}}^{\rho,\rho'}(h,r,x,y)] + \alpha\sqrt{\frac{(\kappa\Lambda/\rho)^2}{m}} + (2\beta c + \alpha)\sqrt{\frac{(\kappa'\Lambda'/\rho')^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

where $\kappa^2 = \sup_{x\in\mathcal{X}} K(x,x)$ and $\kappa'^2 = \sup_{x\in\mathcal{X}} K'(x,x)$.

Proof. By standard kernel-based bounds on Rademacher complexity [25], we have that $\mathfrak{R}_m(\mathcal{H}) \leq \Lambda\sqrt{\frac{\mathrm{Tr}[\mathbf{K}]}{m}} \leq \sqrt{\frac{(\kappa\Lambda)^2}{m}}$ and similarly $\mathfrak{R}_m(\mathcal{R}) \leq \Lambda'\sqrt{\frac{\mathrm{Tr}[\mathbf{K}']}{m}} \leq \sqrt{\frac{(\kappa'\Lambda')^2}{m}}$. Applying this bounds to Corollary 1 completes the proof. □

This learning bound guides directly the definition of our first algorithm based on the L_{MH} (see full version [7] for details) resulting in the following optimization:

$$\min_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{\xi}} \ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda'}{2}\|\boldsymbol{u}\|^2 + \sum_{i=1}^{m}\xi_i \quad \text{subject to: } \xi_i \geq c(1 - \beta(\boldsymbol{u}\cdot\boldsymbol{\Phi}'(x_i) + b')),$$
$$\text{and} \quad \xi_i \geq 1 + \frac{\alpha}{2}\big(\boldsymbol{u}\cdot\boldsymbol{\Phi}'(x_i) + b' - y_i\boldsymbol{w}\cdot\boldsymbol{\Phi}(x_i) - b\big), \xi_i \geq 0,$$

where $\lambda, \lambda' \geq 0$ are parameters and b and b' are explicit offsets for the linear functions h and r. Similarly, we use the learning bound to derive a second algorithm based on the loss L_{PH} (see full paper [7]). We have implemented and tested the dual of both algorithms, which we will refer to as CHR algorithms (short for convex algorithms using $\mathcal{H}$ and $\mathcal{R}$ families). Both the primal and dual optimization are standard QP problems whose solution can be readily found via both general-purpose and specialized QP solvers. The flexibility of the kernel choice and the QP formulation for both primal and dual are key advantages of the CHR algorithms. In Section 6 we report experimental results with these algorithms as well as the details of our implementation.

5 Confidence-based rejection algorithms

In this section, we explore different algorithms for the confidence-based rejection model (Section 2.2). We thus consider a rejection function $r(x) = |h(x)| - \gamma$ that abstains on points classified with confidence less than a given threshold γ.

Table 1. For the DHL algorithm and the CHR algorithm of L_{MH} with cost values $c = 0.25$, we report the mean and standard deviations on the test set of the following quantities: the left two columns contain the rejection loss, the next two columns the fraction of points rejected, followed by two columns with the classification error on the non-rejected points. The rightmost column provides the error on the non-rejected points of the DHL algorithm if its rejection threshold is changed so it rejects the same fraction of points as the CHR algorithm.

Data Sets	Rejection loss DHL	Rejection loss CHR	Fraction rejected DHL	Fraction rejected CHR	Non-rejected error DHL	Non-rejected error CHR	Non-rejected err (incr. thrh.) DHL
cod	$0.176 \pm .030$	$0.098 \pm .037$	$0.186 \pm .055$	$0.024 \pm .028$	$0.130 \pm .043$	$0.092 \pm .039$	$0.186 \pm .033$
skin	$0.158 \pm .041$	$0.043 \pm .020$	$0.093 \pm .033$	$0.052 \pm .027$	$0.135 \pm .037$	$0.030 \pm .024$	$0.135 \pm .041$
bank	$0.061 \pm .022$	$0.030 \pm .006$	$0.066 \pm .016$	$0.036 \pm .022$	$0.045 \pm .018$	$0.021 \pm .008$	$0.044 \pm .016$
haber	$0.261 \pm .033$	$0.211 \pm .037$	$0.875 \pm .132$	$0.439 \pm .148$	$0.043 \pm .027$	$0.102 \pm .048$	$0.252 \pm .110$
pima	$0.241 \pm .025$	$0.171 \pm .017$	$0.055 \pm .007$	$0.700 \pm .055$	$0.227 \pm .025$	$0.043 \pm .023$	$0.112 \pm .060$
australian	$0.115 \pm .026$	$0.111 \pm .021$	$0.136 \pm .008$	$0.172 \pm .024$	$0.081 \pm .025$	$0.068 \pm .023$	$0.349 \pm .100$
liver	$0.236 \pm .040$	$0.248 \pm .005$	$0.397 \pm .047$	$0.980 \pm .019$	$0.136 \pm .044$	$0.003 \pm .006$	$0.292 \pm .120$

The most standard algorithm in this setting is the DHL algorithm, which is based on a double hinge loss, a hinge-type convex surrogate that has favorable consistency properties. The double hinge loss, L_{DHinge}, is an upper bound of the rejection loss only when $0 \leq \gamma \leq 1 - c$, making DHL algorithm only valid for these restricted γ values. Moreover, it is important to note that the hinge loss is in fact a tighter convex upper bound than the double hinge loss for these possible values of γ. We have $L_\gamma(h) \leq L_{\text{Hinge}}(h) \leq L_{\text{DHinge}}(h)$ where $L_\gamma(h) = 1_{yh(x)\leq 0} 1_{|h(x)|>\gamma} + c(x) 1_{|h(x)|\leq\gamma}$ is the rejection loss in this setting. Thus, a natural alternative to the DHL algorithm is simply minimizing the hinge loss. The DHL solves a QCQP optimization problem while the natural alternative solve a standard SVM-type dual.

The aforementioned confidence based algorithms only apply for $\gamma \in [0, 1-c]$ but a robust surrogate should majorate the rejection loss L_γ for all possible values. In [7] we present an algorithm that upper-bounds the rejection error for all values of $\gamma \in [0, 1]$. We provide further details of all these confidence-based algorithm as well as report several experimental results in [7]. While the alternative algorithms we described are based on tighter surrogate losses for the rejection loss than that of DHL, empirical evidence suggests that DHL outperforms these alternatives. Thus, in the experiments with our CHR algorithm, we will use DHL as the baseline for comparison (Section 6).

6 Experiments

In this section, we present the results of several experiments comparing our CHR algorithms with the DHL algorithm. All algorithms were implemented using CVX [8]. We tested the algorithms on seven data sets from the UCI data repository, specifically `australian`, `cod`, `skin`, `liver`, `banknote`, `haberman`, and `pima`. For each data set, we performed standard 5-fold cross-validation. We randomly divided the data into training, validation and test set in the ratio 3:1:1. We then repeated the experiments five times where each time we used a different random partition.

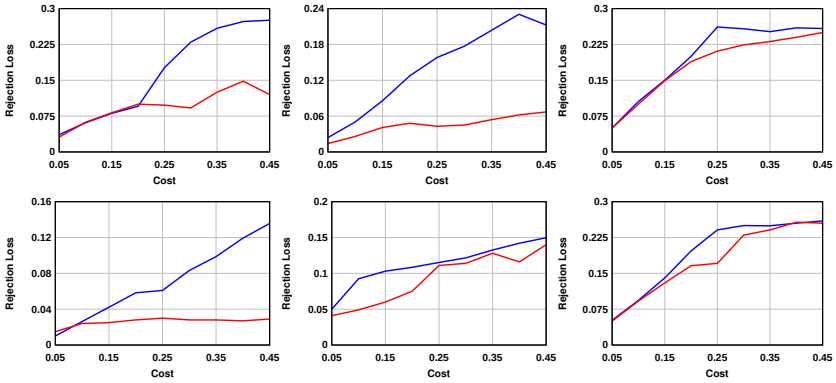

Fig. 4. Average rejection loss on the test set as a function of cost c for the DHL algorithm and the CHR algorithm for six datasets and polynomial kernels. The blue line is the DHL algorithm while the red line is the CHR algorithm based on L_{MH}. The figures on the top starting from the left are for the `cod`, `skin`, and `haberman` data set while the figures on the bottom are for `banknote`, `australian` and `pima` data sets. These figures show that the CHR algorithm outperforms the DHL algorithm for most values of cost, c, across all data sets.

The cost values ranged over $c \in \{0.05, 0.1, \ldots, 0.5\}$ and the kernels for both algorithms were polynomial kernels of degree $d \in \{1, 2, 3\}$ and Gaussian kernels with widths in the set $\{1, 10, 100\}$. The regularization parameters λ, λ' for the CHR algorithms varied over $\lambda, \lambda' \in \{10^i : i = -5, \ldots, 5\}$ and the threshold γ for DHL ranged over $\gamma \in \{0.08, 0.16, \ldots, 0.96\}$.

For each fixed value of c, we chose the parameters with the smallest average rejection loss on the validation set. For these parameter values, Table 1 shows the corresponding rejection loss on the test set for the CHR algorithm based on L_{MH} and the DHL algorithm both with cost $c = 0.25$. The table also shows the fraction of points rejected by each algorithm and the classification error on non-rejected points (see full paper version [7] for similar tables for all cost values). The rejection loss results of Table 1 show that the CHR algorithm yields an improvement in the rejection loss over the DHL algorithm. These findings are statistically significant at the 1% level or higher with one-sided paired t–test for all data sets except for the `liver` and `australian` data sets. Table 1 also reveals that the DHL algorithm rejects at a different rate than the CHR algorithm and often predicts the wrong label on the non-rejected points at a much higher rate. In order to level the playing field for the two algorithms, for the optimal settings of the DHL algorithm, we changed the rejection threshold till the fraction rejected by the DHL algorithm matched the fraction rejected by the CHR algorithm and recorded the error on the remaining non-rejected points. These results are included in the right-most column of Table 1 and demonstrate that the CHR algorithm rejects the hard cases and obtains a significantly better error rate on the remaining ones. In Figure 4, we show the rejection loss as a function of the cost for six of our data sets. These plots demonstrate that the

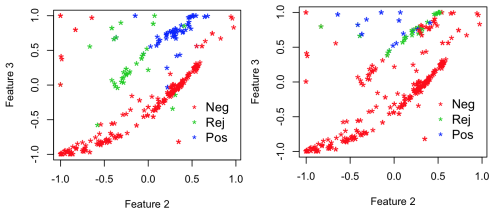

Fig. 5. The left figure shows CHR's classification of sample test points from the `skin` dataset with respect to different feature vectors. The right figure shows their classification by DHL and demonstrates how DHL rejects in areas of low confidence.

difference in accuracy between the two algorithms holds consistently for almost all values of c across all the data sets.

We also analyzed the rejection regions of the two algorithms. Unlike the DHL algorithm, we found that the CHR algorithms do not restrict their rejection regions to only areas of low confidence. On the other hand, the DHL algorithm only rejects around the boundary of the classification surface, see Figure 5. In [7], we further analyze the difference between the rejection functions found by the two algorithms. We also provide more results for the CHR algorithm including results for the CHR algorithm based on L_{PH}. We find that on average the CHR with L_{MH} performs slightly better than the CHR with L_{PH} as is expected since the loss L_{PH} is an upper bound of the loss L_{MH}.

7 Conclusion

We presented a detailed study of the problem of learning with rejection, which is a key question in a number of applications. We gave a general formulation of the problem for which we provided a theoretical analysis, including generalization guarantees, the derivation of different convex surrogates that are calibrated and consistent, and margin bounds that helped us devise new algorithms. The empirical results we reported demonstrate the effectiveness of our algorithms in several datasets. Our general formulation can further inspire the design of other algorithms as well as new theoretical insights and studies, one such a potential area being active learning. Furthermore, a natural extension of our framework is to include a constraint on the maximum fraction of points that can be rejected. Such an additional constraint will require new algorithms and generalization bounds.

Acknowledgments. This work was partly funded by NSF IIS-1117591, CCF-1535987, and DGE-1342536.

References

[1] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 2008.
[2] A. Bounsiar, E. Grall, and P. Beauseroy. Kernel based rejection method for supervised classification. In *WASET*, 2007.
[3] H. L. Capitaine and C. Frelicot. An optimum class-rejective decision rule and its evaluation. In *ICPR*, 2010.

[4] K. Chaudhuri and C. Zhang. Beyond disagreement-based agnostic active learning. In *NIPS*, 2014.
[5] C. Chow. An optimum character recognition system using decision function. *IEEE T. C.*, 1957.
[6] C. Chow. On optimum recognition error and reject trade-off. *IEEE T. C.*, 1970.
[7] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *arXiv*, 2016.
[8] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0, Aug. 2012.
[9] G. DeSalvo, M. Mohri, and U. Syed. Learning with deep cascades. In *ALT*, 2015.
[10] B. Dubuisson and M. Masson. Statistical decision rule with incomplete knowledge about classes. In *Pattern Recognition*, 1993.
[11] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *JMLR*, 2010.
[12] R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *NIPS*, 2011.
[13] C. Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.
[14] Y. Freund, Y. Mansour, and R. Schapire. Generalization bounds for averaged classifiers. *Ann. Stat.*, 2004.
[15] G. Fumera and F. Roli. Support vector machines with embedded reject option. In *ICPR*, 2002.
[16] G. Fumera, F. Roli, and G. Giacinto. Multiple reject thresholds for improving classification reliability. In *ICAPR*, 2000.
[17] Y. Grandvalet, J. Keshet, A. Rakotomamonjy, and S. Canu. Suppport vector machines with a reject option. In *NIPS*, 2008.
[18] R. Herbei and M. Wegkamp. Classification with reject option. *Can. J. Stat.*, 2005.
[19] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.
[20] T. Landgrebe, D. Tax, P. Paclik, and R. Duin. Interaction between classification and reject performance for distance-based reject-option classifiers. *PRL*, 2005.
[21] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York, 1991.
[22] M. Littman, L. Li, and T. Walsh. Knows what it knows: A framework for self-aware learning. In *ICML*, 2008.
[23] P. M. Long and R. A. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *ICML (3)*, pages 801–809, 2013.
[24] I. Melvin, J. Weston, C. S. Leslie, and W. S. Noble. Combining classifiers for improved classification of proteins from sequence or structure. *BMCB*, 2008.
[25] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.
[26] C. S. Pereira and A. Pires. On optimal reject rules and ROC curves. *PRL*, 2005.
[27] T. Pietraszek. Optimizing abstaining classifiers using ROC. In *ICML*, 2005.
[28] D. Tax and R. Duin. Growing a multi-class classifier with a reject option. In *Pattern Recognition Letters*, 2008.
[29] F. Tortorella. An optimal reject rule for binary classifiers. In *ICAPR*, 2001.
[30] K. Trapeznikov and V. Saligrama. Supervised sequential classification under budget constraints. In *AISTATS*, 2013.
[31] J. Wang, K. Trapeznikov, and V. Saligrama. An LP for sequential learning under budgets. In *JMLR*, 2014.
[32] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimizations. In *JMLR*, 2010.
[33] M. Yuan and M. Wegkamp. SVMs with a reject option. In *Bernoulli*, 2011.
[34] C. Zhang and K. Chaudhuri. The extended Littlestone's dimension for learning with mistakes and abstentions. In *COLT*, 2016.