Learning with Rejection

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

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

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

1 Introduction

We consider a flexible binary classification scenario where the learner is given the option to reject an instance instead of predicting its label, thereby incurring some pre-specified cost, typically less than that of a random prediction. While classification with a rejection option has received little attention in the past, it is in fact a scenario of great significance that frequently arises in applications. Incorrect predictions can be costly, especially in applications such as medical diagnosis and bioinformatics. In comparison, the cost of abstaining from prediction, which may be that of additional medical tests, or that of routing a call to a customer representative in a spoken-dialog system, is often more acceptable. From a learning perspective, abstaining from fitting systematic outliers can also result in a more accurate predictor. Accurate algorithms for learning with rejection can further be useful to developing solutions for other learning problems such as active learning [2].

One of the most influential works in this area has been that of Bartlett and Wegkamp [1] who studied a natural discontinuous loss function taking into account the cost of a rejection. They used consistency results to define a convex and continuous *Double Hinge Loss* (DHL) surrogate loss upper-bounding that rejection loss, which they also used to derive an algorithm. A series of follow-up articles further extended this publication, including [8] which used the same convex surrogate while focusing on the l_1 penalty. Grandvalet et al. [5] derived a convex surrogate based on [1] that aims at estimating conditional probabilities

only in the vicinity of the threshold points of the optimal decision rule. They also provided some preliminary experimental results comparing the DHL algorithm and their variant with a naive rejection algorithm. Under the same rejection rule, Yuan and Wegkamp [7] studied the infinite sample consistency for classification with a reject option.

In this paper, we introduce a novel framework for classification with a rejection option that consists of simultaneously learning a pair of functions (h, r): a predictor h along with a rejection function r, each selected from a different hypothesis set. This is a more general framework than that the special case of confidence-based rejection studied by Bartlett and Wegkamp [1] and others, where the rejection function is constrained to be a thresholded function of the predictor's scores. Our novel framework opens up a new perspective on the problem of learning with rejection for which we present a full theoretical 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 [1]. Section 3 provides generalization guarantees for learning with rejection. It also analyzes two convex surrogates of the loss along with consistency results and provides margin-based learning guarantees. Note that many of the proofs are omitted due to space limitations. In Section 4, we present an algorithm with kernel-based hypothesis sets derived from our learning bounds. Lastly, we report the results of several experiments comparing the performance of our algorithms with that of DHL.

2 Learning problem

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

2.1 General rejection model

The learning scenario we consider is that of binary classification with rejection. Let ® denote the rejection symbol. For any given instance $x \in \mathcal{X}$, the learner has the option of abstaining or *rejecting* that instance and returning the symbol ®, or assigning to it a label $\widehat{y} \in \{-1, +1\}$. If the learner rejects an instance, it incurs some loss $c(x) \in [0, 1]$; if it does not reject but assigns an incorrect label,

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

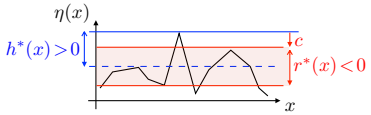

Fig. 1. Mathematical expression and illustration of the optimal classification and rejection function for the Bayes solution. Note, as c increases, the rejection region shrinks.

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 $\text{sign}(h)$ and where $r\colon \mathcal{X} \to \mathbb{R}$ is a function determining the points $x \in \mathcal{X}$ to be rejected according to $r(x) \leq 0$.

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

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

for any pair of functions (h, r) and labeled sample $(x, y) \in \mathcal{X} \times \{-1, +1\}$, thus extending the loss function considered by [1]. In what follows, we assume for simplicity that c is a constant function, though part of our analysis is applicable to the general case. Observe that for $c \geq \frac{1}{2}$, on average, there is no incentive for rejection since a random guess can never incur an expected cost of more than $\frac{1}{2}$. For biased distributions, one may further limit c to the fraction of the smallest class. For $c = 0$, we obtain a trivial solution by rejecting all points, so we restrict c to the case of $c \in]0, \frac{1}{2}[$.

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

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

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

2.2 Confidence-based rejection model

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

This specific choice of the rejection function r is natural when considering the Bayes solution (h^*, r^*) of the learning problem, that is the one where the distribution $\mathcal{D}$ is known. Indeed, for any $x \in \mathcal{X}$, let $\eta(x)$ be defined by $\eta(x) = \mathbb{P}[Y = +1|x]$. As shown in [1], we can choose the Bayes solution h^* and r^* as in Figure 1, which also provides an illustration of confidence-based rejection.

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 $\gamma > 0$.

However, when predictors are selected out of a limited subset $\mathcal{H}$ of all measurable functions over $\mathcal{X}$, requiring the rejection function r to be defined as $r(x) = |h(x)| - \gamma$, for some $h \in \mathcal{H}$, can be too restrictive. Consider, for example, the case where $\mathcal{H}$ is a family of linear functions. 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 $(h, r) \in \mathcal{H} \times \mathcal{R}$:*

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

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

3.2 Convex surrogate losses

We first present general convex upper bounds on the rejection loss. Let $u \mapsto \Phi(-u)$ and $u \mapsto \Psi(-u)$ be convex functions upper-bounding $1_{u \leq 0}$. Since for any

$a, b \in \mathbb{R}$, $\max(a,b) = \frac{a+b+|b-a|}{2} \geq \frac{a+b}{2}$, the following inequalities hold with $\alpha > 0$ and $\beta > 0$:

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

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

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

3.3 Consistency results

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

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

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

1. *the surrogate loss L_{MH} is calibrated with respect to the Bayes classifier: $\operatorname{sign}(h^*) = \operatorname{sign}(h^*_M)$ and $\operatorname{sign}(r^*) = \operatorname{sign}(r^*_M)$;*

2. *furthermore, the following equality holds for the infima over pairs of measurable functions:*

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

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

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

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

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 $L_{\mathrm{MH}}^{\rho,\rho'}(h,r,x,y) = \max\big(\max\big(1+\frac{\alpha}{2}\big(\frac{r(x)}{\rho'} - \frac{yh(x)}{\rho}\big), 0\big), \max\big(c\,\big(1-\beta\frac{r(x)}{\rho'}\big), 0\big)\big)$. 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 4. *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 $(h,r) \in \mathcal{H} \times \mathcal{R}$:*

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

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

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

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

Then, via [6], the bound of Corollary 1 can be shown to hold uniformly for all $\rho, \rho' \in (0,1)$, at the price of a term in $O\Big(\sqrt{\frac{\log\log 1/\rho}{m}} + \sqrt{\frac{\log\log 1/\rho'}{m}}\Big)$.

4 Algorithms for kernel-based hypotheses

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

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

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

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

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

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

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

This learning bound guides directly the definition of our first algorithm based on the L_{MH} (see full version [3] for details) resulting in the following optimization:

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

where $\lambda, \lambda' \geq 0$ are parameters and b and b' are explicit offsets for the linear functions h and r. Similarly, we use the learning bound to derive a second algorithm based on the loss L_{PH} (see full paper [3]). 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 5 we report experimental results with these algorithms as well as the details of our implementation.

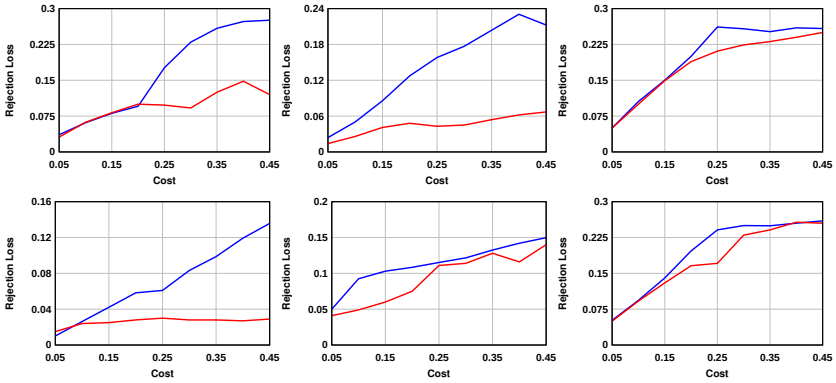

Fig. 3. Average rejection loss on the test set as a function of cost c for the DHL algorithm and the CHR algorithm for six datasets and polynomial kernels. The blue line is the DHL algorithm while the red line is the CHR algorithm based on L_{MH}. The figures on the top starting from the left are for the `cod`, `skin`, and `haberman` data set while the figures on the bottom are for `banknote`, `australian` and `pima` data sets. These figures show that the CHR algorithm outperforms the DHL algorithm for most values of cost, c, across all data sets.

5 Experiments

In this section, we present the results of several experiments comparing our CHR algorithms with the DHL algorithm. All algorithms were implemented using CVX [4]. We tested the algorithms on seven data sets from the UCI data repository, specifically `australian`, `cod`, `skin`, `liver`, `banknote`, `haberman`, and `pima`. For each data set, we performed standard 5-fold cross-validation. We randomly divided the data into training, validation and test set in the ratio 3:1:1. We then repeated the experiments five times where each time we used a different random partition.

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

For each fixed value of c, we chose the parameters with the smallest average rejection loss on the validation set. For these parameter values, Figure 3 shows the corresponding rejection loss on the test set for the CHR algorithm based on L_{MH} and the DHL algorithm 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. Thus, the CHR algorithm yields an improvement in the rejection loss over the DHL algorithm.

References

[1] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 2008.

[2] K. Chaudhuri and C. Zhang. Beyond disagreement-based agnostic active learning. In *NIPS*, 2014.

[3] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *ALT*, 2016.

[4] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0, Aug. 2012.

[5] Y. Grandvalet, J. Keshet, A. Rakotomamonjy, and S. Canu. Suppport vector machines with a reject option. In *NIPS*, 2008.

[6] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

[7] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimizations. In *JMLR*, 2010.

[8] M. Yuan and M. Wegkamp. SVMs with a reject option. In *Bernoulli*, 2011.