Theory and Algorithms for Learning with Rejection in Binary Classification

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

[1*]Google Research, 111 8th Ave, New York, 10011, NY, USA.
[2]CIMS, NYU, 251 Mercer St, New York, 10012, NY, USA.

*Corresponding author(s). E-mail(s): giuliad@google.com;
Contributing authors: corinna@google.com; mohri@google.com;

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 our general framework with the special case of confidence-based rejection for which we also devise alternative loss functions and algorithms. 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.

Keywords: rejection, abstention, kernels, confidence-based models

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 [Hamid et al., 2017]. 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 [Chaudhuri and Zhang, 2014].

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 [1957, 1970] 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 [Fumera et al., 2000, Pereira and Pires, 2005, Tortorella, 2001], while others used rejection options or *punting* to reduce misclassification rate [Bounsiar et al., 2007, Fumera and Roli, 2002, Landgrebe et al., 2005, Melvin et al., 2008, Pietraszek, 2005], 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. [2004] 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 [2005] 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 [2008] 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 Yuan and Wegkamp [2011] which used the same convex surrogate while focusing on the l_1 penalty. Grandvalet et al. [2008] derived a convex surrogate based on Bartlett and Wegkamp [2008] 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 [2010] studied the infinite sample consistency for classification with a reject option. While in this work, we focus solely on binary classification, several works analyzed the learning with rejection framework in the multi-class classification setting focusing on different specific aspects: confidence-based predictors [Ni et al., 2019], a cross-entropy based loss function [Mozannar and Sontag, 2020a], and a one-vs-all classifier [Verma and Nalisnic, 2022].

Using a different approach based on active learning, El-Yaniv and Wiener [2010] studied the trade-off between the coverage and accuracy of classifiers and, in a subsequent paper El-Yaniv and Wiener [2011] 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 [Capitaine and Frelicot., 2010, Dubuisson and Masson, 1993, Tax and Duin, 2008], reinforcement learning [Littman et al., 2008], and in online learning [Zhang and Chaudhuri, 2016]. Several other papers that were published during the review process of our submission are closely aligned with our main topic [Cao et al., 2022, Mao et al., 2023a,b,c, Mohri et al., 2023, Mozannar and Sontag, 2020b, Mozannar et al., 2023, Narasimhan et al., 2023]. However, due to publications constraints, a detailed discussion of these papers will be deferred to our future work.

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 [Trapeznikov and Saligrama, 2013, Wang et al., 2014]. 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 [Elkan, 2001]. 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 $\{-1, 1, \mathtt{r}\}$, with $\mathtt{r}$ the rejection symbol. But, $\mathtt{r}$ is not a class and there is no natural distribution over that set in our scenario. For further discussion on this connection to cost-sensitive learning, please see Appendix C.

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 [2008] 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 analysis, 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 [2008]. 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.

The learning scenario we consider is that of binary classification with rejection. Let $\mathtt{r}$ 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 $\mathtt{r}$, 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 $\operatorname{sgn}(h)$ 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 $\mathtt{r}$. Its natural loss function L is defined by

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

where $\operatorname{sgn}(h(x)) = 1_{h(x) \geq 0} - 1_{h(x) < 0}$ or, equivalently, $\operatorname{sgn}(h(x)) = +1$ if $h(x) \geq 0$ and $\operatorname{sgn}(h(x)) = -1$ if $h(x) < 0$. This loss holds 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 Bartlett and Wegkamp [2008]. 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 $\frac{1}{2}$. 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 $c \in \left]0, \frac{1}{2}\right[$.

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

$$h^*(x) = \eta(x) - \tfrac{1}{2} \quad \text{and} \quad r^*(x) = |h^*(x)| - \left(\tfrac{1}{2} - c\right).$$

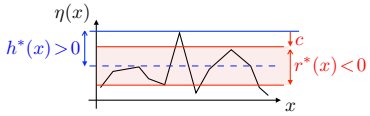

Fig. 1 Mathematical expression and illustration of the optimal classification and rejection function for the Bayes solution. Note, as c increases, the rejection region shrinks.

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 $R(h, r)$

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

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: $\widehat{R}_S(h,r) = \mathbb{E}_{(x,y)\sim S}\left[1_{y\neq \mathrm{sgn}(h(x))} 1_{r(x)>0} + c 1_{r(x)\leq 0}\right]$.

2.1 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. [Bartlett and Wegkamp, 2008], 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. For any $x \in \mathcal{X}$, the misclassification loss of h^* is $\mathbb{E}\left[1_{y\neq \mathrm{sgn}(h^*(x))}|x\right] = \min\{\eta(x), 1 - \eta(x)\}$. 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

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 $\gamma > 0$.

too restrictive. Consider, for example, the case where $\mathcal{H}$ is a family of linear functions. 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 $\mathcal{L}_{\mathcal{H},\mathcal{R}} = \{(x, y) \mapsto L(h, r, x, y), h \in \mathcal{H}, r \in \mathcal{R}\}$. Since the loss function L takes values in $[0, 1]$, by the general Rademacher complexity bound [Koltchinskii and Panchenko, 2002], 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:

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

By Lemma 2 (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,

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{(h,r)\in\mathcal{H}\times\mathcal{R}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i 1_{h(x_i)\neq y_i}1_{r(x_i)=+1}\right]$$
$$\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i 1_{h(x_i)\neq y_i}\right]+\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i 1_{r(x_i)=+1}\right].$$

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 DeSalvo et al. [2015].

Lemma 2 *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: $\widehat{\mathfrak{R}}_S(\mathcal{F})\leq 2\left(\widehat{\mathfrak{R}}_S(\mathcal{F}_1)+\widehat{\mathfrak{R}}_S(\mathcal{F}_2)\right)$.*

The theorem gives generalization guarantees for learning with a family of predictors $\mathcal{H}$ and rejection function $\mathcal{R}$ mapping to $\{-1,+1\}$. 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 $\beta>0$:

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

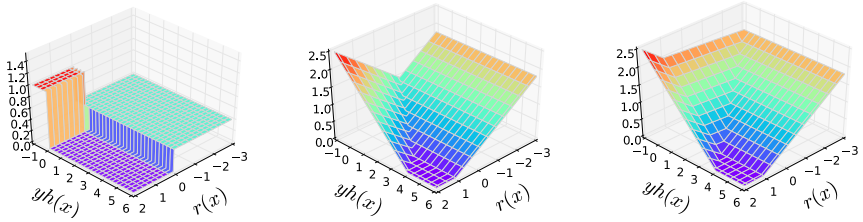

Fig. 3 From the left, the figures show the rejection loss L, the convex surrogate loss L_{MH}, and the convex surrogate loss L_{PH} as a function of $yh(x)$ and $r(x)$, for the cost value $c = 0.4$. The convex surrogates have a steeper left surface reflecting the rejection loss's penalty of incorrectly 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}.

$$\leq \Phi\left(\tfrac{\alpha}{2}\left(r(x) - yh(x)\right)\right) + c\,\Psi(-\beta r(x)). \tag{4}$$

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:

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

3.3 Consistency results

In this section, we present a series of theoretical results related to the consistency of the convex surrogate loss, L_{MH}. We first show that this convex surrogate is calibrated and then we show that the excess risk of the rejection loss can be bounded by the excess risk of the surrogate loss. See Appendix A for the proofs of these theorems. We also prove a general realizable $(\mathcal{H}, \mathcal{R})$-consistency theorem that holds for both of our surrogate losses.

3.3.1 Calibration

Below, we show that the constants $\alpha > 0$ and $\beta > 0$ in the definition of the surrogate loss can be chosen so that the surrogate loss is calibrated with respect to the Bayes solution. Let $(h^*_{\mathrm{M}}, r^*_{\mathrm{M}})$ be a pair attaining the infimum of the expected surrogate loss $\mathbb{E}_{(x,y)}\left[L_{\mathrm{MH}}(h, r, x, y)\right]$ over all measurable functions. Recall from Section 2, the Bayes classifier is denoted by (h^*, r^*). The theorem 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 3 *Let (h^*_M, r^*_M) denote a pair attaining the infimum of the expected surrogate loss, $\mathbb{E}_{(x,y)}\left[L_{\mathrm{MH}}(h^*_M, r^*_M, x, y)\right] = \inf_{(h,r)\in meas} \mathbb{E}_{(x,y)}\left[L_{\mathrm{MH}}(h, r, x, y)\right]$. Then, for $\beta = \frac{1}{1-2c}$ and $\alpha = 1$,*

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

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

The theorem implies that the classification and rejection regions characterized by (h^*, r^*) and (h^*_M, r^*_M) are equal, that is minimizing the expected surrogate loss results in the same type of classification and rejection of the input space as minimizing the expected rejection loss. It also shows that the minimum value of the expected surrogate loss can be scaled by $3-2c$ to match the minimum value of the expected rejection loss.

3.3.2 Excess risk bound

Here, we show an upper bound on the excess risk of the rejection loss in terms of the excess risk of the surrogate loss. 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^*_M denote $\inf_{(h,r)} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)]$.

Theorem 4 *Let $R_M(h, r) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\mathrm{MH}}(h, r, x, y)]$ denote the expected surrogate loss of a pair (h, r). Then, the excess risk of (h, r) is upper bounded by its surrogate excess error as follows:*

$$R(h, r) - R^* \leq \frac{1}{(1 - c)(1 - 2c)} \left(R_M(h, r) - R^*_M\right).$$

The theorem implies that if we can minimize the excess risk based on the surrogate loss, then we are also minimizing the excess risk based on the rejection loss. Thus, in conjunction with the calibration results, these guarantees indicate that the surrogate loss L_{MH} admits several favorable properties that precisely match the rejection loss's behavior.

3.3.3 $(\mathcal{H}, \mathcal{R})$-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 [2013] 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 5 *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 $1_{y\neq\operatorname{sgn}(h(x))}1_{r(x)>0} + c1_{r(x)\leq 0} \leq 1_{yh(x)\leq 0}1_{r(x)>0} + c1_{r(x)\leq 0} \leq \Phi(-yh(x), -r(x))$ and since μh^* is in $\mathcal{H}$ for any $\mu > 0$, the following holds for any $\mu > 0$:

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

Now, $u \mapsto \Phi(-\mu yh^*(x), -r^*(x))$ is bounded for $yh^*(x) > 0$ and $r^*(x) > 0$; since we have that $\lim_{\mu\to+\infty} \Phi(-\mu yh^*(x), -r^*(x)) = 0$, by Lebesgue's dominated convergence theorem $\lim_{\mu\to+\infty} \mathbb{E}\left[\Phi(-\mu yh^*(x), -r^*(x))|r^*(x) > 0\right] = 0$. 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

$$L_{\text{MH}}^{\rho,\rho'}(h, r, x, y) =$$

$$\max\left(\max\left(1+\tfrac{\alpha}{2}\left(\tfrac{r(x)}{\rho'}-\tfrac{yh(x)}{\rho}\right),0\right),\max\left(c\left(1-\beta\tfrac{r(x)}{\rho'}\right),0\right)\right).$$

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 6 *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 $\mathcal{L}_{\text{MH},\mathcal{H},\mathcal{R}}$ be the family of functions defined by $\mathcal{L}_{\mathcal{H},\mathcal{R}} = \{(x,y) \mapsto \min\left(L_{\text{MH}}(h,r,x,y),1\right), h \in \mathcal{H}, r \in \mathcal{R}\}$. Since $\min(L_{\text{MH}},1)$ is bounded by one, by the general Rademacher complexity generalization bound [Koltchinskii and Panchenko, 2002], with probability at least $1-\delta$ over the draw of a sample S, the following holds:

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

Observe that we can express L_{MH} as follows:

$$\max\left(\max\left(1+\tfrac{\alpha}{2}(r(x)-yh(x)),0\right),\max\left(c\,(1-\beta r(x)),0\right)\right).$$

Therefore, since for any $a, b \in \mathbb{R}$, $\min\left(\max(a,b),1\right) = \max\left(\min(a,1),\min(b,1)\right)$, we can rewrite $\min(L_{\text{MH}},1)$ as:

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

Since it holds that $u \mapsto \min\left(\max\left(1+\frac{\alpha u}{2},0\right),1\right)$ is $\frac{\alpha}{2}$-Lipschitz and also that $u \mapsto \min\left(\max(c\,(1-\beta u),0),1\right)$ is $c\beta$-Lipschitz, by Talagrand's contraction lemma [Ledoux and Talagrand, 1991],

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

which completes the proof. □

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

Corollary 7 *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 $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

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

Then, via Koltchinskii and Panchenko [2002], the bound of Corollary 7 can be shown to hold uniformly for all $\rho, \rho' \in (0, 1)$, at the price of a term in $O\left(\sqrt{\frac{\log\log 1/\rho}{m}} + \sqrt{\frac{\log\log 1/\rho'}{m}}\right)$. This corollary will be used to derive our algorithms that learn the pair (h, r) from two kernel-based function classes.

3.5 Confidence-based rejection learning bounds

In this section, we present a learning guarantee for a natural class of confidence-based rejection algorithms based on a two-stage procedure. In the first stage, these algorithms learn a predictor h_0 by using a standard classification algorithm (i.e. SVMs, Adaboost etc.). Then, in the second stage, a threshold $\gamma \geq 0$ is chosen from a family of thresholds Γ such that it minimizes the confidence-based rejection loss $L(h, r_\gamma, x, y)$ where $r_\gamma(x) = |h(x)| - \gamma$. The theorem below bounds the generalization error of the pair (h_0, r_γ) for all possible $\gamma \in \Gamma$ where γ_{min} and γ_{max} denote the minimum and maximum γ values in the Γ set.

Theorem 8 *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\gamma \in \Gamma$*

$$R\left(h_0, r_\gamma\right) \leq \widehat{R}_S\left(h_0, r_\gamma\right) + \frac{\gamma_{\max} - \gamma_{\min}}{\sqrt{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{5}$$

where $r_\gamma = |h_0| - \gamma$.

Proof By standard generalization bounds for Rademacher complexity [Mohri et al., 2012], for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\gamma \in \Gamma$:

$$R\left(h_0, r_\gamma\right) \leq \widehat{R}_S\left(h_0, r_\gamma\right) + 2\mathfrak{R}_m\left(\Gamma\right) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

The empirical Rademacher complexity of Γ can be expressed as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\Gamma) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\gamma\in\Gamma}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\gamma\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\gamma\in\Gamma}\frac{\gamma}{m}\sum_{i=1}^{m}\sigma_i \,\middle|\, \sum_{i=1}^{m}\sigma_i \geq 0\right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\gamma\in\Gamma} -\frac{\gamma}{m}\sum_{i=1}^{m}\sigma_i \,\middle|\, \sum_{i=1}^{m}\sigma_i < 0\right] \\
&= \frac{\gamma_{\max} - \gamma_{\min}}{2m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^{m}\sigma_i\right|\right].
\end{aligned}$$

Thus, using Jensen's inequality, it can be upper-bounded as follows:

$$\widehat{\mathfrak{R}}_S(\Gamma) \leq \frac{\gamma_{\max} - \gamma_{\min}}{2m} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[|\sum_{i=1}^{m} \sigma_i|^2 \right]} = \frac{\gamma_{\max} - \gamma_{\min}}{2m} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_{i=1}^{m} \sigma_i^2 \right]} \leq \frac{\gamma_{\max} - \gamma_{\min}}{2\sqrt{m}}.$$

We then obtain the result of the theorem by taking the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$ and using it to bound $\mathfrak{R}_m(\Gamma)$ in the generalization bound above. □

The theorem justifies the second-stage of these algorithms whenever $\gamma_{\max} - \gamma_{\min}$ is bounded and shows that the generalization guarantee is independent of the number of threshold values used. In other words, numerous thresholds values can be used without affecting learnability, but note that as the number of threshold values increases it will become computationally expensive to find the best threshold.

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 7 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, the confidence-based models can be captured by an appropriate choice of $\boldsymbol{\Phi}'$. That is whenever K' is a second degree-kernel, the area where the rejection function $r \in \mathcal{R}$ abstains must be an ellipsoid. In this space of functions, there then exists a degenerate ellipsoid (i.e. two parallels lines) that corresponds to a rejection region characterized by the confidence-based rejection function, $r(x) = |h(x)| - \gamma$.

Corollary 9 *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 $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

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

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 [Mohri et al., 2012], 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 7 completes the proof. □

This learning bound guides directly the definition of our first algorithm based on the L_{MH} (see Appendix D for details) resulting in the following optimization:

$$\begin{aligned}
&\min_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{\xi}} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda'}{2}\|\boldsymbol{u}\|^2 + \sum_{i=1}^{m}\xi_i \\
&\text{subject to } \xi_i \geq c\left(1 - \beta\left(\boldsymbol{u}\cdot\boldsymbol{\Phi}'(x_i) + b'\right)\right), \\
&\qquad\qquad \xi_i \geq 1 + \frac{\alpha}{2}\left(\boldsymbol{u}\cdot\boldsymbol{\Phi}'(x_i) + b' - y_i\boldsymbol{w}\cdot\boldsymbol{\Phi}(x_i) - b\right), \xi_i \geq 0, i \in [1,m]
\end{aligned}$$

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 an algorithm based on the loss L_{PH} with the following primal optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{\xi},\boldsymbol{\xi}'} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda'}{2}\|\boldsymbol{u}\|^2 + \sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\xi'_i \\
&\text{subject to } \xi'_i \geq c(1 - \beta(\boldsymbol{u}\cdot\boldsymbol{\Phi}'(x_i) + b')), \\
&\qquad\qquad \xi_i \geq 1 + \frac{\alpha}{2}(\mathbf{u}\cdot\boldsymbol{\Phi}'(x_i) + b' - y_i(\boldsymbol{w}\cdot\boldsymbol{\Phi}(x_i) + b)), \xi_i \geq 0, \\
&\qquad\qquad \xi'_i \geq 0, i \in [1,m].
\end{aligned}$$

The dual formulations are given in Appendix D. 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 based on the confidence-based rejection model (Section 2.1). We thus consider a rejection function $r(x) = |h(x)| - \gamma$ that abstains on points classified with confidence less than a given threshold γ.

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

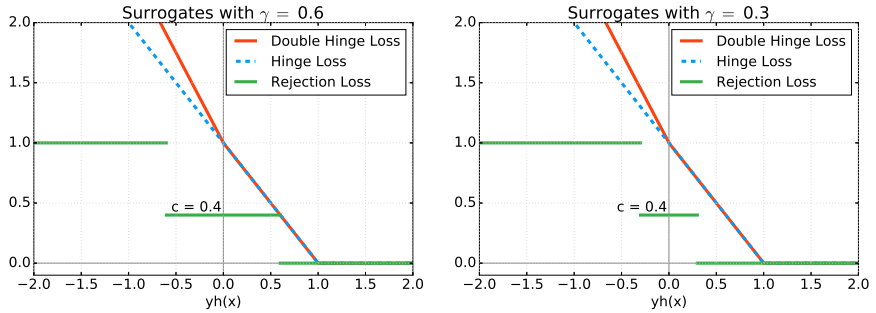

Fig. 4 Convex surrogates for confidence-based algorithms for $\gamma \in [0, 1 - c]$. The figures show the double hinge loss, hinge loss, and rejection loss for $\gamma = 0.6$ (left) and $\gamma = 0.3$ (right) for cost value $c = 0.4$. The hinge loss is clearly the tighter convex upper bound of the rejection loss.

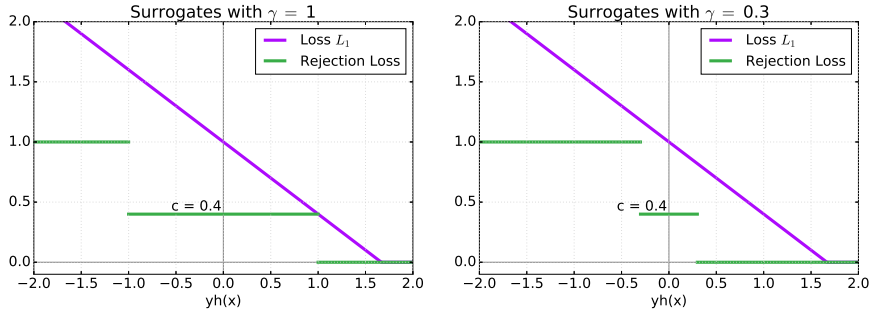

Fig. 5 Convex surrogate for confidence-based algorithms for $\gamma \in [0, 1]$. The left figure has threshold $\gamma = 1$ while the right figure has threshold $\gamma = 0.3$ for cost value $c = 0.4$.

the rejection loss only when $0 \leq \gamma \leq 1 - c$ (see Figure 4), 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. See Appendix E for the primal and dual formulations.

The aforementioned confidence based algorithms only apply for $\gamma \in [0, 1 - c]$. Here, we present an algorithm that upper bounds the rejection error for all values of $\gamma \in [0, 1]$ since a robust surrogate should majorate the rejection loss, L_γ, for all its possible values. This surrogate is a loss function passing through point $(1, c)$ given by

$$L_1 = \max(1 - y(1 - c)(\mathbf{w} \cdot \mathbf{\Phi}(x)), 0), \tag{6}$$

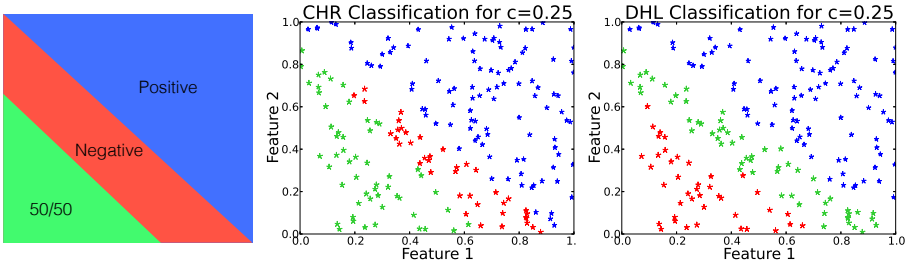

Fig. 6 The left figure illustrates the synthetic data, the middle figure shows how CHR classified the test points, and the right figure shows how DHL classified the same set of points with green denoting rejection.

see Figure 5 for the plot of this function. The main advantage of this loss is that it holds for all values of $\gamma \in [0, 1]$ giving a greater flexibility in choosing the width of the rejection region. Based on this loss, the optimization problem is a QCQP and for the primal and dual formulation, see Appendix E.

6 Experiments

In this section, we present the results of several experiments comparing our CHR algorithms with the confidence-based algorithms. All algorithms were implemented using CVX [CVX Research, 2012]. For the experiments below, the cost of the rejection ranged over $c \in \{0.05, 0.1, \ldots, 0.5\}$. 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 confidence-based algorithms ranged over $\gamma \in \{0.08, 0.16, \ldots, 0.96\}$. 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.

We first ran initial experiments testing the confidence-based algorithms. While the alternative algorithms we described in Section 5 are based on tighter surrogate losses for the rejection loss than that of DHL, empirical evidence suggests that DHL outperforms these alternatives (see Appendix E). Furthermore, we ran experiments comparing the two CHR algorithms which show that on average the CHR with L_{MH} performs slightly better than the CHR with L_{PH} (see Appendix F). Thus, in this section we report the results of several experiments comparing the best of each type of algorithm: CHR based on L_{MH} and the DHL algorithm.

We first tested DHL and CHR based on L_{MH} on a `synthetic` data which we generated by uniformly sampling one thousand points in the unit square such that the points above the $y = -x + 1$ are labeled positive, between $y = -x + 1$ and $y = -x + 0.5$ are labeled negative, and below $y = -x + 0.5$ are randomly assigned a label (see Figure 6). For this set of experiments, we used polynomial kernels of degree $d \in \{1, 2, 3\}$ for both algorithms. Figure 6 clearly shows how CHR finds the best possible rejection region, meaning the region where we randomly assigned a label, while DHL is unable attain this region since the

Table 1 Results for the `synthetic` data as a function of cost c. The last three columns are the degree of the polynomial kernels used by the classifier function for DHL and CHR and by the rejection function for CHR.

Cost	Rejection loss DHL	Rejection loss CHR	Classifier deg. DHL	Classifier deg. CHR	Rejector deg. CHR
0.05	0.154 ± 0.146	0.026 ± 0.002	1	3	2
0.1	0.103 ± 0.011	0.041 ± 0.011	1	2	3
0.15	0.113 ± 0.011	0.033 ± 0.004	1	2	3
0.2	0.130 ± 0.014	0.051 ± 0.016	1	2	3
0.25	0.145 ± 0.021	0.067 ± 0.011	1	2	3
0.3	0.188 ± 0.036	0.077 ± 0.013	1	2	3
0.35	0.248 ± 0.064	0.085 ± 0.017	3	2	3
0.4	0.323 ± 0.066	0.088 ± 0.013	3	2	3
0.45	0.382 ± 0.059	0.088 ± 0.014	3	2	1

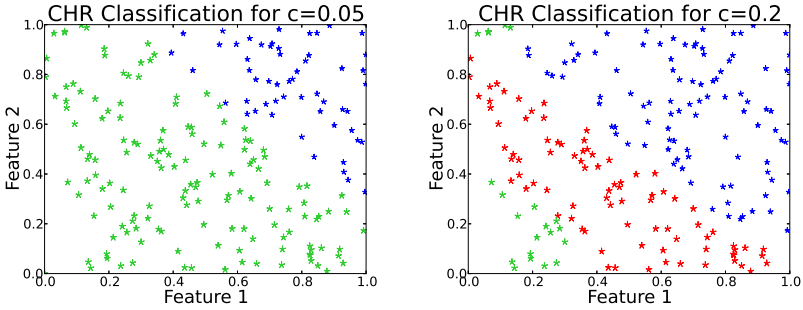

Fig. 7 CHR classification for rejection cost of $c = 0.05$ and $c = 0.2$.

points that it chooses to reject are restricted between its positive and negative labeled points.

For the parameters with the smallest rejection loss on the validation set, we provide the average rejection loss on the test set in Table 1, which shows that CHR substantially outperforms DHL. We also report the degree of the polynomial kernels used by each algorithm. Figure 7 shows the effects of cost c on CHR's classification surface. As we increase the cost c, less points are rejected as we would expect and moreover, the points that are rejected are chosen in the correct region, that is points below $y = -x + \frac{1}{2}$.

We then tested the algorithms on nine data sets from the UCI data repository, specifically `australian`, `cod`, `skin`, `liver`, `banknote`, `haberman`, `pima`, `monk`, and `transfusion`. Table 2 shows the sample size and number of features for each data set used in our experiments. For these experiments, we used polynomial kernels of degree $d \in \{1, 2, 3\}$ as well as Gaussian kernels with widths in the set $\{1, 10, 100\}$ for both algorithms.

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 3 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 rejection loss results of Table 3 show that the CHR algorithm yields a substantial improvement over the DHL algorithm. These findings are statistically significant at the 1% level

Table 2 Sample size and the number of features for each data set.

Data Sets	Sample Size	Feature
`synthetic`	1, 000	2
`australian`	690	14
`liver`	345	6
`cod`	369	8
`skin`	400	3
`banknote`	1, 372	4
`haberman`	306	3
`pima`	768	8
`monk`	124	6
`transfusion`	748	4

Table 3 Experimental results on the test set for the DHL algorithm and the CHR_{MH} algorithm for the fixed cost value $c = 0.25$.

Data	Rejection loss DHL	Rejection loss CHR_{MH}	Fraction rejected DHL	Fraction rejected CHR_{MH}	Non-rejected error DHL	Non-rejected error CHR_{MH}	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	$0.236 \pm .040$	$0.248 \pm .005$	$0.397 \pm .047$	$0.980 \pm .019$	$0.136 \pm .044$	$0.003 \pm .006$	$0.292 \pm .120$
monk	$0.326 \pm .061$	$0.242 \pm .016$	$0.184 \pm .134$	$0.776 \pm .300$	$0.280 \pm .063$	$0.048 \pm .072$	$0.144 \pm .101$
transfusion	$0.240 \pm .034$	$0.176 \pm .026$	$0.748 \pm .303$	$0.420 \pm .120$	$0.053 \pm .053$	$0.071 \pm .045$	$0.134 \pm .111$

or higher with one-sided paired t–test for all data sets except for the `liver` and `australian` data sets.

Table 3 also includes the fraction of rejected points and the standard classification error on non-rejected points. We can see from the table that the fraction of points that are rejected depends primarily on the dataset and on the algorithm. For almost all datasets, the DHL algorithm predicts at a higher rate the wrong label on non-rejected points compared to the CHR algorithm. In order to level the playing field for the two algorithms, the right most column of Table 3 was calculated as follows. First, we matched the fraction rejected of the DHL algorithm with fraction rejected of the CHR algorithm by varying the rejection threshold value, γ. Second, we recorded in the table the error on the remaining non-rejected points. This then shows that the CHR algorithm not only rejects the more difficult to classify sample points, but also obtains a significantly better error rate on the remaining points.

We now provide a series of figures that highlight different properties of the two algorithms for the UCI datasets. In Figure 8, we show the rejection loss as a function of the cost for six of our data sets. These plots demonstrate that the difference in accuracy between the two algorithms holds consistently for almost all values of c across all the data sets. Figure 9 shows that, as we would except, that the confidence-based algorithm only rejects on the boundary of the surface while the CHR algorithm does not restrict its rejection regions to only areas of low confidence for the `skin` dataset. Figure 10 shows the conditional probabilities over the test set of the CHR algorithm and the DHL algorithm

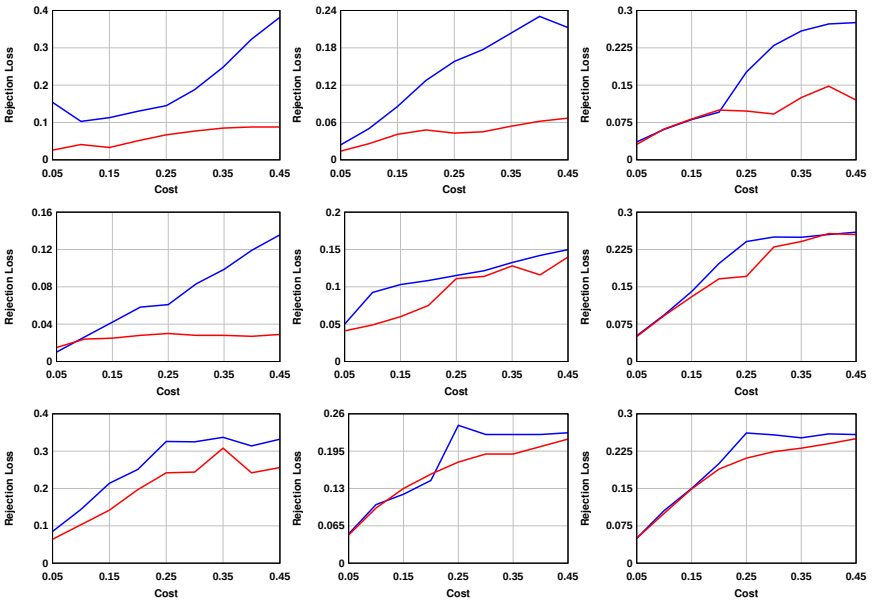

Fig. 8 Average rejection loss on the test set as a function of cost c for the DHL algorithm (blue line) and the CHR algorithm (red line) for six datasets. The first row shows the `synthetic`, `skin`, and `cod`, the middle row shows the `banknote`, `australian` and `pima`, and bottom row shows the `monk`, `transfusion`, and `haberman`.

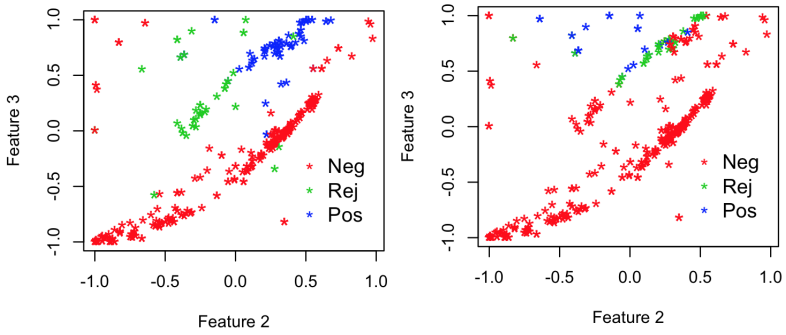

Fig. 9 The left figure shows CHR's classification of sample test points from the `skin` dataset with respect to different feature vectors. The right figure shows their classification by DHL and demonstrates how DHL rejects in areas of low confidence.

as a function of a feature vector for the `australian` data set. It indicates that DHL does not reject nearly as much as CHR, yet the sum of the positive and rejected points for the two algorithms is the almost the same (grey line).

In the appendix, we provide a series of tables that further analyze the difference between the two algorithms. Tables F2 report the average rejection loss with standard deviations on the test set of each algorithm for different cost values across the data sets. Note that the rejection loss equals the associated

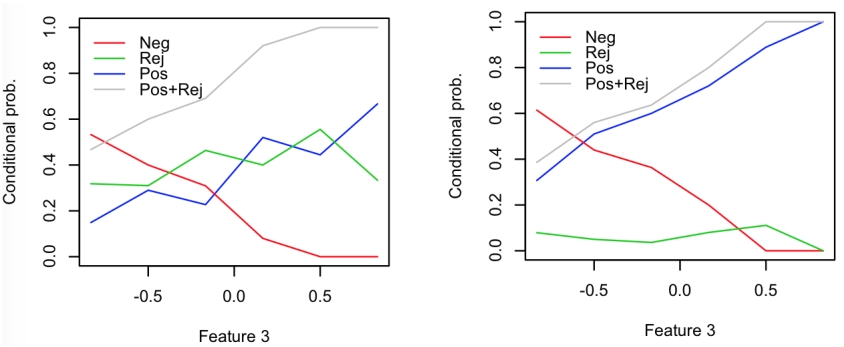

Fig. 10 The figures show the conditional probabilities of the CHR algorithm (left) and DHL algorithm (right) as a function of a feature vector for the `australian` test set.

cost when the algorithm rejects all the points (i.e. see DHL algorithm on `australian` with $c = 0.05$). Tables F3 show the average fraction of the test points rejected by each algorithm for different values of c across the data sets. As the cost of rejection c increases, less points are rejected by both algorithms. Tables F4 provide the classification error on the non-rejected points for all the algorithms. For most values of c, the CHR algorithm classifies correctly the non-rejected points at a higher rate than the DHL algorithm.

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.

Declarations

- Funding: None
- Competing Interests: None.

Appendix A Consistency of convex surrogates

In this appendix, we derive two theorems about the consistency of the convex surrogate, L_{MH}. The first theorem shows that the convex surrogate is calibrated with respect to the Bayes solution and the second theorem upper bounds the excess risk of the rejection loss by the excess risk of the surrogate loss. For both theorems, we will analyze the expected surrogate loss, which can be written in terms of $\eta(x)$:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L_{\text{MH}}(h,r,x,y)] = \mathop{\mathbb{E}}_{x}\left[\eta(x)\phi(-h(x),r(x)) + (1-\eta(x))\phi(h(x),r(x))\right], \tag{A1}$$

where $\phi(-h(x),r(x)) = \max\left(1+\frac{1}{2}\left(r(x)-h(x)\right), c\left(1-\frac{1}{1-2c}r(x)\right), 0\right)$. For simplicity, we also define

$$\mathcal{L}_\phi(\eta(x),h(x),r(x)) = \eta(x)\phi(-h(x),r(x)) + (1-\eta(x))\phi(h(x),r(x)). \tag{A2}$$

The idea behind the proof of the first theorem below is to find the minimizer of $(u,v)\mapsto\mathcal{L}_\phi(\eta(x),u,v)$ for any fixed x in order to then re-write it in terms of the infimum of the expected rejection loss.

Theorem 3. *Let (h^*_M, r^*_M) denote a pair attaining the infimum of the expected surrogate loss, $\mathbb{E}_{(x,y)}[L_{\text{MH}}(h^*_M, r^*_M, x, y)] = \inf_{(h,r)\in meas}\mathbb{E}_{(x,y)}[L_{\text{MH}}(h,r,x,y)]$. 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:*

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

Proof Since the infimum of the expected surrogate loss is over all measurable functions, to determine $(h^*_{\text{M}}, r^*_{\text{M}})$ it suffices to find, 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

$$\begin{aligned}
&\mathcal{L}_\phi(\eta(x),3-2c,1-2c) = (3-2c)(1-\eta(x))\\
&\mathcal{L}_\phi(\eta(x),-3+2c,1-2c) = (3-2c)\eta(x)\\
&\mathcal{L}_\phi(\eta(x),0,(2c-2)(1-2c)) = (3-2c)c.
\end{aligned}$$

Thus, we can conclude that the minimum of $\mathcal{L}_\phi(\eta(x),u,v)$ is attained at $(3-2c)[\eta(x)1_{\eta(x)<c} + c1_{c\leq\eta(x)\leq 1-c} + (1-\eta(x))1_{\eta(x)>1-c}]$, which completes the proof. Below, for completeness, we show how to solve three of these LPs

where $\mathcal{L}_\phi(\eta(x), h, r) = 0$, $\mathcal{L}_\phi(\eta(x), h, r) = c\left(1 - \frac{1}{1-2c}r\right)$, and $\mathcal{L}_\phi(\eta(x), h, r) = \eta(x)\left(1 + \frac{1}{2}(r-h)\right) + (1-\eta(x))\left(1 + \frac{1}{2}(r+h)\right)$.

1. For $\mathcal{L}_\phi(\eta(x), h, r) = 0$, we have the following optimization problem

$$\begin{aligned}&\min_{(h,r)} 0\\ &\text{subject to: } c\left(1 - \frac{1}{1-2c}r\right) \leq 0, 1 + \frac{1}{2}(r-h) \leq 0, 1 + \frac{1}{2}(r+h) \leq 0\end{aligned}$$

   Now the constraint $c\left(1 - \frac{1}{1-2c}r\right) \leq 0$ implies that $r \geq (1-2c)c > 0$. If we sum the remaining constraints $1 + \frac{1}{2}(r-h) \leq 0, 1 + \frac{1}{2}(r+h) \leq 0$, they imply that $r \leq -2$. Thus, this LP is not feasible.
2. For $\mathcal{L}_\phi(\eta(x), h, r) = c(1 - \frac{1}{1-2c}r)$, we have the following optimization problem

$$\begin{aligned}&\min_{(h,r)} c\left(1 - \frac{1}{1-2c}r\right)\\ &\text{subject to: } c\left(1 - \frac{1}{1-2c}r\right) \geq 0, 1 + \frac{1}{2}(r-h) \leq c\left(1 - \frac{1}{1-2c}r\right),\\ &\qquad 1 + \frac{1}{2}(r+h) \leq c\left(1 - \frac{1}{1-2c}r\right)\end{aligned}$$

   Summing the last two constraints and solving for r, we have that $r \leq 2(c-1)(1-2c) \leq 0$. Since this optimization problem, we want to maximize r, we can conclude that $r^*_M = 2(c-1)(1-2c)$ and that $h^*_M = 0$.
3. For $\mathcal{L}_\phi(\eta(x), h, r) = \eta(x)(1 + \frac{1}{2}(r-h)) + (1-\eta(x))(1 + \frac{1}{2}(r+h))$, we have the following problem

$$\begin{aligned}&\min_{(h,r)} \eta(x)\left(1 + \frac{1}{2}(r-h)\right) + (1-\eta(x))\left(1 + \frac{1}{2}(r+h)\right)\\ &\text{subject to: } 1 + \frac{1}{2}(r-h) \geq 0, 1 + \frac{1}{2}(r-h) \geq c\left(1 - \frac{1}{1-2c}r\right),\\ &\qquad 1 + \frac{1}{2}(r+h) \geq 0, 1 + \frac{1}{2}(r+h) \geq c\left(1 - \frac{1}{1-2c}r\right),\end{aligned}$$

   By simplifying the constraints, we can see that the feasibility region of the optimization problem has to be between the lines $2+r \geq h$ and $h \geq -(2+r)$ and between $2(1-c)+\frac{1}{1-2c}r \geq h$ and $h \geq -\left(2(1-c) + \frac{1}{1-2c}r\right)$. Notice that $-(2+r) = r+2$ at $r = -2$ and that $-\left(2(1-c) + \frac{1}{1-2c}r\right) = 2(1-c)+\frac{1}{1-2c}r$ at $r = 2(c-1)(1-2c)$. Since $-2 \leq 2(c-1)(1-2c)$ for $0 < c < 0.5$, we have that -2 is not in the feasibility region of the optimization problem. Thus one of the optimality points is at $r = 2(c-1)(1-2c)$ and $h = 0$.

We also have that $2 + r = 2(1 - c) + \frac{1}{1-2c}r$ at the point $r = 1 - 2c$ and $h = 3 - 2c$. Similarly, $-(2 + r) = -\left(2(1 - c) + \frac{1}{1-2c}r\right)$ at the point $r = 1 - 2c$ and $h = -(3 - 2c)$. Thus all the optimality points are in the set $(h, r) \in \{(0, 2(c - 1)(1 - 2c)), (3 - 2c, 1 - 2c), (-(3 - 2c), 1 - 2c)\}$

□

We now provide a proof of the excess risk bound of our convex surrogate. It consists of analyzing three cases when $\mathcal{L}^*(\eta(x)) = c$, $\mathcal{L}^*(\eta(x)) = \eta(x)$, and $\mathcal{L}^*(\eta(x)) = 1 - \eta(x)$ and then using the calibration results of the previous theorem.

Theorem 4. *Let $R_M(h, r) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\text{MH}}(h, r, x, y)]$ denote the expected surrogate loss of a pair (h, r). Then, the excess risk of (h, r) is upper bounded by its surrogate excess error as follows:*

$$R(h, r) - R^* \leq \frac{1}{(1 - c)(1 - 2c)} \left(R_M(h, r) - R_M^*\right).$$

Proof Conditioning on the label y and using the fact that the infimum is over all measurable functions, we can switch the infimum and expectation as follows:

$$\begin{aligned} R(h, r) - R(h^*, r^*) = \mathop{\mathbb{E}}_{x} \big[&(\eta(x) - \mathcal{L}^*(\eta(x)))1_{\text{sgn}(h)\neq 1, r>0} \\ &+ (1 - \eta(x) - \mathcal{L}^*(\eta(x)))1_{\text{sgn}(h)\neq -1, r>0} + (c - \mathcal{L}^*(\eta(x)))1_{r\leq 0}\big] \end{aligned} \tag{A3}$$

where $\mathcal{L}^*(\eta(x)) = \eta(x)1_{\eta(x)<c} + c1_{c\leq\eta(x)\leq 1-c} + (1 - \eta(x))1_{\eta(x)>1-c}$. We can thus focus on minimizing the components inside the expectation for a fixed x. From the calibration theorem, we have that $\mathcal{L}_\phi^*(\eta(x)) = (3 - 2c)\mathcal{L}^*(\eta(x))$. Since $\mathcal{L}^*(\eta(x))$ admits three values, we can consider the following three cases: $\mathcal{L}^*(\eta(x)) = c$, $\mathcal{L}^*(\eta(x)) = \eta(x)$, and $\mathcal{L}^*(\eta(x)) = 1 - \eta(x)$. Below, we describe one such case, but the remaining can be analyzed by a similar reasoning. When $c \leq \eta(x) \leq 1 - c$, we have that $\mathcal{L}^*(\eta(x)) = c$ and so $r^* \leq 0$. Since by the calibration theorem, $\text{sign}(r^*) = \text{sign}(r_M^*)$, we have that $r_M^* \leq 0$ as well as $\mathcal{L}_\phi{}^*(\eta(x)) = (3 - 2c)c$. Under this case, the Equation A3 can be written as $R(h, r) - R(h^*, r^*) = \mathbb{E}_x\left((\eta(x) - c)1_{\text{sgn}(h)\neq 1, r>0} + (1 - \eta(x) - c)1_{\text{sgn}(h)\neq -1, r>0}\right)$. Note that these indicator functions on the right hand side are mutually exclusive, thus we can just show that each component is bounded. Since for the value of $\eta(x)$ and c that satisfy $c \leq \eta(x) \leq 1 - c$, we have $(\eta(x) - c)1_{\text{sgn}(h)\neq 1, r>0} \leq (\eta(x) - c)1_{h<0, r>0}$ and $(1 - \eta(x) - c)1_{\text{sgn}(h)\neq -1, r>0} \leq (1 - \eta(x) - c)1_{h\geq 0, r>0}$. Thus, for the first component, we want to show that

$$(\eta(x) - c)1_{h<0, r>0} \leq \frac{1}{(1 - c)(1 - 2c)} \left(\mathcal{L}_\phi(\eta(x), h, r) - (3 - 2c)c\right) 1_{h<0, r>0}$$

and for the second component, we want to show that

$$(1 - \eta(x) - c)1_{h\geq 0, r>0} \leq \frac{1}{(1 - c)(1 - 2c)} \left(\mathcal{L}_\phi(\eta(x), h, r) - (3 - 2c)c\right) 1_{h\geq 0, r>0}.$$

We will prove that the bound holds for each component if there exists a constant $\kappa > 0$ such that inequality $1 - 2c \leq \kappa\big(1 - (3 - 2c)c\big)$ holds. Since $\frac{1-2c}{1-(3-2c)c} = \frac{1}{1-c}$,

we can easily conclude that $\kappa := \frac{1}{1-c}$. Now since $\frac{1}{(1-c)(1-2c)} \geq \frac{1}{1-c}$, we have the inequality of the theorem under this case.

Focusing on the second component, we proceed by first finding the minimum of the $\mathcal{L}_\phi(\eta(x), h, r)1_{h\geq 0, r>0}$ and then show that the inequality is satisfied. The optimality points of minimizing $\mathcal{L}_\phi(\eta(x), h, r)1_{h\geq 0, r>0}$ are $(h, r) \in \{(3-2c, 1-2c), (0,0), (2(1-c), 0), (0, 1-2c)\}$. Evaluating $\mathcal{L}_\phi$ at these optimal points, we have that $\mathcal{L}_\phi(\eta(x), 3-2c, 1-2c) = (3-2c)(1-\eta(x))$, $\mathcal{L}_\phi(\eta(x), 0, 1-2c) = \frac{3}{2} - c$, $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$, and $\mathcal{L}_\phi(\eta(x), 2(1-c), 0) = 1 + (1-2\eta(x))(1-c)$. Now since $(3-2c)(1-\eta(x)) \geq 1 + (1-2\eta(x))(1-c)$ for $c \leq \eta(x) \leq 1-c$ and since $\frac{3}{2} - c > 1$ for $c < \frac{1}{2}$, we can exclude $(2(1-c), 0)$ and $(0, 1-2c)$. Thus, depending on the sign of $\eta(x)$, the minimum is attained at $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$ or at $\mathcal{L}_\phi(\eta(x), 0, 1-2c) = 1 + (1-2\eta(x))(1-c)$. For $\mathcal{L}_\phi(\eta(x), 0, 1-2c) = 1 + (1-2\eta(x))(1-c)$, the inequality $1-\eta(x)-c \leq 1+(1-2\eta(x))(1-c)-(3-2c)c$ holds for all $c \leq \eta(x) \leq 1-c$. While for $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$, since $c \leq \eta(x)$, we have that $1-\eta(x)-c \leq 1-2c \leq \kappa\big(1-(3-2c)c\big)$ holds.

Now for the first component, we again proceed by first finding the minimum of the $\mathcal{L}_\phi(\eta(x), h, r)1_{h<0, r>0}$ and then by showing the inequality is satisfied. By similar reasoning as the calibration theorem, we have that the optimality points are $(h, r) \in \{(-(3-2c), 1-2c), (0,0), (-2(1-2c), 0), (0, 1-2c)\}$. Evaluating $\mathcal{L}_\phi$ at these points, we have that $\mathcal{L}_\phi(\eta(x), -(3-2c), 1-2c) = (3-2c)\eta(x)$, $\mathcal{L}_\phi(\eta(x), 0, 1-2c) = \frac{3}{2} - c$, $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$, and $\mathcal{L}_\phi(\eta(x), -2(1-c), 0) = \eta(x)(2-c)+(1-\eta(x))c$. By the similar reasoning as above, we can again exclude the points $(-(3-2c), 1-2c)$ and $(0, 1-2c)$. Depending on the sign of $\eta(x)$, the minimum is attained at $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$ or at $\mathcal{L}_\phi(\eta(x), -2(1-c), 0) = \eta(x)(2-c)+(1-\eta(x))c$. For all $c \leq \eta(x) \leq 1-c$, we have that the inequality $\eta(x) - c \leq \eta(x)(2-c) + (1-\eta(x))c$ holds. Now for $\mathcal{L}_\phi(\eta(x), 0, 0) = 1$, we have that $\eta(x) - c \leq 1 - 2c \leq \kappa\big(1-(3-2c)c\big)$. □

Appendix B Alternative convex surrogate functions

Alternative convex surrogate functions can be found using a concave lower bound formula described here. Let $u \mapsto \Phi(u)$ and $u \mapsto \Psi(u)$ be strictly increasing concave functions lower bounding $1_{u>0}$. Then, the following inequalities hold:

$$
\begin{aligned}
&L(h, r, x, y) \\
&\leq 1_{yh(x)\leq 0} 1_{r(x)>0} + c\, 1_{r(x)\leq 0} \\
&= \big(1 - 1_{yh(x)>0}\big)\, 1_{r(x)>0} + c\, 1_{r(x)\leq 0} \\
&= 1_{r(x)>0} - 1_{yh(x)>0} 1_{r(x)>0} + c\, 1_{r(x)\leq 0} \\
&= \big(1 - 1_{r(x)\leq 0}\big) - 1_{yh(x)>0} 1_{r(x)>0} + c\, 1_{r(x)\leq 0} \\
&= 1 - (1-c) 1_{r(x)\leq 0} - 1_{yh(x)>0} 1_{r(x)>0} \\
&= 1 - (1-c) 1_{r(x)\leq 0} - 1_{\min(yh(x), r(x))>0} \\
&\leq 1 - (1-c)\Phi(-r(x)) - \Psi(\min(yh(x), r(x))) \\
&= 1 - (1-c)\Phi(-r(x)) - \min(\Psi(yh(x)), \Psi(r(x))).
\end{aligned} \tag{B4}
$$

The last term of right - hand side of (B4) defines a convex function of h and r since the minimum of two concave functions is concave.

Appendix C Connections to cost-sensitive learning

In this section, we draw connections between the cost-sensitive learning framework and learning with rejection.

The standard cost-sensitive algorithms and theory are designed for unknown distributions; however, in our setting, there is some prior information about the distribution since the rejection label has measure zero, a fact that should be exploited to derive a finer analysis. Moreover, using cost-sensitive algorithms for the rejection setting might not produce any interesting solution since they would treat rejection as any other label and since it is unclear how they would perform with a label for which there is no training data [Beygelzimer et al., 2008, 2005, Lin, 2014]. To elaborate on this, we first introduce a natural model for multi-class classification with rejection which can be viewed as an instance of cost-sensitive models and discuss its properties. The hypothesis set commonly adopted in multi-class classification is that of scoring functions: a *scoring function* $h(\cdot, y)\colon \mathcal{X} \to \mathbb{R}$ is learned for each class $y \in \mathcal{Y}$ and the class predicted for $x \in \mathcal{X}$ is the one with the highest score, that is $\operatorname{argmax}_{y\in\mathcal{Y}} h(x, y)$. This is also the hypothesis set adopted in the more complex multi-class classification scenario of structured prediction where misclassification is cost-sensitive: the loss $L(y, y')$ of predicting $y' \in \mathcal{Y}$ instead of the correct class $y \in \mathcal{Y}$ depends on the pair (y, y').

This suggests a natural model for multi-class classification with rejection. As in the standard multi-class case, we can introduce a scoring function for rejection $r(x) = h(x, \mathtt{r})$, where $\mathtt{r}$ is the rejection symbol. The *label* predicted, which is either a regular class label or the label $\mathtt{r}$ with the semantics of rejection, is the one with the highest score:

$$\mathsf{h}(x) = \operatorname*{argmax}_{y\in\mathcal{Y}\cup\{\mathtt{r}\}} h(x, y).$$

Thus, the rejection function r is implicitly defined by $r(x) = \max_{y\in\mathcal{Y}} h(x, y) - h(x, \mathtt{r})$ and the rejection loss can be expressed by

$$L(h, r, x, y) = 1_{h(x,y)\leq \max_{y'\neq y} h(x,y')} 1_{h(x,\mathtt{r})<\max_{\mathtt{y}\in\mathcal{Y}} \mathtt{h(x,y)}} + c1_{h(x,\mathtt{r})\geq\max_{\mathtt{y}\in\mathcal{Y}} \mathtt{h(x,y)}}.$$

This loss can be upper bounded by the convex surrogate

$$L_{\mathrm{SH}}(h, r, x, y) = \max\Big(0, 1-[h(x, y)-\max_{y'\neq y} h(x, y')], c\big(1-[h(x, y)-h(x, \mathtt{r})]\big)\Big),$$

which is closely related to the loss function used in StructSVM. Using L_{SH} and linear functions $h(x, y) = \boldsymbol{w}_y \cdot \boldsymbol{\Phi}(x)$ for each class $y \in \mathcal{Y} \cup \{\mathtt{r}\}$ with a norm-2 regularization leads to an algorithm defined by the following optimization problem

$$\begin{aligned} \min_{\mathbf{W}, \boldsymbol{w}_r, \boldsymbol{\xi}} \quad & \frac{\lambda}{2} \sum_{l=1}^{k} {\boldsymbol{w}_l}^2 + \frac{\lambda'}{2} {\boldsymbol{w}_r}^2 + \sum_{i=1}^{m} \xi_i \\ \text{subject to: } & \xi_i \geq c(1 - \boldsymbol{w}_{y_i} \cdot \Phi(x_i) + \boldsymbol{w}_r \cdot \Phi(x_i)), \\ & \xi_i \geq 1 - \boldsymbol{w}_{y_i} \cdot \Phi(x_i) + \boldsymbol{w}_l \cdot \Phi(x_i), \\ & \xi_i \geq 0, i \in [1, m], \forall l \in \mathcal{Y} - \{y_i\}, \end{aligned}$$

where $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$.

In principle, one can use the theory and learning bounds from structured prediction to derive the optimization problem above, but in the absence of rejection labels in the data, there is no incentive for the rejection scoring function to be large. More precisely, suppose that the dataset has only positive features so that $\Phi(x_i)$ has only positive elements. Now, considering the constraints of the optimization problem, $\boldsymbol{w}_r$ appears only in $\xi \geq c(1 - \boldsymbol{w}_{y_i} \cdot \Phi(x_i) + \boldsymbol{w}_r \cdot \Phi(x_i))$ and as a consequence of $\Phi(x_i)$ being positive, these constraints will push $\boldsymbol{w}_r$ to be negative. Combining this with the fact that the objective is to minimize ${\boldsymbol{w}_r}^2$, the optimization problem will find a solution such that w_r is small and negative. One may also see this directly by looking at the KKT conditions for $\boldsymbol{w}_r$ of the optimization problem. Thus, for positive $\Phi(x_i)$ the score for the rejection label will be a small negative number while scores of the other class-labels could be positive. This implies that this method is likely not abstain very often. Thus, while very natural, this cost-sensitive formulation does not lead to a useful algorithm in this scenario. One may seek to modify the objective function to promote larger values for the scoring functions but our attempts typically led to non-convex functions and the absence of an $\mathtt{r}$ label in the training sample remained a problem.

There are existing cost-sensitive algorithms that can be used in the rejection setting [Beygelzimer et al., 2008, 2005, Lin, 2014], which are based on reductions stemming from the work of Langford and Beygelzimer [2005]. However, their guarantees are based on relating the difference of the generalization error and the Bayes optimal error of the cost sensitive problem to that of reduced binary problem by paying a multiplication factor that usually depends on the quality of the reduction, which results in a quantity that is not easy to compare to. Furthermore, as argued by Tu and Lin [2010], these algorithms can be quite complicated both in terms of their encoding structure and their algorithmic procedure since they reduce the cost-sensitive problem first to a weighted binary classification, that is then converted into a binary classification problem via the Costing algorithm of Zadrozny et al. [2003], and which in turn is solved by a standard algorithm for binary classification. Note that the convex surrogate loss approach described in the previous paragraph is closer

in nature to the cost-sensitive work of Tu and Lin [2010], but their algorithm does not apply to the rejection setting.

The analysis of calibration in Ramaswamy and Agarwal [2016] is not helpful for the analysis of learning with rejections since the main point of their paper is consistency guarantees and analysing a notion they introduce, *convex calibration dimension of the loss matrix*, which chracterizes when it is possible to design a convex surrogate that is calibrated. Instead, we would need guarantees and an analysis for the convex surrogate, L_{MH}, and our main concern is not consistency. Additionally, the size of our loss matrix as defined in Ramaswamy and Agarwal [2016] is small, thus the analysis of the dimensionality is not relevant and in fact their bound for the rejection loss is not tight.

Appendix D Algorithms with kernel-based hypotheses

In this section, we provide further details related to the algorithms with kernel-based hypothesis.

D.1 Optimization problems

We derive the optimization problem first for loss L_{MH} and then for loss L_{PH}. We find that both the primal and dual optimization problems for L_{MH} and L_{PH} are QPs.

Firstly, by the generalization of the Corollary 9 to a uniform bound over $\rho, \rho' \in (0,1)$ and by picking $\Lambda = 1$ and $\Lambda' = 1$, we have that, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $\rho, \rho' \in (0,1)$, $\mathcal{H} = \{\boldsymbol{x} \to \boldsymbol{w} \cdot \Phi(x) \colon \|\boldsymbol{w}\| \leq 1\}$ and $\mathcal{R} = \{\boldsymbol{x} \to \boldsymbol{u} \cdot \Phi'(x) \colon \|\boldsymbol{u}\| \leq 1\}$:

$$R(h,r) \leq \frac{1}{m}\sum_{i=1}^{m} \max\left(1 + \tfrac{\alpha}{2}\left(\tfrac{\boldsymbol{u}\cdot\Phi'(x_i)}{\rho'} - \tfrac{y\boldsymbol{w}\cdot\Phi(x_i)}{\rho}\right), c\left(1 - \tfrac{\beta\boldsymbol{u}\cdot\Phi'(x_i)}{\rho'}\right), 0\right)$$
$$+ \alpha\sqrt{\tfrac{(\kappa/\rho)^2}{m}} + (2\beta c + \alpha)\sqrt{\tfrac{(\kappa'/\rho')^2}{m}} + C(\rho, \rho', m, \delta),$$

where $C(\rho, \rho', m, \delta) = \sqrt{\frac{\log\frac{1}{\delta}}{2m}} + \sqrt{\frac{\log\log 1/\rho}{m}} + \sqrt{\frac{\log\log 1/\rho'}{m}}$. Secondly, under binary classification, the functions h/ρ and r/ρ admit the same generalization error as h and r for any $\rho \in (0,1)$ and $\rho' \in (0,1)$. Thus, with probability at least $1-\delta$, the following holds for all $\rho \in (0,1)$, $\rho' \in (0,1)$, $h \in \mathcal{H} = \{\boldsymbol{x} \to \boldsymbol{w} \cdot \Phi(x) : \|\boldsymbol{w}\| \leq \frac{1}{\rho}\}$ and $r \in \mathcal{R} = \{\boldsymbol{x} \to \boldsymbol{u} \cdot \Phi'(x) : \|\boldsymbol{u}\| \leq \frac{1}{\rho'}\}$

$$R(h,r) \leq \frac{1}{m}\sum_{i=1}^{m} \max\left(1 + \frac{\alpha}{2}\big(\boldsymbol{u}\cdot\Phi'(x_i) - y\boldsymbol{w}\cdot\Phi(x_i)\big), c\big(1 - \beta\boldsymbol{u}\cdot\Phi'(x_i)\big), 0\right)$$

$$+ \alpha\sqrt{\frac{(\kappa/\rho)^2}{m}} + (2\beta c + \alpha)\sqrt{\frac{(\kappa'/\rho')^2}{m}} + C(\rho, \rho', m, \delta).$$

For any $\rho \in (0,1)$ and $\rho' \in (0,1)$, the sum term on the right hand side depends on $\boldsymbol{w}$ and $\boldsymbol{u}$ and so the bound leads to the following optimization problem

$$\min_{\substack{\|\boldsymbol{w}\|^2 \leq \frac{1}{\rho^2} \\ \|\boldsymbol{u}\|^2 \leq \frac{1}{\rho'^2}}} \frac{1}{m}\sum_{i=1}^{m} \max\Big(1 + \frac{\alpha}{2}\big(\boldsymbol{u}\cdot\Phi'(x_i) - y\boldsymbol{w}\cdot\Phi(x_i)\big), c\left(1 - \beta\boldsymbol{u}\cdot\Phi'(x_i)\right), 0\Big).$$

Lastly, we introduce slack variables ξ_i for $i \in [1,m]$ along with Lagrange multipliers $\lambda \geq 0$ and $\lambda' \geq 0$ so that the primal optimization problem for L_{MH} is as follows:

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

where we explicitly mark both the offset b of classifier $h(x)$ and offset b' of rejection function $r(x)$. Since $K(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$ and $K'(x_i,x_j) = \Phi'(x_i)\cdot\Phi'(x_j)$, the dual optimization problem is given by the following:

$$\begin{aligned}
\max_{\boldsymbol{\eta},\boldsymbol{\zeta}} \quad & \lambda\lambda'\sum_{i=1}^{m}\eta_i + \lambda\lambda' c\sum_{i=1}^{m}\zeta_i - \frac{\alpha^2\lambda'}{8}\sum_{i,j=1}^{m}\eta_i\eta_j y_i y_j K(x_i,x_j) \\
& - \frac{\lambda}{2}\sum_{i,j=1}^{m}\Big(\frac{\alpha\eta_i}{2} - c\beta\zeta_i\Big)\Big(\frac{\alpha\eta_j}{2} - c\beta\zeta_j\Big)K'(x_i,x_j) \\
\text{subject to} \quad & \sum_{i=1}^{m}\eta_i y_i = 0, \sum_{i=1}^{m}\Big(\frac{\alpha\eta_i}{2} - c\beta\zeta_i\Big) = 0, \\
& \eta_i \geq 0, \zeta_i \geq 0, \eta_i + \zeta_i \leq 1, i \in [1,m].
\end{aligned}$$

By a similar reasoning as above, we derive the optimization problem for the surrogate loss L_{PH}. By introducing slack variables ξ_i for $i \in [1,m]$ as well as Lagrange multipliers $\lambda \geq 0$ and $\lambda' \geq 0$, we have the following primal optimization problem for L_{PH}:

$$\begin{aligned}
\min_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{\xi},\boldsymbol{\xi}'} \quad & \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda'}{2}\|\boldsymbol{u}\|^2 + \sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\xi'_i \\
\text{subject to} \quad & \xi'_i \geq c(1 - \beta(\boldsymbol{u}\cdot\Phi'(x_i) + b')),
\end{aligned}$$

$$\xi_i \geq 1 + \frac{\alpha}{2}(\mathbf{u} \cdot \Phi'(x_i) + b' - y_i(\boldsymbol{w} \cdot \Phi(x_i) + b)),$$
$$\xi_i \geq 0, \xi'_i \geq 0, i \in [1, m].$$

Since $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ and $K'(x_i, x_j) = \Phi'(x_i) \cdot \Phi'(x_j)$, the dual optimization problem of L_{PH} is given by the following:

$$\begin{aligned}\max_{\boldsymbol{\eta},\boldsymbol{\zeta}} \ & \lambda\lambda' \sum_{i=1}^{m} \eta_i + \lambda\lambda' c \sum_{i=1}^{m} \zeta_i - \frac{\alpha^2\lambda'}{8} \sum_{i,j=1}^{m} \eta_i\eta_j y_i y_j K(x_i, x_j) \\ & - \frac{\lambda}{2} \sum_{i,j=1}^{m} \Big(\frac{\alpha\eta_i}{2} - c\beta\zeta_i\Big)\Big(\frac{\alpha\eta_j}{2} - c\beta\zeta_j\Big) K'(x_i, x_j) \\ \text{subject to } & \sum_{i=1}^{m} \eta_i y_i = 0, \sum_{i=1}^{m} \Big(\frac{\alpha\eta_i}{2} - c\beta\zeta_i\Big) = 0, \\ & 0 \leq \eta_i \leq 1, 0 \leq \zeta_i \leq 1, i \in [1, m].\end{aligned}$$

Appendix E Confidence-based rejection algorithms

In this section, we present the optimization problems studied in Section 5 and then report experimental results that compares these different confidence-based rejection algorithms.

E.1 Optimization problems

We first consider the algorithms that hold for $\gamma \in [0, 1-c]$. The DHL algorithm of Bartlett and Wegkamp [2008] solves the following optimization problem

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\beta}} \ & \sum_{i=1}^{m} \xi_i + \frac{1-2c}{c}\beta_i \\ \text{subject to } & \sum_{i=1,j=1}^{m} \alpha_i\alpha_j K(x_i, x_j) \leq (1-c)^2, \xi_i \geq 1 - y\sum_{i=1}^{m} \alpha_i K(x_i, x) \wedge \xi_i \geq 0, \\ & \beta_i \geq -y\sum_{i=1}^{m} \alpha_i K(x_i, x) \wedge \beta_i \geq 0, i \in [1, m].\end{aligned}$$

The optimization problem based on the hinge loss is given by

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\xi}} \ & \sum_{i=1}^{m} \xi_i \\ \text{subject to } & \sum_{i=1,j=1}^{m} \alpha_i\alpha_j K(x_i, x_j) \leq (1-c)^2,\end{aligned}$$

$$\xi_i \geq 1 - y \sum_{i=1}^{m} \alpha_i K(x_i, x) \wedge \xi_i \geq 0, i \in [1, m],$$

and its dual formulation is as follows

$$\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{\eta}, \zeta} \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} \eta_i - \zeta r^2 \\ &\text{subject to } 0 \leq \zeta, 0 \leq \alpha_i, 0 \leq \eta_i, 0 \leq \alpha_i + \eta_i \leq 1, i \in [1, m] \\ &\sum_{i=1}^{m} (\alpha_i + a\eta_i)(\alpha_j + a\eta_j) y_i y_j K(x_i, x_j) = (\zeta r)^2 \end{aligned}$$

The above shows that the optimization problem solved by DHL is QCQP while the optimization problem based on the hinge loss is a QP.

We now show the optimization problem based on the loss L_1 that holds for all $\gamma \in [0, 1]$:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\xi}} \sum_{i=1}^{m} \xi_i \\ &\text{subject to } \sum_{i=1, j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \leq 1, \\ &\xi_i \geq 0, \xi_i \geq 1 - y(1 - c) \sum_{i=1}^{m} \alpha_i K(x_i, x), i \in [1, m]. \end{aligned}$$

Its dual formulation is as follows

$$\begin{aligned} &\max_{\boldsymbol{\alpha}, \zeta} \sum_{i} \alpha_i - \zeta (1 - c)^2 \\ &\text{subject to } \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) = \zeta^2 (1 - c)^2, 1 \geq \alpha \geq 0, \zeta \geq 0, i \in [1, m] \end{aligned}$$

where we note that this optimization problem is a QCQP.

E.2 Empirical comparison of confidence-based rejection algorithms

We tested the confidence-based algorithms on four data sets from the UCI repository: `australian`, `cod`, `skin`, and `liver`. Table E1 shows the average rejection loss along with the standard deviations for the Hinge loss, L_1 loss, and the DHL confidence-based algorithms across the four data sets for the nine cost values c. These results show that the DHL algorithm outperforms the Hinge and L_1 algorithms across four data sets for most values of c. They were obtained using standard 5-fold cross-validation: For each data set, we split

Table E1 Average rejection loss along with the standard deviations for confidence-based algorithms described in Section 5 across the four data sets for the nine cost values c.

Cost	australian Hinge Loss	australian L_1 Loss	australian DHL	liver Hinge	liver L_1 Loss	liver DHL
0.05	0.147 ± 0.042	0.147 ± 0.042	0.112 ± 0.033	0.338 ± 0.146	0.313 ± 0.138	0.061 ± 0.014
0.1	0.149 ± 0.044	0.149 ± 0.044	0.120 ± 0.024	0.350 ± 0.119	0.262 ± 0.129	0.111 ± 0.024
0.15	0.150 ± 0.047	0.150 ± 0.047	0.128 ± 0.025	0.373 ± 0.087	0.297 ± 0.119	0.166 ± 0.022
0.2	0.148 ± 0.049	0.148 ± 0.049	0.130 ± 0.036	0.381 ± 0.066	0.289 ± 0.056	0.206 ± 0.017
0.25	0.148 ± 0.047	0.148 ± 0.047	0.134 ± 0.038	0.414 ± 0.063	0.346 ± 0.051	0.236 ± 0.040
0.3	0.151 ± 0.049	0.151 ± 0.049	0.137 ± 0.038	0.401 ± 0.035	0.374 ± 0.061	0.263 ± 0.042
0.35	0.149 ± 0.046	0.149 ± 0.046	0.141 ± 0.039	0.431 ± 0.022	0.402 ± 0.031	0.273 ± 0.041
0.4	0.148 ± 0.046	0.148 ± 0.046	0.148 ± 0.042	0.431 ± 0.022	0.431 ± 0.022	0.337 ± 0.048
0.45	0.150 ± 0.046	0.150 ± 0.046	0.150 ± 0.046	0.426 ± 0.026	0.426 ± 0.026	0.430 ± 0.017

Cost	cod Hinge Loss	cod L_1 Loss	cod DHL	skin Hinge	skin L_1 Loss	skin DHL
0.05	0.192 ± 0.177	0.154 ± 0.084	0.044 ± 0.034	0.137 ± 0.077	0.083 ± 0.047	0.024 ± 0.016
0.1	0.191 ± 0.071	0.175 ± 0.085	0.077 ± 0.028	0.115 ± 0.044	0.110 ± 0.066	0.061 ± 0.031
0.15	0.235 ± 0.032	0.214 ± 0.064	0.123 ± 0.030	0.153 ± 0.063	0.111 ± 0.033	0.091 ± 0.031
0.2	0.268 ± 0.072	0.223 ± 0.046	0.175 ± 0.031	0.133 ± 0.045	0.168 ± 0.051	0.128 ± 0.036
0.25	0.263 ± 0.026	0.273 ± 0.050	0.204 ± 0.026	0.196 ± 0.053	0.221 ± 0.078	0.158 ± 0.041
0.3	0.273 ± 0.026	0.285 ± 0.035	0.230 ± 0.022	0.179 ± 0.058	0.209 ± 0.048	0.177 ± 0.044
0.35	0.274 ± 0.025	0.274 ± 0.025	0.259 ± 0.029	0.209 ± 0.151	0.213 ± 0.068	0.204 ± 0.056
0.4	0.276 ± 0.025	0.276 ± 0.025	0.273 ± 0.026	0.213 ± 0.068	0.195 ± 0.035	0.231 ± 0.067
0.45	0.276 ± 0.025	0.276 ± 0.025	0.276 ± 0.025	0.214 ± 0.068	0.202 ± 0.091	0.215 ± 0.066

the data randomly into training, test, and validation test in the ration 3:1:1. We allowed the threshold γ to vary in $\{0.1, 0.2, \ldots, 0.9\}$ and the cost c values range in $c \in \{0.05, 0.1 \ldots, 0.45\}$. All the kernels are polynomial degree kernels where the degree d is in $\{1, 2, 3\}$ and unlike the experiments done in Section 6, we did not use Gaussian kernels for this initial set of experiments. For a fixed cost value c, we find the combination of parameters (γ, d) with smallest average rejection loss on the validation set and report the average rejection loss for these parameters on the test set. Overall, these results show that DHL outperforms the other algorithms on three out of four datasets for most of the values of c. While the other algorithms are seemingly plausible alternatives, we find through these preliminary results that the DHL is the superior algorithm under this confidence based setting.

Appendix F Experiments comparing DHL and CHR algorithms

In the following pages, we provide the results of the several experiments described in Section 6. As a short summary, the CHR algorithm achieves a better performance across all data sets for most values of cost c. Tables F2 shows the average rejection loss with standard deviations on the test set. The CHR_{MH} stands for the CHR algorithm based on L_{MH}. Tables F3 reports the average fraction of the test points rejected. Tables F4 provide the classification error on the non-rejected points.

F.1 Experiments comparing CHR algorithms

In this section, we show the results of some initial experiments comparing the two CHR Algorithms. Let CHR_{PH} stand for the CHR algorithm based on L_{PH}. The experimental set-up is exactly the same as in Section 6 except that we just used polynomial kernels of degree $d \in \{1, 2, 3\}$. Table F5 shows the average rejection loss with standard deviations on the test set for both algorithms. We

find that on average the $\mathrm{CHR}_{\mathrm{MH}}$ performs slightly better than the $\mathrm{CHR}_{\mathrm{PH}}$ as is expected since the loss L_{PH} is an upper bound of the loss L_{MH}.

References

Bartlett, P., Wegkamp, M.: Classification with a reject option using a hinge loss. JMLR (2008)

Beygelzimer, A., Langford, J., Ravikumar, P.: Error correcting tournaments. In: Arxiv (2008)

Beygelzimer, A., Dani, V., Hayes, T., Langford, J., Zadrozny, B.: Error limiting reductions between classification tasks. In: ICML (2005)

Bounsiar, A., Grall, E., Beauseroy, P.: Kernel based rejection method for supervised classification. In: WASET (2007)

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., Sugiyama, M.: Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In: Advances in Neural Information Processing Systems (2022)

Capitaine, H.L., Frelicot., C.: An optimum class-rejective decision rule and its evaluation. In: ICPR (2010)

Chaudhuri, K., Zhang, C.: Beyond disagreement-based agnostic active learning. In: NIPS (2014)

Chow, C.K.: An optimum character recognition system using decision function. IEEE T. C. (1957)

Chow, C.K.: On optimum recognition error and reject trade-off. IEEE T. C. (1970)

CVX Research, I.: CVX: Matlab Software for Disciplined Convex Programming, version 2.0 (2012)

DeSalvo, G., Mohri, M., Syed, U.: Learning with deep cascades. In: ALT (2015)

Dubuisson, B., Masson, M.: Statistical decision rule with incomplete knowledge about classes. In: Pattern Recognition (1993)

El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. JMLR (2010)

El-Yaniv, R., Wiener, Y.: Agnostic selective classification. In: NIPS (2011)

Elkan, C.: The foundations of cost-sensitive learning. In: IJCAI (2001)

Freund, Y., Mansour, Y., Schapire, R.: Generalization bounds for averaged classifiers. Annals of Statistics (2004)

Fumera, G., Roli, F.: Support vector machines with embedded reject option. In: ICPR (2002)

Fumera, G., Roli, F., Giacinto, G.: Multiple reject thresholds for improving classification reliability. In: ICAPR (2000)

Grandvalet, Y., Keshet, J., Rakotomamonjy, A., Canu, S.: Suppport vector machines with a reject option. In: NIPS (2008)

Hamid, K., Asif, A., Abbasi, W., Sabih, D., Minhas, F.: Machine learning with abstention for automated liver disease diagnosis. In: FIT (2017)

Herbei, R., Wegkamp, M.: Classification with reject option. Can. J. Stat. (2005)

Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics (2002)

Landgrebe, T., Tax, D., Paclik, P., Duin, R.: Interaction between classification and reject performance for distance-based reject-option classifiers. PRL (2005)

Langford, J., Beygelzimer, A.: Sensitive error correcting output codes. In: COLT (2005)

Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes. Springer, New York (1991)

Lin, H.-T.: Reduction from cost-sensitive multiclass classification to one-versus-one binary classification. In: JMLR (2014)

Littman, M., Li, L., Walsh, T.: Knows what it knows: A framework for self-aware learning. In: ICML (2008)

Long, P.M., Servedio, R.A.: Consistency versus realizable H-consistency for multiclass classification. In: ICML (2013)

Mao, A., Mohri, M., Zhong, Y.: Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. CoRR **to appear** (2023a)

Mao, A., Mohri, M., Zhong, Y.: Theoretically grounded loss functions and algorithms for score-based multi-class abstention. CoRR **to appear** (2023b)

Mao, A., Mohri, C., Mohri, M., Zhong, Y.: Two-stage learning to defer with multiple experts. In: NeurIPS (2023c)

Melvin, I., Weston, J., Leslie, C.S., Noble, W.S.: Combining classifiers for improved classification of proteins from sequence or structure. BMCB (2008)

Mohri, C., Andor, D., Choi, E., Collins, M.: Learning to reject with a fixed predictor: Application to decontextualization. CoRR **abs/2301.09044** (2023)

Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Boston (2012)

Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: ICML (2020a)

Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: International Conference on Machine Learning, pp. 7076–7087 (2020b)

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., Sontag, D.: Who should predict? exact algorithms for learning to defer to humans. In: International Conference on Artificial Intelligence and Statistics, pp. 10520–10545 (2023)

Narasimhan, H., Menon, A.K., Jitkrittum, W., Kumar, S.: Learning to reject meets ood detection: Are all abstentions created equal? arXiv preprint arXiv:2301.12386 (2023)

Ni, C., Charoenphakdee, N., Honda, J., Sugiyama, M.: On possibility and impossibility of multiclass classification with rejection. (2019)

Pereira, C.S., Pires, A.: On optimal reject rules and ROC curves. PRL (2005)

Pietraszek, T.: Optimizing abstaining classifiers using ROC. In: ICML (2005)

Ramaswamy, H., Agarwal, S.: Convex calibration dimension for multiclass loss matrices. In: JMLR (2016)

Tax, D., Duin, R.: Growing a multi-class classifier with a reject option. In: Pattern Recognition Letters (2008)

Tortorella, F.: An optimal reject rule for binary classifiers. In: ICAPR (2001)

Trapeznikov, K., Saligrama, V.: Supervised sequential classification under budget constraints. In: AISTATS (2013)

Tu, H.-H., Lin, H.-T.: One-sided support vector regression for multiclass cost-sensitive classification. In: ICML (2010)

Verma, R., Nalisnic, E.: Calibrated learning to defer with one-vs-all classifiers.

In: ICML (2022)

Wang, J., Trapeznikov, K., Saligrama, V.: An LP for sequential learning under budgets. In: JMLR (2014)

Yuan, M., Wegkamp, M.: Classification methods with reject option based on convex risk minimizations. In: JMLR (2010)

Yuan, M., Wegkamp, M.: SVMs with a reject option. In: Bernoulli (2011)

Zadrozny, B., Langford, J., Abe, N.: Cost sensitive learning by cost- proportionate example weighting. In: ICDM (2003)

Zhang, C., Chaudhuri, K.: The extended Littlestone's dimension for learning with mistakes and abstentions. In: COLT (2016)

Table F2 Average rejection loss along with the standard deviations on the test set for the DHL algorithm and the CHR_{MH} algorithm across different costs.

Cost	`australian` CHR_{MH}	`australian` DHL	`liver` CHR_{MH}	`liver` DHL
0.05	0.041 ± 0.023	0.05 ± 0	0.055 ± 0.012	0.05 ± 0
0.1	0.049 ± 0.046	0.092 ± 0.02	0.100 ± 0.001	0.108 ± 0.019
0.15	0.060 ± 0.082	0.103 ± 0.022	0.151 ± 0.006	0.152 ± 0.018
0.2	0.075 ± 0.071	0.108 ± 0.024	0.197 ± 0.003	0.206 ± 0.017
0.25	0.111 ± 0.021	0.115 ± 0.026	0.248 ± 0.005	0.236 ± 0.04
0.3	0.114 ± 0.107	0.122 ± 0.025	0.299 ± 0.011	0.263 ± 0.042
0.35	0.128 ± 0.019	0.133 ± 0.023	0.322 ± 0.024	0.273 ± 0.041
0.4	0.116 ± 0.073	0.142 ± 0.023	0.343 ± 0.068	0.35 ± 0.039
0.45	0.140 ± 0.028	0.15 ± 0.046	0.343 ± 0.070	0.399 ± 0.031

Cost	`cod` CHR_{MH}	`cod` DHL	`skin` CHR_{MH}	`skin` DHL
0.05	0.031 ± 0.009	0.036 ± 0.007	0.014 ± 0.004	0.024 ± 0.016
0.1	0.062 ± 0.017	0.061 ± 0.01	0.026 ± 0.008	0.05 ± 0.006
0.15	0.082 ± 0.023	0.081 ± 0.018	0.041 ± 0.016	0.086 ± 0.025
0.2	0.100 ± 0.024	0.096 ± 0.024	0.048 ± 0.021	0.128 ± 0.036
0.25	0.098 ± 0.037	0.176 ± 0.03	0.043 ± 0.020	0.158 ± 0.041
0.3	0.092 ± 0.029	0.23 ± 0.022	0.045 ± 0.019	0.177 ± 0.044
0.35	0.125 ± 0.049	0.259 ± 0.029	0.054 ± 0.017	0.204 ± 0.056
0.4	0.148 ± 0.043	0.273 ± 0.026	0.062 ± 0.024	0.231 ± 0.067
0.45	0.120 ± 0.021	0.276 ± 0.025	0.067 ± 0.024	0.213 ± 0.068

Cost	`bank` CHR_{MH}	`bank` DHL	`haber` CHR_{MH}	`haber` DHL
0.05	0.015 ± 0.003	0.01 ± 0.006	0.050 ± 0.000	0.05 ± 0
0.1	0.024 ± 0.006	0.026 ± 0.012	0.100 ± 0.000	0.106 ± 0.008
0.15	0.025 ± 0.006	0.042 ± 0.02	0.149 ± 0.002	0.15 ± 0
0.2	0.028 ± 0.006	0.058 ± 0.02	0.189 ± 0.015	0.2 ± 0
0.25	0.030 ± 0.006	0.061 ± 0.022	0.211 ± 0.037	0.261 ± 0.033
0.3	0.028 ± 0.008	0.083 ± 0.025	0.224 ± 0.022	0.258 ± 0.025
0.35	0.028 ± 0.010	0.099 ± 0.028	0.231 ± 0.029	0.252 ± 0.015
0.4	0.027 ± 0.007	0.119 ± 0.028	0.240 ± 0.033	0.26 ± 0.021
0.45	0.029 ± 0.008	0.136 ± 0.027	0.250 ± 0.031	0.258 ± 0.022

Cost	`pima` CHR_{MH}	`pima` DHL
0.05	0.050 ± 0.000	0.052 ± 0.003
0.1	0.092 ± 0.006	0.094 ± 0.006
0.15	0.130 ± 0.009	0.14 ± 0.017
0.2	0.166 ± 0.011	0.197 ± 0.022
0.25	0.171 ± 0.017	0.241 ± 0.025
0.3	0.230 ± 0.011	0.25 ± 0.023
0.35	0.241 ± 0.024	0.25 ± 0.027
0.4	0.257 ± 0.024	0.255 ± 0.028
0.45	0.255 ± 0.030	0.26 ± 0.034

Table F2 Average rejection loss along with the standard deviations on the test set for the DHL algorithm and the CHR_{MH} algorithm across different costs.

Cost	`monk` CHR_{MH}	`monk` DHL	`transfusion` CHR_{MH}	`transfusion` DHL
0.05	0.064 ± 0.034	0.085 ± 0.079	0.049 ± 0.003	0.051 ± 0.002
0.1	0.103 ± 0.007	0.144 ± 0.083	0.096 ± 0.007	0.102 ± 0.015
0.15	0.142 ± 0.023	0.214 ± 0.131	0.130 ± 0.024	0.12 ± 0.03
0.2	0.197 ± 0.023	0.251 ± 0.098	0.155 ± 0.020	0.144 ± 0.041
0.25	0.242 ± 0.016	0.326 ± 0.061	0.176 ± 0.026	0.24 ± 0.034
0.3	0.244 ± 0.338	0.325 ± 0.085	0.190 ± 0.038	0.224 ± 0.05
0.35	0.308 ± 0.052	0.337 ± 0.056	0.190 ± 0.037	0.224 ± 0.05
0.4	0.242 ± 0.335	0.314 ± 0.107	0.203 ± 0.047	0.224 ± 0.05
0.45	0.256 ± 0.352	0.332 ± 0.105	0.216 ± 0.034	0.227 ± 0.047

Cost	`synthetic` CHR_{MH}	`synthetic` DHL
0.05	0.026 ± 0.002	0.154 ± 0.146
0.1	0.041 ± 0.011	0.103 ± 0.011
0.15	0.033 ± 0.004	0.113 ± 0.011
0.2	0.051 ± 0.016	0.13 ± 0.014
0.25	0.067 ± 0.011	0.145 ± 0.021
0.3	0.077 ± 0.013	0.188 ± 0.036
0.35	0.085 ± 0.017	0.248 ± 0.064
0.4	0.088 ± 0.013	0.323 ± 0.066
0.45	0.088 ± 0.014	0.382 ± 0.059

Table F3 Average fraction of points rejected along with the standard deviations for the DHL algorithm and the CHR_{MH} algorithm across different costs.

	`australian`	`australian`	`liver`	`liver`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.799 ± 0.446	1 ± 0	0.977 ± 0.022	1 ± 0
0.1	0.245 ± 0.284	0.258 ± 0.092	0.997 ± 0.006	0.968 ± 0.071
0.15	0.400 ± 0.548	0.165 ± 0.044	0.986 ± 0.014	0.919 ± 0.024
0.2	0.013 ± 0.016	0.143 ± 0.02	0.986 ± 0.014	0.887 ± 0.066
0.25	0.172 ± 0.024	0.136 ± 0.008	0.980 ± 0.019	0.397 ± 0.047
0.3	0.299 ± 0.275	0.106 ± 0.019	0.919 ± 0.030	0.142 ± 0.059
0.35	0.168 ± 0.029	0.097 ± 0.008	0.565 ± 0.310	0.142 ± 0.024
0.4	0.004 ± 0.010	0.072 ± 0.017	0.574 ± 0.511	0.128 ± 0.047
0.45	0.035 ± 0.078	0.004 ± 0.006	0.046 ± 0.012	0.029 ± 0.031

	`cod`	`cod`	`skin`	`skin`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.511 ± 0.163	0.665 ± 0.054	0.280 ± 0.081	0.18 ± 0.044
0.1	0.570 ± 0.171	0.584 ± 0.054	0.260 ± 0.084	0.503 ± 0.062
0.15	0.527 ± 0.131	0.503 ± 0.063	0.260 ± 0.080	0.405 ± 0.074
0.2	0.297 ± 0.082	0.357 ± 0.049	0.102 ± 0.053	0.1 ± 0.025
0.25	0.024 ± 0.028	0.186 ± 0.055	0.052 ± 0.027	0.093 ± 0.033
0.3	0.038 ± 0.035	0.027 ± 0.021	0.050 ± 0.029	0.09 ± 0.051
0.35	0.124 ± 0.060	0.014 ± 0.01	0.062 ± 0.060	0.075 ± 0.051
0.4	0.154 ± 0.055	0 ± 0	0.067 ± 0.034	0.033 ± 0.011
0.45	0.003 ± 0.006	0 ± 0	0.065 ± 0.039	0 ± 0

	`bank`	`bank`	`haber`	`haber`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.272 ± 0.031	0.071 ± 0.025	1.000 ± 0.000	1 ± 0
0.1	0.183 ± 0.035	0.094 ± 0.03	1.000 ± 0.000	0.99 ± 0.015
0.15	0.131 ± 0.041	0.106 ± 0.022	0.993 ± 0.015	1 ± 0
0.2	0.036 ± 0.015	0.105 ± 0.016	0.846 ± 0.211	1 ± 0
0.25	0.036 ± 0.022	0.066 ± 0.016	0.439 ± 0.148	0.875 ± 0.132
0.3	0.070 ± 0.024	0.058 ± 0.017	0.275 ± 0.120	0.007 ± 0.015
0.35	0.066 ± 0.031	0.053 ± 0.012	0.249 ± 0.088	0.092 ± 0.1
0.4	0.062 ± 0.022	0.048 ± 0.005	0.207 ± 0.158	0.01 ± 0.009
0.45	0.061 ± 0.022	0.052 ± 0.012	0.141 ± 0.081	0.013 ± 0.014

	`pima`	`pima`
Cost	CHR_{MH}	DHL
0.05	1.000 ± 0.000	0.982 ± 0.012
0.1	0.762 ± 0.054	0.766 ± 0.04
0.15	0.709 ± 0.065	0.564 ± 0.105
0.2	0.700 ± 0.055	0.355 ± 0.03
0.25	0.643 ± 0.068	0.055 ± 0.007
0.3	0.469 ± 0.144	0.045 ± 0.02
0.35	0.049 ± 0.034	0.038 ± 0.024
0.4	0.056 ± 0.060	0.034 ± 0.014
0.45	0.010 ± 0.013	0.026 ± 0.009

Table F3 Average fraction of points rejected along with the standard deviations for the DHL algorithm and the CHR_{MH} algorithm across different costs.

Cost	`monk` CHR_{MH}	`monk` DHL	`transfusion` CHR_{MH}	`transfusion` DHL
0.05	0.960 ± 0.049	0.576 ± 0.108	0.917 ± 0.121	0.989 ± 0.014
0.1	0.952 ± 0.107	0.48 ± 0.075	0.943 ± 0.079	0.863 ± 0.174
0.15	0.896 ± 0.149	0.36 ± 0.117	0.753 ± 0.147	0.428 ± 0.04
0.2	0.944 ± 0.088	0.896 ± 0.211	0.597 ± 0.176	0.285 ± 0.027
0.25	0.776 ± 0.300	0.184 ± 0.134	0.420 ± 0.120	0.748 ± 0.303
0.3	0.040 ± 0.089	0.176 ± 0.1	0.381 ± 0.127	0 ± 0
0.35	0.240 ± 0.376	0.208 ± 0.077	0.177 ± 0.093	0 ± 0
0.4	0.024 ± 0.054	0.064 ± 0.022	0.168 ± 0.136	0 ± 0
0.45	0.072 ± 0.161	0.08 ± 0.049	0.127 ± 0.069	0.011 ± 0.024

Cost	`synthetic` CHR_{MH}	`synthetic` DHL
0.05	0.495 ± 0.039	0.888 ± 0.153
0.1	0.370 ± 0.078	0.26 ± 0.03
0.15	0.164 ± 0.017	0.237 ± 0.044
0.2	0.117 ± 0.027	0.238 ± 0.043
0.25	0.084 ± 0.027	0.228 ± 0.049
0.3	0.060 ± 0.020	0.208 ± 0.037
0.35	0.023 ± 0.025	0.1 ± 0.018
0.4	0.004 ± 0.009	0.077 ± 0.008
0.45	0.000 ± 0.000	0.065 ± 0.015

Table F4 Average classification error on non-rejected points along with the standard deviations for the DHL algorithm and the CHR_{MH} algorithm across different costs.

	`australian`	`australian`	`liver`	`liver`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.001 ± 0.003	0 ± 0	0.006 ± 0.013	0 ± 0
0.1	0.025 ± 0.026	0.067 ± 0.013	0.000 ± 0.000	0.012 ± 0.026
0.15	0.000 ± 0.000	0.078 ± 0.022	0.003 ± 0.006	0.014 ± 0.02
0.2	0.072 ± 0.069	0.08 ± 0.023	0.000 ± 0.000	0.029 ± 0.029
0.25	0.068 ± 0.023	0.081 ± 0.025	0.003 ± 0.006	0.136 ± 0.044
0.3	0.025 ± 0.027	0.09 ± 0.022	0.023 ± 0.017	0.22 ± 0.047
0.35	0.070 ± 0.020	0.099 ± 0.024	0.125 ± 0.113	0.223 ± 0.042
0.4	0.114 ± 0.071	0.113 ± 0.027	0.113 ± 0.144	0.299 ± 0.055
0.45	0.125 ± 0.043	0.148 ± 0.044	0.322 ± 0.065	0.386 ± 0.042

	`cod`	`cod`	`skin`	`skin`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.005 ± 0.007	0.003 ± 0.006	0.000 ± 0.000	0.015 ± 0.016
0.1	0.005 ± 0.007	0.003 ± 0.006	0.000 ± 0.000	0 ± 0
0.15	0.003 ± 0.006	0.005 ± 0.012	0.003 ± 0.006	0.025 ± 0.031
0.2	0.041 ± 0.038	0.024 ± 0.024	0.028 ± 0.027	0.108 ± 0.034
0.25	0.092 ± 0.039	0.13 ± 0.043	0.030 ± 0.024	0.135 ± 0.037
0.3	0.081 ± 0.029	0.222 ± 0.023	0.030 ± 0.024	0.15 ± 0.035
0.35	0.081 ± 0.053	0.254 ± 0.028	0.033 ± 0.023	0.178 ± 0.045
0.4	0.086 ± 0.039	0.273 ± 0.026	0.035 ± 0.024	0.218 ± 0.063
0.45	0.119 ± 0.022	0.276 ± 0.025	0.037 ± 0.020	0.213 ± 0.068

	`bank`	`bank`	`haber`	`haber`
Cost	CHR_{MH}	DHL	CHR_{MH}	DHL
0.05	0.001 ± 0.003	0.007 ± 0.005	0.000 ± 0.000	0 ± 0
0.1	0.006 ± 0.006	0.017 ± 0.01	0.000 ± 0.000	0.007 ± 0.009
0.15	0.006 ± 0.002	0.026 ± 0.018	0.000 ± 0.000	0 ± 0
0.2	0.021 ± 0.004	0.037 ± 0.017	0.020 ± 0.027	0 ± 0
0.25	0.021 ± 0.008	0.045 ± 0.018	0.102 ± 0.048	0.043 ± 0.027
0.3	0.007 ± 0.003	0.066 ± 0.021	0.141 ± 0.039	0.256 ± 0.022
0.35	0.005 ± 0.003	0.08 ± 0.028	0.144 ± 0.032	0.22 ± 0.034
0.4	0.002 ± 0.003	0.1 ± 0.028	0.157 ± 0.068	0.256 ± 0.022
0.45	0.001 ± 0.002	0.112 ± 0.024	0.187 ± 0.043	0.252 ± 0.025

	`pima`	`pima`
Cost	CHR_{MH}	DHL
0.05	0.000 ± 0.000	0.003 ± 0.004
0.1	0.016 ± 0.004	0.017 ± 0.007
0.15	0.023 ± 0.010	0.056 ± 0.031
0.2	0.026 ± 0.011	0.126 ± 0.026
0.25	0.043 ± 0.023	0.227 ± 0.025
0.3	0.090 ± 0.044	0.236 ± 0.021
0.35	0.223 ± 0.025	0.236 ± 0.024
0.4	0.235 ± 0.034	0.242 ± 0.029
0.45	0.251 ± 0.033	0.248 ± 0.036

Table F4 Average classification error on non-rejected points along with the standard deviations for the DHL algorithm and the CHR_{MH} algorithm across different costs.

Cost	`monk` CHR_{MH}	`monk` DHL	`transfusion` CHR_{MH}	`transfusion` DHL
0.05	0.016 ± 0.036	0.056 ± 0.083	0.003 ± 0.006	0.001 ± 0.003
0.1	0.008 ± 0.018	0.096 ± 0.088	0.001 ± 0.003	0.016 ± 0.032
0.15	0.008 ± 0.018	0.16 ± 0.147	0.017 ± 0.028	0.056 ± 0.035
0.2	0.008 ± 0.018	0.072 ± 0.14	0.036 ± 0.036	0.087 ± 0.046
0.25	0.048 ± 0.072	0.28 ± 0.063	0.071 ± 0.045	0.053 ± 0.053
0.3	0.232 ± 0.325	0.272 ± 0.095	0.076 ± 0.036	0.224 ± 0.05
0.35	0.224 ± 0.134	0.264 ± 0.061	0.128 ± 0.037	0.224 ± 0.05
0.4	0.232 ± 0.325	0.288 ± 0.115	0.136 ± 0.071	0.224 ± 0.05
0.45	0.224 ± 0.318	0.296 ± 0.092	0.159 ± 0.041	0.223 ± 0.052

Cost	`synthetic` CHR_{MH}	`synthetic` DHL
0.05	0.001 ± 0.002	0.11 ± 0.153
0.1	0.004 ± 0.005	0.077 ± 0.013
0.15	0.008 ± 0.004	0.077 ± 0.013
0.2	0.028 ± 0.016	0.082 ± 0.014
0.25	0.046 ± 0.011	0.088 ± 0.02
0.3	0.059 ± 0.011	0.126 ± 0.035
0.35	0.077 ± 0.023	0.213 ± 0.066
0.4	0.086 ± 0.014	0.292 ± 0.064
0.45	0.088 ± 0.014	0.353 ± 0.053

Table F5 Average rejection loss along with the standard deviations on the test set the $\mathrm{CHR}_{\mathrm{MH}}$ algorithm and the $\mathrm{CHR}_{\mathrm{PH}}$ algorithm across the seven data sets for the nine cost values c using polynomial kernels.

Cost	australian $\mathrm{CHR}_{\mathrm{PH}}$	australian $\mathrm{CHR}_{\mathrm{MH}}$	liver $\mathrm{CHR}_{\mathrm{PH}}$	liver $\mathrm{CHR}_{\mathrm{MH}}$
0.05	0.055 ± 0.010	0.011 ± 0.025	0.062 ± 0.010	0.054 ± 0.008
0.1	0.080 ± 0.016	0.046 ± 0.063	0.109 ± 0.011	0.109 ± 0.011
0.15	0.092 ± 0.019	0.032 ± 0.071	0.160 ± 0.011	0.151 ± 0.008
0.2	0.101 ± 0.019	0.070 ± 0.096	0.208 ± 0.011	0.202 ± 0.012
0.25	0.117 ± 0.026	0.088 ± 0.120	0.249 ± 0.013	0.260 ± 0.016
0.3	0.124 ± 0.022	0.083 ± 0.076	0.293 ± 0.019	0.301 ± 0.028
0.35	0.137 ± 0.024	0.135 ± 0.024	0.336 ± 0.036	0.333 ± 0.015
0.4	0.137 ± 0.034	0.129 ± 0.015	0.365 ± 0.069	0.359 ± 0.015
0.45	0.149 ± 0.037	0.082 ± 0.182	0.372 ± 0.079	0.401 ± 0.052

Cost	cod $\mathrm{CHR}_{\mathrm{PH}}$	cod $\mathrm{CHR}_{\mathrm{MH}}$	skin $\mathrm{CHR}_{\mathrm{PH}}$	skin $\mathrm{CHR}_{\mathrm{MH}}$
0.05	0.043 ± 0.005	0.021 ± 0.048	0.013 ± 0.004	0.014 ± 0.004
0.1	0.083 ± 0.021	0.037 ± 0.083	0.026 ± 0.008	0.026 ± 0.008
0.15	0.141 ± 0.018	0.081 ± 0.014	0.036 ± 0.013	0.020 ± 0.029
0.2	0.132 ± 0.050	0.096 ± 0.132	0.045 ± 0.014	0.032 ± 0.045
0.25	0.159 ± 0.060	0.047 ± 0.106	0.041 ± 0.022	0.033 ± 0.048
0.3	0.191 ± 0.060	0.104 ± 0.142	0.057 ± 0.018	0.066 ± 0.091
0.35	0.241 ± 0.052	0.052 ± 0.117	0.062 ± 0.018	0.066 ± 0.019
0.4	0.227 ± 0.026	0.112 ± 0.157	0.064 ± 0.017	0.044 ± 0.052
0.45	0.222 ± 0.020	0.094 ± 0.134	0.070 ± 0.013	0.060 ± 0.028

Cost	banknote $\mathrm{CHR}_{\mathrm{PH}}$	banknote $\mathrm{CHR}_{\mathrm{MH}}$	haberman $\mathrm{CHR}_{\mathrm{PH}}$	haberman $\mathrm{CHR}_{\mathrm{MH}}$
0.05	0.015 ± 0.003	0.009 ± 0.021	0.059 ± 0.008	0.059 ± 0.009
0.1	0.021 ± 0.004	0.021 ± 0.007	0.129 ± 0.032	0.108 ± 0.013
0.15	0.028 ± 0.010	0.030 ± 0.067	0.178 ± 0.058	0.178 ± 0.059
0.2	0.029 ± 0.004	0.023 ± 0.004	0.206 ± 0.006	0.214 ± 0.022
0.25	0.031 ± 0.007	0.025 ± 0.006	0.224 ± 0.026	0.244 ± 0.019
0.3	0.032 ± 0.009	0.029 ± 0.007	0.242 ± 0.028	0.232 ± 0.023
0.35	0.037 ± 0.010	0.036 ± 0.008	0.246 ± 0.034	0.246 ± 0.020
0.4	0.041 ± 0.013	0.032 ± 0.008	0.247 ± 0.020	0.250 ± 0.020
0.45	0.038 ± 0.013	0.030 ± 0.007	0.265 ± 0.024	0.259 ± 0.032

Cost	pima $\mathrm{CHR}_{\mathrm{PH}}$	pima $\mathrm{CHR}_{\mathrm{MH}}$
0.05	0.064 ± 0.020	0.035 ± 0.032
0.1	0.106 ± 0.006	0.102 ± 0.003
0.15	0.178 ± 0.051	0.158 ± 0.013
0.2	0.203 ± 0.042	0.163 ± 0.091
0.25	0.230 ± 0.040	0.237 ± 0.030
0.3	0.235 ± 0.019	0.199 ± 0.113
0.35	0.270 ± 0.010	0.268 ± 0.020
0.4	0.258 ± 0.021	0.274 ± 0.020
0.45	0.250 ± 0.021	0.194 ± 0.267