Regression with Cost-based Rejection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning with rejection is an important framework that can refrain from making predictions to avoid critical mispredictions by balancing between the rejection costs and prediction errors. Previous cost-based studies only focused on the classification setting, which cannot handle the continuous and infinite target space in the regression setting. In this paper, we investigate a novel regression problem called regression with cost-based rejection, where the model can reject to make predictions on some examples given certain rejection costs. To solve this problem, we first formulate the expected risk for this problem and then derive the Bayes optimal solution, which shows that the optimal model should reject to make predictions on the examples whose variance is larger than the rejection cost when the mean squared error is used as the evaluation metric. Furthermore, we propose to train the model by a surrogate loss function that considers rejection as binary classification and provides conditions for the consistency, where consistency implies that the Bayes optimal can be recovered by our proposed surrogate loss. Extensive experiments demonstrate the effectiveness of our proposed method.

1 Introduction

In machine learning, the learned model from training data is expected to make predictions on unknown test data as accurately as possible. However, it would be unreasonable for the learned model to make predictions on all the test instances, as there may exist some difficult instances that the learned model cannot give an accurate prediction. Incorrect predictions can cause severe consequences and even can be life-threatening, especially in risk-sensitive applications such as healthcare management, autonomous driving, and product inspection [4, 18, 33, 10]. Therefore, the *learning with rejection* (LwR) framework was extensively investigated, which aims to provide a reject option to not make a prediction in order to prevent critical false predictions at a pre-defined rejection cost [9, 8]. In this case, the LwR model can be learned by balancing the rejection cost and the prediction error.

So far, most of the existing studies on LwR have focused on the classification setting, i.e., *classification with rejection* (CwR) [8, 3, 40, 10, 5, 11, 13, 17, 34]. In the CwR setting, there is a pre-determined rejection cost c for each instance, which must be smaller than the classification error 1. A typical approach for CwR is the *confidence-based approach* [21, 3, 40, 33, 6]. The main idea is to use the real-valued output of the classifier as the confidence score and decide whether to reject the prediction based on the confidence score and the given rejection cost c. Another effective approach is *classifier-rejector approach* [10, 11], which simultaneously trains a classifier and a rejector, and this approach achieves state-of-the-art performance in binary classification.

Despite many previous studies on LwR, they only focused on the classification setting, which cannot handle the continuous and infinite target space in the regression setting. In many real-world scenarios, regression tasks with continuous real-valued targets can be commonly encountered. However, even state-of-the-art regression models may make incorrect predictions, and blindly trusting the model results may lead to critical consequences, especially in risk-sensitive applications. Therefore, it is necessary to consider adding a rejection option for the regression problem to not make predictions in order to avoid critical mispredictions. To this end, many studies have been conducted on *selective regression* [41, 25, 38, 19, 24] that trains a regression model with a reject option given a fixed reject rate of predictions. However, this selective regression setting fails to consider the cost-based rejection scenario where a certain cost could be incurred if the model chooses to refrain from making a prediction for a certain instance.

In this paper, we provide the first attempt to investigate a novel regression setting called *regression with cost-based rejection* (RcR), where the model could reject to make predictions on some instances at certain costs to avoid critical mispredictions. To solve the RcR problem, we first formulate the expected risk and then derive the Bayes optimal solution, which shows that the optimal model should reject to make predictions on the examples whose variance is larger than the rejection cost when the popular mean squared error is used as the regression loss. However, it is difficult to directly optimize the expected risk to derive the optimal solution, since the variance of the instances cannot be easily accessed. Therefore, we propose a surrogate loss function to train the model that considers the rejection behavior as a binary classification and we provide theoretical analyses to show that the Bayes optimal solution can be recovered by minimizing our surrogate loss under mild conditions. Our main contributions can be summarized as follows:

- We formulate the expected risk for regression with cost-based rejection and derive the Bayes optimal solution, which shows that the example whose variance is greater than the rejection cost should be rejected for prediction when the mean squared error is used as the regression loss.
- We propose a surrogate loss function considering rejection as a binary classification process and give a condition of regressor-consistent that the classification calibrated binary classification loss is always greater than 0. In that condition, the optimal regressor can be derived by our method.
- We propose a definition of rejector-calibration and show that our method is rejector-calibration when the regressor-consistent condition is satisfied. Based on this, we further propose a weaker version of the condition allowing the classification calibrated binary classification loss to be greater than or equal to 0. In the weakened condition, the regression consistency can only be satisfied in the accepted instances, and regressor-consistent is still satisfied.
- We derive the theoretical analysis of the regret transfer and estimation error bounds for our proposed method, and extensive experiments demonstrate the effectiveness of our method.

2 Preliminaries

In this section, we introduce preliminaries of ordinary regression and classification with rejection.

2.1 Ordinary Regression

For the ordinary regression problem, let the feature space be $\mathcal{X} \in \mathbb{R}^d$ and the label space be $\mathcal{Y} \in \mathbb{R}$. Let us denote by $(\boldsymbol{x}, y)$ an example including an instance x and a real-valued label y. Each example $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is assumed to be independently sampled from an unknown data distribution with probability density $p(\boldsymbol{x}, y)$. For the regression task, we aim to learn a regression model $h : \mathcal{X} \mapsto \mathbb{R}$ that minimizes the following expected risk:

$$R(L) = \mathbb{E}_{p(\boldsymbol{x},y)}[L(h(\boldsymbol{x}), y)], \tag{1}$$

where $\mathbb{E}_{p(\boldsymbol{x},y)}$ denotes the expectation over the data distribution $p(\boldsymbol{x}, y)$ and $L : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ is a conventional loss function (such as mean squared error and mean absolute error) for regression, which measures how well a model estimates a given real-valued label.

2.2 Classification with Rejection

A widely studied framework in classification with rejection is the cost-based framework [8, 15] that aims to train a classifier $f : \mathcal{X} \mapsto \mathcal{Z}^{®}$ that can reject to make a prediction, where ® denotes the reject option. The evaluation metric of this task is the zero-one-c loss ℓ_{01c} defined as follows:

$$\ell_{01c}(f(\boldsymbol{x}), z) = \begin{cases} c, & f(\boldsymbol{x}) = ®, \\ \ell_{01}(f(\boldsymbol{x}, z), & \text{otherwise}, \end{cases} \tag{2}$$

Then, the expected risk with ℓ_{01c} can be represented as follows:

$$R_{01c}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell_{01c}(f(\boldsymbol{x}), y)], \tag{3}$$

The optimal solution for classification with rejection $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} R_{01c}(f)$ known as Chow's rule [8] can be expressed as follows:

Definition 1. *(Chow's Rule [8]) A classifier $f : \mathcal{X} \to \mathcal{Z}^{®}$ is the optimal solution of expected risk (3) if and only if the following conditions are almost satisfied:*

$$f(\boldsymbol{x}) = \begin{cases} ®, & \max_z \eta_z(\boldsymbol{x}) \leq 1 - c, \\ \operatorname{argmax}_z \eta_z(\boldsymbol{x}), & \text{otherwise}, \end{cases} \tag{4}$$

where $\eta_z(\boldsymbol{x}) = p(z|\boldsymbol{x})$ denotes the class-prior estimation (CPE) [35, 39]. Chow's rule shows that CwR can be solved when $\boldsymbol{\eta}(\boldsymbol{x})$ is known. However, the estimation of the posterior probability is difficult especially when using deep neural networks [22].

3 Regression with Cost-based Rejection

Let $\mathcal{X} \in \mathbb{R}^d$ be the d-dimensional feature space and $\mathcal{Y} \in \mathbb{R}$ be the label space. Suppose the training set is denoted by $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and each training example $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ is assumed to be sampled from an unknown data distribution with probability density $p(\boldsymbol{x}, y)$. In the regression with cost-based rejection (RcR) setting, for a given instance $\boldsymbol{x}$, the learner has the option ® to reject making a prediction or to make a regression prediction. If the learner rejects an instance, the cost is a non-negative loss $c(\boldsymbol{x})$. The goal of RcR is to induce a pair (h, r) where $h : \mathcal{X} \mapsto \mathbb{R}$ is a regressor to predict the accepted instance and $r : \mathcal{X} \mapsto \mathbb{R}$ is a rejector to determine whether to reject an instance. The evaluation metric of this task is the following loss function $\mathcal{L}(h, r, c, \boldsymbol{x}, y)$:

$$\mathcal{L}(h, r, c, \boldsymbol{x}, y) = \begin{cases} L(h(\boldsymbol{x}), y), & r(\boldsymbol{x}) > 0, \\ c(\boldsymbol{x}), & \text{otherwise}, \end{cases} \tag{5}$$

where $L(h(\boldsymbol{x}), y)$ is a conventional regression loss function (e.g., mean squared error).

In what follows, we will present a Bayes optimal solution to the RcR problem and provide a surrogate loss function to train the regressor-rejector.

3.1 Bayes Optimal Solution

In this paper, we only discuss the case where the loss function $L(h(\boldsymbol{x}), y)$ is the mean squared error (MSE), which is the most widely used regression loss function. The expected risk of $\mathcal{L}(h, r, c, \boldsymbol{x}, y)$ over the data distribution can be represented as follows:

$$R_{\mathrm{RcR}}(h, r) = \mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}(h, r, c, \boldsymbol{x}, y)]. \tag{6}$$

Let us denote by $(h^\star, r^\star) = \operatorname{argmin}_{(h,r)} R_{\mathrm{RcR}}(h, r)$ the optimal pair of expected risk R_{RcR} and we use $\mathbb{E}_{p(y|\boldsymbol{x})}[y] = \int_{\mathcal{Y}} p(y|\boldsymbol{x}) y \mathrm{d}y$ and $\mathbb{D}_{p(y|\boldsymbol{x})}[y] = \int_{\mathcal{Y}} p(y|\boldsymbol{x})(y - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 \mathrm{d}y$ represent the expectation and variance of y over the distribution $p(y|\boldsymbol{x})$. For a given cost function $c(\boldsymbol{x})$, we have the following theorem:

Theorem 2. *Suppose the hypothesis space $\mathcal{H}$ and $\mathcal{R}$ is strong enough [16, 29] (i.e., the optimal solution $(h^\star, r^\star) = \operatorname{argmin}_{h \in \mathcal{H}, r \in \mathcal{R}} R_{\mathrm{RcR}}(h, r)$ leads to $R_{\mathrm{RcR}}(h^\star, r^\star) = 0$). For a given instance $\boldsymbol{x}$ and the Bayes optimal pair $(h^\star, r^\star)$ of risk R_{RcR}, the following equality holds:*

$$\begin{cases} h^\star(\boldsymbol{x}) = \mathbb{E}_{p(y|\boldsymbol{x})}[y], \\ r^\star(\boldsymbol{x}) = \mathbb{I}\left(c(\boldsymbol{x}) - \mathbb{D}_{p(y|\boldsymbol{x})}[y]\right). \end{cases} \tag{7}$$

The proof of Theorem 2 is provided in Appendix A. Theorem 2 shows the expected optimal pair $(h^\star, r^\star)$ of risk R_{RcR} where the rejector $r^\star$ should reject making a prediction if the variance of the distribution of labels y associated with x is so large that it exceeds a given rejection cost $c(\boldsymbol{x})$. This is intuitive and easy to understand. Unfortunately the probability density function $p(y|\boldsymbol{x})$ is usually unknown, meaning that obtaining the variance $\mathbb{D}_{p(y|\boldsymbol{x})}[y]$ and expectation $\mathbb{E}_{p(y|\boldsymbol{x})}[y]$ is difficult or even impossible. If the variance and expectation can be obtained, most of the regression tasks can be easily solved. Many previous studies adopted specific assumptions to avoid this problem (e.g., homoscedasticity [24, 37, 36] and heteroscedasticity) [26, 27, 7, 28], while all of them have certain constraints. Therefore, the key challenge of RcR is how to learn the optimal solution $(h^\star, r^\star)$ without the expectation and the variance.

3.2 Surrogate Loss Function of Training Regressor-Rejector

From Theorem 2, we know how the optimal pair $(h^\star, r^\star)$ makes rejection and prediction for an unknown instance, but since the expectation and the variance are difficult to obtain, we cannot directly derive the optimal regressor and rejector. Let us reconsider the RcR loss function $\mathcal{L}(h, r, c, \boldsymbol{x}, y)$ by the following equation:

$$\mathcal{L}(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2 \mathbb{I}[r(\boldsymbol{x}) > 0] + c(\boldsymbol{x})\mathbb{I}[r(\boldsymbol{x}) \leq 0], \tag{8}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. We cannot directly derive a regressor h and a rejector r by the above loss since the loss function contains non-convex and discontinuous parts $\mathbb{I}[r(\boldsymbol{x}) > 0]$ and $\mathbb{I}[r(\boldsymbol{x}) \leq 0]$. In order to efficiently optimize the target loss, using surrogate loss is preferred. It is noteworthy that the behavior of the rejector is similar to binary classification due to the only two options reject and accept. We may consider it directly as a binary classification where $\mathcal{Z} = \{+1, -1\}$, $+1$ means accept and -1 means reject. Then we have the following surrogate loss function:

$$\psi(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2 \ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1), \tag{9}$$

where $\ell(\cdot)$ is an arbitrary binary classification loss function such as hinge loss. Then the expected risk with our surrogate loss ψ can be represented as follows:

$$R^{\psi}_{\mathrm{RcR}}(h, r) = \mathbb{E}_{p(\boldsymbol{x},y)}[\psi(h, r, c, \boldsymbol{x}, y)]. \tag{10}$$

The intuition behind this is that when the squared error is less than the given cost, we expect its weight $\ell(r(\boldsymbol{x}), -1)$ to be larger i.e. the smaller $\ell(r(\boldsymbol{x}), +1)$ is. It is worth noting that not all binary classification losses are valid, and in the following sections we will show the conditions for our method to satisfy consistency.

4 Theoretical Analysis

4.1 Regressor-Consistent and Rejector-Calibration

The rejector-calibration we are talking about here is the classification-calibration [1, 42, 14] due to the fact that the rejector is actually a classifier. The notion of calibration for surrogate loss is defined as the minimum requirement to ensure that a risk-minimizing classifier satisfies the Bayes optimal classifier, which is a pointwise version of consistency, implying that the minimization of surrogate loss yields a target loss for each possible instance. We further give the definition of rejector calibration.

Definition 3. *(Rejector-Calibration) We say a surrogate loss Φ is rejector-calibration if and only if for the optimal regressor $r^\star_\Phi = \mathrm{argmin}_{r\in\mathcal{R}} R^{\Phi}_{\mathrm{RcR}}(h^\star, r)$, we have $\mathrm{sign}(r^\star_\Phi(\boldsymbol{x})) = \mathrm{sign}(r^\star(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathcal{X}$ such that $r^\star(\boldsymbol{x}) \neq 0$.*

The definition of rejector calibration indicates that we do not need to obtain the optimal rejector based on the difficult to obtain variance, we just need to ensure that our rejector makes the same decisions as the optimal rejector.

We say a method is regressor-consistent, meaning that the regressor h learned by the method converges to the optimal regressor $h^\star$. Here we demonstrate that our method is regressor-consistent and we have the following theorem:

Theorem 4. *Suppose the classification calibrated binary classification loss $\ell(r(\boldsymbol{x}), z)$ can be achieved: $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(r(\boldsymbol{x}), z) > 0$. For given non-negative cost $c(\boldsymbol{x})$, the optimal regressor $h^\star_\psi = \mathrm{argmin}_{h\in\mathcal{H}} R^{\psi}_{\mathrm{RcR}}(h, r)$ is equivalent to the optimal regressor $h^\star = \mathrm{argmin}_{h\in\mathcal{H}} R_{\mathrm{RcR}}(h, r)$.*

The proof of Theorem 4 is provided in Appendix B.1. Theorem 4 shows that the optimal regressor h learned from our method can converge to the optimal regressor $h^\star$. Then we demonstrate that our method is rejector-calibration. We have the following theorem:

Theorem 5. *Suppose the classification calibrated binary classification loss $\ell(r(\boldsymbol{x}), z)$ can be achieved: $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(r(\boldsymbol{x}), z) > 0$. For the given non-negative cost $c(\boldsymbol{x})$, the optimal rejector $r_\psi^\star = \mathrm{argmin}_{r \in \mathcal{R}} R_{\mathrm{RcR}}^\psi(h, r)$ satisfies $\mathrm{sign}(r_\psi^\star(\boldsymbol{x})) = \mathrm{sign}(r^\star(\boldsymbol{x}))$ where $r^\star$ is the optimal rejector of R_{RcR}.*

The proof of Theorem 5 is provided in Appendix B.2. Theorem 5 shows that our method is rejector-consistent in the condition that $\ell(r(\boldsymbol{x}), z) > 0$ holds. When $\ell(r(\boldsymbol{x}), z) \leq 0$, the regressor will show abandonment and aversion to some instances, the condition implying that the regressor needs to ensure that the autonomous learning capability avoids being fully controlled by the rejector. It is worth noting that there is a special case $\ell(r(\boldsymbol{x}), -1) = 0$, in which case the regressor actually ignores the instance. Here we show a weakened version of consistency, we have the following theorem:

Theorem 6. *Suppose the classification calibrated binary classification loss $\ell(r(\boldsymbol{x}), z)$ can be achieved: $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(r(\boldsymbol{x}), z) \geq 0$. For given non-negative cost $c(\boldsymbol{x})$, the optimal pair $(h_\psi^\star, r_\psi^\star) = \mathrm{argmin}_{(h,r) \in \mathcal{H} \times \mathcal{R}} R_{\mathrm{RcR}}^\psi(h, r)$ satisfies rejector-calibration and satisfies the regressor-consistent for all $\forall \boldsymbol{x} \in \mathcal{X}$, $r^\star(\boldsymbol{x}) > 0$, where $r^\star$ is the optimal rejector of R_{RcR}.*

The proof of Theorem 6 is provided in Appendix B.3. Theorem 6 gives a weakened version of consistency, where regressor-consistent is satisfied only for accepted samples.

4.2 Regret Transfer and Estimation Error Bounds

In the previous section, we have given the Bayes consistency analysis of our method, i.e., if the minimizer of our proposed risk can be the optimal one in Theorem 2. However, such a result does not guarantee the performance of models which are close to but not the minimizer of the R_{RcR}^ψ, which occurs commonly since we usually minimize the empirical risk in practice. We give a guarantee for such cases by showing the following regret transfer bound:

Theorem 7. *For any classification calibrated binary classification loss ℓ, suppose that the variance $\mathbb{D}_{p(y|\boldsymbol{x})}[y] \leq M$ almost surely, the following bound holds:*

$$R_{\mathrm{RcR}}(h, r) - R_{\mathrm{RcR}}^* \leq \xi(C(R_{\mathrm{RcR}}^\psi(h, r) - R_{\mathrm{RcR}}^{\psi *})),$$

where * denotes the minimum w.r.t. h and r. $C = M + c$, and ξ is a function where $\xi(0) = 0$. For example, when ℓ is sigmoid loss and hinge loss, $\xi(u) = u$. When ℓ is logistic loss or square loss, $\xi(u) = \sqrt{u}$.

The proof of Theorem 7 is provided in Appendix C.1. This theorem guarantees that even if the obtained (h, r) is not exactly the minimizer of R_{RcR}^ψ, we can also expect them to have a good performance as long as they have low R_{RcR}^ψ. Then we can further get the following estimation error bound:

Theorem 8. *Suppose the hypothesis space $\mathcal{H}$ and $\mathcal{R}$ is strong enough. Given empirical risk minimizer $\hat{h}$ and $\hat{r}$, there exists $\alpha_1, \alpha_2 > 0$ that make the following bound holds with probability at least $1 - \delta$:*

$$R_{\mathrm{RcR}}(\hat{h}, \hat{r}) - R_{\mathrm{RcR}}^* \leq \xi\left(C\left(\alpha_1 \mathfrak{R}_n(\mathcal{H}) + \alpha_2 \mathfrak{R}_n(\mathcal{R}) + \sqrt{\frac{\log(1/\delta)}{2n}}\right)\right),$$

where n is the $i.i.d.$ sample size and $\mathfrak{R}_n$ is the Rademacher complexity [2].

The proof of Theorem 8 is provided in Appendix C.2. Given the fact that Rademacher complexity usually decays at the rate of $\mathcal{O}(1/n)$, we can finally conclude that the performance of our model can approximate its optimal performance with the increasing size of the training set.

5 Experiments

5.1 Implementation Details

When using deep neural networks as the model and using gradient descent optimization, we consider a possible scenario where the regressor h predicts any instance $\boldsymbol{x}$ with such a large error that

$\ell(h(\boldsymbol{x}), y) >> c(\boldsymbol{x})$. In this case the rejector r expects to reject all instances to make the empirical risk minimal. However, when the rejector r converges quickly to reject all train instances, i.e., $\ell(r(\boldsymbol{x}), -1) \to 0$ for all train instances, the surrogate loss ψ will be constant equal to $c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)$. At that point the gradient of the regressor h suffers from gradient vanishing. The main reason for this situation is that the regressor h has not learned the distribution of the label, but the rejector r has converged, which means that the regressor is not ready. Fortunately, we can avoid such a situation by training the rejector after the regressor is ready, and we name such a method Slow-Start. Specifically, Slow-Start prioritizes training the regressor h without training the rejector r, and then co-trains the regressor h and rejector r when the regressor h is capable of making predictions.

5.2 Datasets and Backbone Models

We conduct experiments on seven datasets, including one computer vision dataset (AgeDB [32]), one healthcare dataset (BreastPathQ [30]), and five datasets from the UCI Machine Learning Repository [12] (Abalone, Airfoil, Auto-mpg, Housing and Concrete). For each dataset, we randomly split the original dataset into training, validation, and test sets by the proportions of 60%, 20%, and 20%, respectively. It is worth noting that our approach has no restrictions on the regressor h and rejector r, so h and r can be two separate parts or share parameters.

AgeDB is a regression dataset on age prediction [20] collected by [32]. It contains 16.4K face images with a minimum age of 0 and a maximum age of 101. Age prediction is not an easy task, especially when only a single photo is available. Lighting, clothing, makeup, and facial expressions all tend to affect the intuitive age, and even friends can hardly say they can identify the age in a photo. Rejecting predictions for photos with complex environments can avoid large errors. We employ ResNet-50 [23] as our backbone network for AgeDB, and the regressor h and rejector r share parameters. We use the Adam optimizer to train our method for 100 epochs where the slow-start is set to 40 epochs, the initial learning rate of 10^{-3} and fix the batch size to 256.

BreastPathQ [30] is a healthcare dataset collected at the Sunnybrook Health Sciences Centre, Toronto. The dataset contains 2579 patch images, each patch has been assigned a tumor cellularity score score of 0 to 1 by 1 expert pathologist. Currently, this task is performed manually and relies upon expert interpretation of complex tissue structures. Moreover, cancer cellularity scoring is extremely risky and the use of automated methods could lead to irreversible disasters. Regression with rejection can improve this problem very well by predicting only the accepted samples and leaving the rejected samples back to the experts for evaluation. We use the same network as AgeDB and train 300 epochs using Adam optimizer where the slow-start is set to 50 epochs, the initial learning rate of 10^{-3} and fix the batch size to 128.

We conducted experiments on five UCI benchmark datasets including Abalone, Airfoil, Auto-mpg, Housing and Concrete. All of these datasets can be downloaded from the UCI Machine Learning [12]. Since our proposed method do not depend on a specific model, and we train two types of base models including the linear model and the multilayer perceptron (MLP) to support the flexibility of our method on choosing a model, where the MLP model is a five-layer (d-20-30-10-1) neural network with a ReLU activation function. For the rejector r and regressor h, we consider them as two separate parts with the same structure. For both the linear model and the MLP model, we use the Adam optimization method with the batch size set to 1024 and the number of training epochs set to 1000 where the slow-start is set to 200 epochs. The learning rate for all UCI benchmark datasets is selected from $\{10^{-1}, 10^{-2}, 10^{-3}\}$.

5.3 Evaluation Metrics

For evaluation metrics, we use the RcR loss (RcRLoss) in Eq. (5) and rejection ratio (RR). In order to further investigate how the model work, we propose additional metrics. Accepted loss (AL) and rejection loss (RL) denote losses on accepted instances and rejected instances, and they are defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i)>0](h(\boldsymbol{x}_i)-y_i)^2}{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i)>0]}$ and $\frac{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i)\leq 0](h(\boldsymbol{x}_i)-y_i)^2}{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i)\leq 0]}$. We also present the false rejection ratio (AR) and false acceptance ratio (RA) similar to false negative and false positive, which denote the ratio of instances that should be accepted that are rejected and the ratio of instances that should be rejected that are accepted, and they are defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i)-y_i)^2<c(\boldsymbol{x}_i)]\mathbb{I}[r(\boldsymbol{x}_i)\leq 0]}{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i)-y_i)^2<c(\boldsymbol{x}_i)]}$ and $\frac{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i)-y_i)^2\geq c(\boldsymbol{x}_i)]\mathbb{I}[r(\boldsymbol{x}_i)>0]}{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i)-y_i)^2\geq c(\boldsymbol{x}_i)]}$. It is worth noting that the optimal pair $(h^\star, r^\star)$ is unknown, so

Table 1: Test performance (mean and std) of our surrogate loss equipped MAE on BreastPathQ. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100 and Sup, RcRLoss, AL and RL are all magnified by a factor of 1000.

Cost	Sup	RcRLoss	AL	RL	RR	AR	RA
5	16.77 (1.22)	4.37 (0.17)	2.70 (1.07)	31.51 (2.29)	72.53 (4.44)	52.61 (5.10)	6.53 (2.66)
10		8.22 (0.70)	5.50 (1.98)	37.14 (4.43)	60.08 (4.34)	43.14 (4.49)	11.01 (4.22)
15		11.11 (0.55)	6.84 (1.43)	40.39 (1.67)	53.49 (3.39)	38.39 (2.86)	15.46 (3.97)
20		13.84 (0.62)	9.53 (1.69)	43.41 (5.34)	40.65 (7.28)	29.98 (6.58)	29.02 (9.81)
25		16.01 (1.32)	12.91 (2.48)	46.62 (9.43)	24.47 (4.26)	17.46 (4.96)	48.97 (8.00)

Table 2: Test performance (mean and std) of our surrogate loss equipped MAE on AgeDB. We repeat the sampling-and-training process 5 times. The metrics RR, AR and RA are scaled to 0-100.

Cost	Sup	RcRLoss	AL	RL	RR	AR	RA
60	100.34 (3.73)	59.80 (0.31)	54.25 (4.41)	156.81 (23.21)	95.40 (2.88)	93.13 (4.30)	2.51 (1.56)
70		69.00 (0.39)	61.56 (4.10)	151.04 (12.05)	86.22 (2.94)	81.41 (3.07)	8.12 (2.49)
80		77.10 (1.72)	67.32 (2.21)	150.52 (12.36)	76.00 (15.71)	70.63 (16.36)	16.11 (13.20)
90		85.36 (2.23)	73.07 (3.21)	162.44 (12.45)	73.38 (11.50)	67.33 (12.07)	17.20 (9.08)
100		92.94 (3.02)	82.89 (7.47)	170.04 (20.53)	58.35 (12.51)	52.15 (11.59)	30.56 (12.48)
110		95.08 (5.62)	79.62 (5.44)	166.07 (13.75)	52.15 (14.96)	46.13 (14.76)	34.38 (13.40)
120		96.80 (7.45)	82.44 (2.40)	173.14 (12.58)	37.11 (22.64)	32.54 (21.42)	51.31 (23.96)

AR and RA are for the current regressor and rejector. We also provide the results under supervised regression method (Sup) that directly trains the model with MSE from fully training set.

5.4 Formulation of Surrogates and Setting of Rejection Cost

In our experiments, we consider a variety of binary classification loss functions, such as mean squared error (MAE), square loss, logistic loss, sigmoid and hinge loss. The rejection cost $c(\boldsymbol{x})$ is considered as a constant, which is the most commonly considered scenario in learning with rejection [5, 6, 33, 10]. For each dataset, we set various rejection cost c including extreme cases and unstressed cases depending on the supervised loss. The *complete* experiments are provided in Appendix D.

5.5 Experimental Performance

Table 1, Table 2, Table 3, and Table 4 show some of the experimental results on the AgeDB, BreastPathQ, and UCI datasets, respectively. From the four tables, we have the following observations: (1) Our proposed method significantly outperforms the supervised regression method in almost all cases, which validates the ability of our method to reject difficult test instances demonstrating the effectiveness of our method. (2) In most cases, the average loss of our method in the accepted test instances (AL) is always smaller than the average loss of the supervised regression model (Sup) in all test instances. This further indicates the ability of our method to identify hard-to-predict samples and reject them. (3) As the rejection cost c increases, we can clearly see the following trends in all datasets: RcR loss (RcRLoss) decreases; Rejection rate (RR) decrease; Accepted test data loss (AL) increases; This is because as the prediction error we can accept increases, the rejector will accept more instances leading to a decrease in the rejection rate. However, the regressor capacity remains

Table 3: Test performance (mean and std) of our surrogate loss on five UCI datasets trained with the MLP model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Sup	RcRLoss	AL	RL	RR	AR	RA
Abalone	3	4.44 (0.46)	2.41 (0.12)	1.99 (0.21)	8.13 (1.08)	42.04 (3.18)	32.82 (3.44)	33.33 (3.22)
	4		2.88 (0.13)	2.30 (0.21)	11.37 (1.70)	33.70 (2.47)	25.56 (2.81)	39.27 (3.71)
	5		3.22 (0.23)	2.66 (0.35)	10.30 (1.25)	23.43 (2.94)	16.83 (2.41)	48.98 (5.90)
	6		3.53 (0.25)	2.93 (0.35)	12.13 (1.69)	19.32 (3.47)	13.81 (3.33)	53.20 (5.67)
Airfoil	9	12.96 (2.60)	7.20 (0.35)	4.23 (0.86)	37.80 (2.95)	62.23 (3.73)	41.49 (5.73)	11.60 (3.29)
	12		8.11 (0.36)	5.39 (0.86)	51.51 (10.51)	40.33 (7.95)	23.37 (7.20)	25.88 (9.04)
	16		9.15 (0.43)	6.84 (0.70)	72.80 (20.79)	24.92 (5.67)	11.92 (6.93)	38.17 (5.02)
	20		11.32 (0.75)	8.83 (1.47)	58.28 (8.87)	21.53 (7.71)	13.70 (5.34)	48.66 (18.08)
	25		11.47 (1.54)	9.24 (1.35)	74.38 (16.07)	14.19 (5.11)	8.08 (3.60)	52.11 (12.32)
	30		11.68 (3.07)	11.17 (3.20)	96.55 (16.60)	2.52 (3.81)	1.38 (1.78)	86.35 (20.06)
Auto-mpg	4	8.34 (2.16)	3.64 (0.29)	2.99 (0.83)	13.98 (4.16)	56.92 (13.00)	46.80 (15.49)	28.74 (10.51)
	6		4.83 (0.93)	3.83 (1.70)	18.04 (5.95)	37.31 (14.10)	29.01 (12.74)	42.42 (19.54)
	8		6.75 (1.93)	6.14 (2.41)	25.59 (12.48)	22.95 (19.88)	19.26 (18.27)	64.99 (23.95)
	10		7.14 (1.64)	6.11 (2.24)	23.29 (9.54)	24.07 (6.58)	17.15 (5.12)	48.47 (15.98)
	13		8.13 (2.41)	7.42 (2.83)	35.49 (23.74)	12.56 (6.83)	10.38 (6.14)	71.52 (14.52)
Housing	9	12.57 (3.43)	8.80 (0.34)	6.25 (3.22)	40.28 (17.30)	84.46 (11.67)	77.60 (15.88)	9.72 (5.91)
	12		9.52 (0.75)	7.40 (1.48)	58.94 (25.98)	44.65 (8.69)	33.30 (8.99)	31.25 (8.64)
	16		10.12 (1.84)	8.35 (1.58)	88.14 (44.53)	22.38 (8.90)	14.21 (6.81)	51.84 (14.41)
	20		10.50 (3.32)	9.59 (3.50)	184.24 (109.35)	8.51 (6.82)	5.81 (5.32)	73.40 (13.11)
Concrete	20	34.44 (3.05)	18.03 (1.32)	13.17 (4.91)	82.17 (14.58)	69.42 (6.92)	54.06 (9.37)	12.34 (4.47)
	30		24.20 (1.85)	19.29 (3.85)	112.13 (30.32)	44.08 (8.81)	27.43 (8.55)	26.80 (7.90)
	40		28.63 (2.56)	23.12 (4.59)	136.51 (46.59)	31.50 (8.98)	18.32 (7.30)	39.49 (12.07)
	50		32.48 (2.76)	27.90 (4.31)	168.19 (41.73)	19.76 (7.54)	10.54 (4.51)	53.82 (13.74)
	60		34.33 (3.50)	30.33 (4.89)	197.26 (49.03)	12.82 (6.62)	5.67 (3.21)	60.95 (14.99)

the same and more instances (containing difficult instances) also face more challenges, so RcRLoss and AL increase but remain smaller than Sup. (4) For setting the rejection cost c we consider many extreme cases, i.e., the rejection cost is much smaller and much larger than the average loss in the supervised regression. In such extreme cases, our approach is still effective to identify and reject difficult test instances. (5) The false acceptance ratio (RA) is usually not large in most cases which verifies that our approach prefers rejection to avoid critical mispredictions.

6 Conclusion

In this paper, we investigated a novel regression problem called regression with cost-based rejection, which aims to learn a model that can reject predictions to avoid critical mispredictions at a certain

Table 4: Test performance (mean and std) of our surrogate loss on five UCI datasets trained with the Linear model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Sup	RcRLoss	AL	RL	RR	AR	RA
Abalone	3	4.92 (0.51)	2.80 (0.09)	2.00 (0.36)	5.88 (0.62)	79.07 (4.68)	72.81 (6.64)	9.92 (1.72)
	4		3.51 (0.14)	2.57 (0.42)	6.34 (0.63)	63.99 (3.67)	57.24 (3.85)	19.14 (4.27)
	5		3.48 (0.30)	2.88 (0.37)	10.93 (2.85)	28.96 (9.71)	21.83 (10.34)	44.72 (8.64)
	6		3.83 (0.26)	3.52 (0.33)	15.00 (3.16)	13.20 (2.82)	8.54 (2.33)	65.48 (5.95)
Airfoil	9	23.32 (1.54)	8.81 (0.27)	6.33 (1.60)	27.22 (2.25)	86.84 (2.31)	80.28 (3.86)	6.65 (1.16)
	12		11.39 (0.40)	7.52 (1.62)	29.39 (2.92)	79.20 (6.42)	71.55 (8.17)	10.80 (4.27)
	16		14.43 (0.84)	10.75 (1.68)	33.24 (2.53)	60.23 (5.69)	51.78 (6.44)	25.06 (5.21)
	20		16.90 (1.02)	12.28 (2.22)	34.24 (2.44)	55.42 (3.93)	47.90 (4.24)	28.38 (5.39)
	25		19.47 (2.25)	15.04 (4.09)	37.36 (3.34)	39.77 (11.48)	33.82 (10.56)	44.50 (15.28)
	30		22.91 (1.49)	21.59 (2.22)	34.64 (6.99)	14.05 (5.68)	12.28 (5.23)	79.58 (8.60)
Auto-mpg	4	11.66 (2.26)	4.07 (0.04)	7.24 (5.03)	13.63 (2.49)	99.85 (0.40)	99.49 (0.94)	3.17 (0.99)
	6		5.97 (1.09)	5.93 (4.00)	16.42 (4.89)	67.18 (11.12)	61.05 (11.77)	25.19 (10.63)
	8		7.29 (1.50)	6.78 (2.67)	21.23 (9.00)	43.33 (6.80)	36.64 (9.99)	44.11 (13.91)
	10		8.17 (1.51)	7.20 (2.28)	22.30 (9.41)	34.61 (7.62)	30.37 (7.93)	53.76 (15.04)
	13		9.47 (1.83)	8.65 (2.33)	34.57 (18.96)	16.41 (7.88)	13.14 (7.85)	69.31 (13.26)
Housing	9	24.08 (5.34)	9.18 (0.53)	7.61 (3.15)	35.54 (13.93)	86.14 (16.32)	81.44 (19.72)	10.83 (10.55)
	12		10.91 (0.76)	9.67 (1.87)	58.37 (19.08)	48.22 (9.97)	39.40 (10.83)	34.95 (10.02)
	16		13.98 (2.97)	12.28 (4.47)	63.37 (18.75)	36.44 (9.35)	30.84 (9.52)	49.01 (10.22)
	20		16.73 (4.61)	14.93 (6.55)	69.52 (23.34)	27.82 (7.06)	22.64 (6.88)	54.32 (11.36)
Concrete	20	111.12 (8.01)	19.85 (0.11)	5.94 (2.66)	115.92 (11.06)	98.91 (0.72)	97.09 (1.92)	0.81 (0.23)
	30		29.93 (0.22)	16.85 (16.90)	114.93 (11.72)	99.24 (0.92)	98.49 (1.84)	1.00 (0.34)
	40		40.18 (1.02)	29.88 (33.68)	117.29 (14.30)	98.71 (2.46)	98.50 (2.37)	1.96 (2.94)
	50		50.44 (1.36)	50.07 (22.99)	130.44 (23.58)	93.98 (6.19)	93.21 (6.77)	5.92 (5.74)
	60		58.69 (2.88)	45.84 (16.77)	150.10 (52.86)	85.33 (9.28)	81.90 (10.75)	10.25 (7.20)

rejection cost. In order to solve this problem, we first formulate the expected risk for regression with cost-based rejection and derive the Bayes optimal solution for the expected risk, which shows that we should reject instances where the variance is greater than the rejection cost. Since the variance is difficult to obtain, we propose a surrogate loss function that considers the rejection process as a binary classifition problem. Further, we provide consistency conditions for our method, implying that the optimal solution can be recovered by our method. More, we propose a weakened version of consistency where regression-consistent is satisfied only in the accepted instances. Finally, we derive the regret transfer and an estimation error bound for our method and conduct extensive experiments on various datasets to demonstrate the effectiveness of our proposed method. We expect that our first study of a simple but theoretically grounded method to regression with rejection will inspire more interesting research work on this new task.

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of the American Statistical Association*, 3:463–482, 2002.

[3] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

[4] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[5] Y. Cao, T. Cai, L. Feng, L. Gu, J. Gu, B. An, G. Niu, and M. Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *NeurIPS*, pages 521–534, 2022.

[6] N. Charoenphakdee, Z. Cui, Y. Zhang, and M. Sugiyama. Classification with rejection based on cost-sensitive classification. In *ICML*, pages 1507–1517, 2021.

[7] K. Chaudhuri, P. Jain, and N. Natarajan. Active heteroscedastic regression. In *ICML*, pages 694–702, 2017.

[8] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[9] C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

[10] C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. In *NeurIPS*, 2016.

[11] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *ALT*, pages 67–82, 2016.

[12] D. Dua and C. Graff. UCI machine learning repository, 2017.

[13] R. El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

[14] J. Finocchiaro, R. Frongillo, and B. Waggoner. An embedding framework for consistent polyhedral surrogates. In *NeurIPS*, 2019.

[15] V. Franc and D. Prusa. On discriminative learning of prediction uncertainty. In *ICML*, pages 1963–1971, 2019.

[16] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

[17] W. Gao and Z.-H. Zhou. On the consistency of auc pairwise optimization. *arXiv preprint arXiv:1208.0645*, 2012.

[18] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

[19] Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *ICML*, pages 2151–2159, 2019.

[20] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[21] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu. Support vector machines with a reject option. In *NeurIPS*, 2008.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[24] C. M. Jarque and A. K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259, 1980.

[25] W. Jiang, Y. Zhao, and Z. Wang. Risk-controlled selective prediction for regression deep neural network models. In *IJCNN*, pages 1–8, 2020.

[26] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heteroscedastic gaussian process regression. In *ICML*, pages 393–400, 2007.

[27] M. Lázaro-Gredilla and M. K. Titsias. Variational heteroscedastic gaussian process regression. In *ICML*, pages 841–848, 2011.

[28] H. Liu, Y.-S. Ong, and J. Cai. Large-scale heteroscedastic regression via gaussian process. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):708–721, 2020.

[29] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama. Progressive identification of true labels for partial-label learning. In *ICML*, 2020.

[30] A. L. Martel, S. Nofech-Mozes, S. Salama, S. Akbar, and M. Peikari. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology [data]. 2019.

[31] A. Maurer. A vector-contraction inequality for rademacher complexities. In *ALT*, pages 3–17, 2016.

[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *CVPRW*, pages 1997–2005, 2017.

[33] C. Ni, N. Charoenphakdee, J. Honda, and M. Sugiyama. On the calibration of multiclass classification with rejection. In *NeurIPS*, 2019.

[34] H. G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*, 2015.

[35] M. D. Reid and R. C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

[36] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21, 1985.

[37] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

[38] Y. Wiener and R. El-Yaniv. Pointwise tracking the optimal regression function. In *NeurIPS*, 2012.

[39] R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016.

[40] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(1), 2010.

[41] A. Zaoui, C. Denis, and M. Hebiri. Regression with reject option and application to knn. In *NeurIPS*, pages 20073–20082, 2020.

[42] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

A Proof of Theorem 2

For an instance $\boldsymbol{x}$, we have the following expected risk for $\boldsymbol{x}$:

$$\begin{aligned}
R_{\mathrm{RcR|x}}(h,r) &= \mathbb{E}_{p(y|\boldsymbol{x})}[\mathcal{L}(h,r,c,\boldsymbol{x},y)] \\
&= \int_{\mathcal{Y}} p(y|\boldsymbol{x})\mathcal{L}(h,r,c,\boldsymbol{x},y)\mathrm{d}y
\end{aligned}$$

If we refuse to make a prediction for $\boldsymbol{x}$, i.e., $r(\boldsymbol{x}) < 0$, the above expected risk transforms into the following equation:

$$\begin{aligned}
R_{\mathrm{RcR|x,r(x)<0}}(h,r) &= \int_{\mathcal{Y}} p(y|\boldsymbol{x})\mathcal{L}(h,r,c,\boldsymbol{x},y)\mathrm{d}y \\
&= \int_{\mathcal{Y}} p(y|\boldsymbol{x})c(\boldsymbol{x})\mathrm{d}y \\
&= c(\boldsymbol{x}).
\end{aligned}$$

If we want to make a prediction for $\boldsymbol{x}$, i.e., $r(\boldsymbol{x}) > 0$, the above expected risk transforms into the following equation:

$$\begin{aligned}
R_{\mathrm{RcR|x,r(x)>0}}(h,r) &= \int_{\mathcal{Y}} p(y|\boldsymbol{x})\mathcal{L}(h,r,c,\boldsymbol{x},y)\mathrm{d}y \\
&= \int_{\mathcal{Y}} p(y|\boldsymbol{x})(h(\boldsymbol{x})-y)^2\mathrm{d}y \\
&= \int_{\mathcal{Y}} p(y|\boldsymbol{x})(h(\boldsymbol{x})^2-2yh(\boldsymbol{x})+y^2)\mathrm{d}y \\
&= \int_{\mathcal{Y}} p(y|\boldsymbol{x})h(\boldsymbol{x})^2\mathrm{d}y - \int_{\mathcal{Y}} p(y|\boldsymbol{x})2yh(\boldsymbol{x})\mathrm{d}y + \int_{\mathcal{Y}} p(y|\boldsymbol{x})y^2\mathrm{d}y \\
&= h^2(\boldsymbol{x}) - 2h(\boldsymbol{x})\mathbb{E}_{p(y|\boldsymbol{x})}[y] + \mathbb{E}_{p(y|\boldsymbol{x})}[y^2] \\
&= h^2(\boldsymbol{x}) - 2h(\boldsymbol{x})\mathbb{E}_{p(y|\boldsymbol{x})}[y] + \mathbb{E}^2_{p(y|\boldsymbol{x})}[y] + \mathbb{D}_{p(y|\boldsymbol{x})}[y] \\
&= (h(\boldsymbol{x}) - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 + \mathbb{D}_{p(y|\boldsymbol{x})}[y]
\end{aligned}$$

When $h(\boldsymbol{x}) = \mathbb{E}_{p(y|\boldsymbol{x})}[y]$ makes $R_{\mathrm{RcR|x,r(x)>0}} = \mathbb{D}_{p(y|\boldsymbol{x})}[y]$ minimum. It is easy to know that $R_{\mathrm{RcR|x,r(x)>0}} < R_{\mathrm{RcR|x,r(x)<0}}$ when $c(\boldsymbol{x}) - \mathbb{D}_{p(y|\boldsymbol{x})}[y] > 0$ and $R_{\mathrm{RcR|x,r(x)>0}} > R_{\mathrm{RcR|x,r(x)<0}}$ when $c(\boldsymbol{x}) - \mathbb{D}_{p(y|\boldsymbol{x})}[y] < 0$ which means that $R_{\mathrm{RcR|x}}$ is minimum when the following equation holds.

$$r^{\star}(\boldsymbol{x}) = \mathbb{I}(c(\boldsymbol{x}) - \mathbb{D}_{p(y|\boldsymbol{x})}[y]).$$

The proof is completed. $\square$

B Proofs of Consistent and Calibration

B.1 Proof of Theorem 4

First, we prove that the optimal regressor $h^\star$ is also the optimal regressor for R^{ψ}_{RcR} as follows.

$$
\begin{aligned}
&R^{\psi}_{\mathrm{RcR}}(h^\star, r)\\
&= \mathbb{E}_{p(\boldsymbol{x},y)}[\psi(h^\star, r, c, \boldsymbol{x}, y)]\\
&= \mathbb{E}_{p(\boldsymbol{x},y)}[(h^\star(\boldsymbol{x}) - y)^2 \ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\\
&= \mathbb{E}_{p(\boldsymbol{x},y)}[(h^\star(\boldsymbol{x})^2 - 2yh^\star(\boldsymbol{x}) + y^2)\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p(\boldsymbol{x}, y)[(h^\star(\boldsymbol{x})^2 - 2yh^\star(\boldsymbol{x}) + y^2)\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}y\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p(y|\boldsymbol{x})p(\boldsymbol{x})[(h^\star(\boldsymbol{x})^2 - 2yh^\star(\boldsymbol{x}) + y^2)\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}y\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}} p(\boldsymbol{x})[(h^\star(\boldsymbol{x})^2 - \int_{\mathcal{Y}} 2yh^\star(\boldsymbol{x})p(y|\boldsymbol{x})\mathrm{d}y + \int_{\mathcal{Y}} y^2 p(y|\boldsymbol{x})\mathrm{d}y)\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}} p(\boldsymbol{x})[(h^\star(\boldsymbol{x})^2 - 2h^\star(\boldsymbol{x})\mathbb{E}_{p(y|\boldsymbol{x})}[y] + \mathbb{E}_{p(y|\boldsymbol{x})}[y^2])\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}} p(\boldsymbol{x})[(h^\star(\boldsymbol{x})^2 - 2h^\star(\boldsymbol{x})\mathbb{E}_{p(y|\boldsymbol{x})}[y] + \mathbb{E}^2_{p(y|\boldsymbol{x})}[y] + \mathbb{D}_{p(y|\boldsymbol{x})}[y])\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}} p(\boldsymbol{x})[((h^\star(\boldsymbol{x})^2 - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 + \mathbb{D}_{p(y|\boldsymbol{x})}[y])\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]\mathrm{d}\boldsymbol{x}\\
&= \int_{\mathcal{X}} \mathbb{D}_{p(y|\boldsymbol{x})}[y]\ell(r(\boldsymbol{x}), -1)p(\boldsymbol{x}) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}
\tag{11}
$$

When $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(\boldsymbol{x}, z) \geq 0$, the risk loss is minimal for an any rejector r. Therefore $h^\star$ is the optimal regressor for risk R^{ψ}_{RcR}. On the other hand, we prove that $h^\star$ is the only optimal regressor if condition: $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(r(\boldsymbol{x}), z) > 0$ is achieved.

Suppose given an instance $\boldsymbol{x}_0$ and a rejector r' such that $\ell(r'(\boldsymbol{x}_0), -1) = 0$. Then we have at least one other regressor h' such that $R^{\psi}_{\mathrm{RcR}}(h', r') = R^{\psi}_{\mathrm{RcR}}(h^\star, r')$ and $h'(\boldsymbol{x}_0) \neq h^\star(\boldsymbol{x}_0)$ due to the following equation holds.

$$
\mathbb{D}_{p(y|\boldsymbol{x}_0)}[y]\ell(r'(\boldsymbol{x}_0), -1) = 0. \tag{12}
$$

Therefore when condition: $\forall \boldsymbol{x} \in \mathcal{X}$, $\ell(r(\boldsymbol{x}), z) > 0$ is achieved, there is one, and only one minimizer of R^{ψ}_{RcR}, which is the same as the optimal regressor $h^\star$. The proof is completed. □

B.2 Proof of Theorem 5

Fixing the regressor h, it is easy to see that the conditional optimal r should have the same sign with $\mathbb{E}_{p(y|\boldsymbol{x})}[(h(\boldsymbol{x})-y)^2]-c(\boldsymbol{x})$ due to the definition of classification calibrated binary loss. Then it is easy to see the rejector calibration holds when $\mathbb{D}_{p(y|\boldsymbol{x})}[y] \geq c(\boldsymbol{x})$ since $\mathbb{E}_{p(y|\boldsymbol{x})}[(h(\boldsymbol{x}) - y)^2] \geq \mathbb{D}_{p(y|\boldsymbol{x})}[y]$. When $\mathbb{D}_{p(y|\boldsymbol{x})}[y] < c(\boldsymbol{x})$, it is easy to show that

$$
\begin{aligned}
&\min_{r(\boldsymbol{x})}(\mathbb{E}_{p(y|\boldsymbol{x})}[(h'(\boldsymbol{x}) - y)^2]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1))\\
\geq &\min_{r(\boldsymbol{x})}(\mathbb{E}_{p(y|\boldsymbol{x})}[(h''(\boldsymbol{x}) - y)^2]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1))
\end{aligned}
$$

when the expected square loss of h' is larger than h''. Furthermore, when $h'(\boldsymbol{x})$'s expected square loss is equal to $c(\boldsymbol{x})$, it can be learned from the property of binary classification calibrated losses that

$$
\begin{aligned}
&\min_{r(\boldsymbol{x})}(\mathbb{E}_{p(y|\boldsymbol{x})}[(h'(\boldsymbol{x}) - y)^2]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1))\\
> &\min_{r(\boldsymbol{x})}(\mathbb{E}_{p(y|\boldsymbol{x})}[(h''(\boldsymbol{x}) - y)^2]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)),
\end{aligned}
$$

and thus we can conclude that the optimal h must have expected square loss that is lower than $c(\boldsymbol{x})$. Then the optimal rejector must have a negative sign, which is the same as the Bayes optimal one. Combining the conclusions above and we can complete the proof. $\square$

B.3 Proof of Theorem 6

We suppose that for any classification calibrated binary classification loss function $\ell(r(\boldsymbol{x}), z)$, when $\ell(r(\boldsymbol{x}), -1) = 0$, $r(\boldsymbol{x}) < 0$, i.e. the classification is correct. Let us go back to the discussion of Eq. (11):

$$\begin{aligned}R^{\psi}_{\mathrm{RwR}}(h^{\star}, r) &= \int_{\mathcal{X}} [((h^{\star}(\boldsymbol{x})^2 - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 + \mathbb{D}_{p(y|\boldsymbol{x})}[y])\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{X}} \mathbb{D}_{p(y|\boldsymbol{x})}[y]\ell(r(\boldsymbol{x}), -1)p(\boldsymbol{x}) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Similar to the proof of Theorem 2, the optimal regressor $h^{\star}$ still minimizes risk loss for any rejector. However, it is easy to know that for a rejector r_0 when there exists an instance $\boldsymbol{x}_0$ such that $\ell(r'(\boldsymbol{x}_0), -1) = 0$, there exists at least one other regressor h' such that $R^{\psi}_{\mathrm{RcR}}(h', r') = R^{\psi}_{\mathrm{RcR}}(h^{\star}, r')$ and $h'(\boldsymbol{x}_0) \neq h^{\star}(\boldsymbol{x}_0)$ due $((h(\boldsymbol{x})^2 - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 + \mathbb{D}_{p(y|\boldsymbol{x})}[y])\ell(r(\boldsymbol{x}), -1) = 0$ holds. Therefore the optimal regressor $h^{\star}$ is not the only optimal solution. Fortunately, we can still show that it is regressor-consistent for some instances in this case.

For a binary classification loss function ℓ, We denote by $\mathcal{X}^1_{\ell}$ the space where $\forall \boldsymbol{x} \in \mathcal{X}^1_{\ell}$, $\ell(r(\boldsymbol{x}), -1) \neq 0$ for any rejector r. Then we have the following equation:

$$\begin{aligned}R^{\psi}_{\mathrm{RwR}}(h, r) &= \int_{\mathcal{X}} [((h(\boldsymbol{x}) - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2 + \mathbb{D}_{p(y|\boldsymbol{x})}[y])\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{X}^1_{\ell}} (h(\boldsymbol{x}) - \mathbb{E}_{p(y|\boldsymbol{x})}[y])^2\ell(r(\boldsymbol{x}), -1)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &+ \int_{\mathcal{X}^1_{\ell}} \mathbb{D}_{p(y|\boldsymbol{x})}[y]\ell(r(\boldsymbol{x}), -1)p(\boldsymbol{x}) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

When for all $\boldsymbol{x} \in \mathcal{X}^1_{\ell}$, we have The above risk is minimised when $h(\boldsymbol{x}) = \mathbb{E}_{p(y|\boldsymbol{x})}[y]$ for all $\boldsymbol{x} \in \mathcal{X}^1_{\ell}$. It is worth noting that when $\ell(r(\boldsymbol{x}), -1) = 0$, $r(\boldsymbol{x}) < 0$, so the rejector remains consistent. The proof is completed. $\square$

C Proofs of Regret Transfer and Estimation Error Bound

C.1 Proof of Theorem 7

Proof. For each point $\boldsymbol{x}$, we can learn that its excess risk can be decomposed below if the model misrejects a sample:

$$\begin{aligned}&\mathbb{E}_{p(y|\boldsymbol{x})}[(h(\boldsymbol{x}) - y)^2]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1) - (h^{\star}(\boldsymbol{x}) - y)^2\ell(r^{\star}(\boldsymbol{x}), -1) - c(\boldsymbol{x})\ell(r^{\star}(\boldsymbol{x}), +1) \\ &\geq D_{p(y|\boldsymbol{x})}[y]\ell(r(\boldsymbol{x}), -1) + c(\boldsymbol{x})\ell(r(\boldsymbol{x}), +1) - D_{p(y|\boldsymbol{x})}[y]\ell(r^{\star}(\boldsymbol{x}), -1) - c(\boldsymbol{x})\ell(r^{\star}(\boldsymbol{x}), +1) \\ &\geq \left(D_{p(y|\boldsymbol{x})}[y] + c(\boldsymbol{x})\right)\xi^{-1}\left(\frac{c(\boldsymbol{x}) - D_{p(y|\boldsymbol{x})}[y]}{(D_{p(y|\boldsymbol{x})}[y] + c(\boldsymbol{x}))}\right)\end{aligned}$$

When a sample is correctly accepted, the lower bound is

$$\left[\mathbb{E}_{p(y|\boldsymbol{x})}[(h(\boldsymbol{x}) - y)^2] - D_{p(y|\boldsymbol{x})}[y]\right]\alpha,$$

where $\alpha = \min_{r(\boldsymbol{x} \leq 0)} \ell(r(\boldsymbol{x}), -1)$ and when it is misaccepted:

$$\left(D_{p(y|\boldsymbol{x})}[y] + c(\boldsymbol{x})\right)\xi^{-1}\left(\frac{D_{p(y|\boldsymbol{x})}[y] - c(\boldsymbol{x})}{(D_{p(y|\boldsymbol{x})}[y] + c(\boldsymbol{x}))}\right) + \left[\mathbb{E}_{p(y|\boldsymbol{x})}[(h(\boldsymbol{x}) - y)^2] - D_{p(y|\boldsymbol{x})}[y]\right]\alpha,$$

When sigmoid loss is used, $\xi(u) = u$, and we can learn that $\alpha = 1/2$, then we can learn that the excess risk of the surrogate at this point is larger can upper bound that of the original loss. When logistic loss is used, $\xi(u) = \sqrt{u}$ and $\alpha = \log 2$, we can use the same method to show that the excess risk of the surrogate can bound the square root of the excess risk of original loss, which concludes the proof.

□

C.2 Proof of Theorem 8

Definition 9. *(Rademacher complexity)* Let $Z_1, \cdots, Z_n$ be n *i.i.d.* random variables drawn from a probability distribution μ and $\mathcal{F} = \{f : Z \to \mathbb{R}\}$ be a class of measurable functions. Then the expected Rademacher complexity of function class $\mathcal{F}$ is given as follow:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{Z_1,\cdots,Z_n\sim\mu}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right], \tag{13}$$

where $\sigma_1, \cdots, \sigma_n$ are the Rademacher variables that take the value from $\{-1, +1\}$ uniformly.

Then we can begin proving Theorem 8.

Proof. Suppose that the loss is bounded by M_1 and ρ-Lipschitz continuous, $|h|$, $c(\boldsymbol{x})$, and $|y|$ is bounded by M_2, then we can learn that the loss is bounded by $C = (4M_2^2 + M_2)M_1$, and is L_1-Lipschitz continuous *w.r.t.* (h, r), where $L_1 = \sqrt{(4M_1^2\rho + M_1\rho)^2 + 16M_1^4M_2^2}$. By applying the McDiarmid's inequality, it is routine to show that the following inequalities hold with probability at least $1 - \frac{\delta}{2}$, respectively:

$$\sup_{h,r\in\mathcal{H},\mathcal{R}}\left(R^{\psi}_{\mathrm{RwR}}(h,r) - \hat{R}^{\psi}_{\mathrm{RwR}}(h,r)\right) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{h,r\in\mathcal{H},\mathcal{R}}\left(R^{\psi}_{\mathrm{RwR}}(h,r) - \hat{R}^{\psi}_{\mathrm{RwR}}(h,r)\right)\right] + C\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

$$\sup_{h,r\in\mathcal{H},\mathcal{R}}\left(\hat{R}^{\psi}_{\mathrm{RwR}}(h,r) - R^{\psi}_{\mathrm{RwR}}(h,r)\right) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{h,r\in\mathcal{H},\mathcal{R}}\left(\hat{R}^{\psi}_{\mathrm{RwR}}(h,r) - R^{\psi}_{\mathrm{RwR}}(h,r)\right)\right] + C\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

By applying Talagrand's contraction lemma [31], we can learn that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{h,r\in\mathcal{H},\mathcal{R}}\left(\hat{R}^{\psi}_{\mathrm{RwR}}(h,r) - R^{\psi}_{\mathrm{RwR}}(h,r)\right)\right] \leq \sqrt{2}L_1(\mathfrak{R}_n(\mathcal{H}) + \mathfrak{R}_n(\mathcal{R}))$$

and this conclusion also holds for another direction. Plugging this conclusion into the former inequalities and using the union bound, we can learn this inequality holds with probability at least $1 - \delta$:

$$\sup_{h,r\in\mathcal{H},\mathcal{R}}\left|\hat{R}^{\psi}_{\mathrm{RwR}}(h,r) - R^{\psi}_{\mathrm{RwR}}(h,r)\right| \leq \sqrt{2}L_1(\mathfrak{R}_n(\mathcal{H}) + \mathfrak{R}_n(\mathcal{R})) + C\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

According to the definition of empirical risk minimization and identifiable condition, we can get the following conclusion:

$$\begin{aligned}
&R^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r}) - \min_{h,r\in\mathcal{H},\mathcal{R}} R^{\psi}_{\mathrm{RwR}}(h,r) = R^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r}) - R^{\psi *}_{\mathrm{RwR}}(h^*,r^*) \\
&= \left(R^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r}) - \hat{R}^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r})\right) + \left(\hat{R}^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r}) - \hat{R}^{\psi}_{\mathrm{RwR}}(h^*,r^*)\right) + \left(\hat{R}^{\psi}_{\mathrm{RwR}}(h^*,r^*) - R^{\psi *}_{\mathrm{RwR}}(h^*,r^*)\right) \\
&\leq \left(R^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r}) - \hat{R}^{\psi}_{\mathrm{RwR}}(\hat{h},\hat{r})\right) + \left(\hat{R}^{\psi}_{\mathrm{RwR}}(h^*,r^*) - R^{\psi *}_{\mathrm{RwR}}(h^*,r^*)\right) \\
&\leq 2\sup_{h,r\in\mathcal{H},\mathcal{R}}\left|\hat{R}^{\psi}_{\mathrm{RwR}}(h,r) - R^{\psi}_{\mathrm{RwR}}(h,r)\right|
\end{aligned}$$

combining Theorem 5 and we can conclude the proof. □

Table 5: Test performance (mean and std) of our surrogate loss equipped hinge loss on BreastPathQ. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100 and Sup, RcRLoss, AL and RL are all magnified by a factor of 1000.

Cost	Sup	RR	AL	RL	Rej	AR	RA
5		4.74 (0.38)	3.51 (1.96)	53.05 (20.33)	80.86 (4.17)	60.37 (6.70)	4.19 (2.36)
10		8.32 (0.21)	4.58 (1.74)	58.90 (13.15)	68.99 (4.71)	46.63 (6.45)	5.91 (2.54)
15	16.77 (1.22)	11.89 (0.31)	6.44 (1.82)	49.42 (8.12)	62.45 (4.76)	46.86 (5.12)	10.69 (3.84)
20		15.07 (0.33)	9.53 (1.15)	49.58 (8.10)	52.33 (3.94)	38.04 (3.47)	17.88 (5.36)
25		16.54 (0.78)	10.36 (2.39)	58.39 (19.85)	41.23 (8.36)	29.71 (7.36)	25.34 (12.36)

Table 6: Test performance (mean and std) of our surrogate loss equipped huber loss on AgeDB. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100.

Cost	Sup	RR	AL	RL	Rej	AR	RA
60		59.99 (0.10)	44.80 (13.97)	177.36 (40.19)	97.30 (2.16)	95.84 (3.19)	1.43 (1.17)
70		70.24 (0.50)	71.81 (4.61)	185.68 (26.75)	92.41 (1.20)	88.90 (1.75)	4.32 (0.74)
80		79.67 (1.40)	76.43 (12.86)	185.14 (18.51)	87.23 (2.08)	82.63 (2.63)	7.83 (1.88)
90	100.34 (3.73)	88.71 (1.08)	76.78 (8.93)	166.84 (5.90)	83.43 (11.01)	79.46 (12.07)	11.19 (9.13)
100		96.95 (0.67)	77.02 (7.46)	182.70 (13.20)	84.78 (6.77)	80.39 (7.29)	9.14 (5.49)
110		104.29 (0.31)	85.84 (6.98)	192.05 (26.75)	73.52 (10.39)	67.73 (10.33)	17.41 (9.56)
120		111.54 (2.23)	92.59 (4.79)	186.50 (13.73)	67.31 (11.60)	61.11 (11.56)	21.74 (10.43)

D Additional Information of Experiments

D.1 Evaluation Metrics

We describe in detail all the evaluation metrics we used in our experiments.

RcR loss. The RcR loss (RcRloss) is the main evaluation metric for RcR. For a given example $(\boldsymbol{x}, y)$ and rejection cost $c(\boldsymbol{x})$, the RcR loss defined as if $r(\boldsymbol{x}) > 0$, $\mathcal{L}(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2$, otherwise $\mathcal{L}(h, r, c, \boldsymbol{x}, y) = c(\boldsymbol{x})$.

Rejection rate. The rejection rate (RR) is defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}) \leq 0]}{n}$. RR indicates the ratio of rejection of our model on the test dataset.

Accepted loss. The accepted loss (AL) is defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i) > 0](h(\boldsymbol{x}_i) - y_i)^2}{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i) > 0]}$. AL denotes the average loss of our regressor on the accepted test dataset.

Rejected loss. The rejected loss (RL) is defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i) \leq 0](h(\boldsymbol{x}_i) - y_i)^2}{\sum_{i=1}^{n} \mathbb{I}[r(\boldsymbol{x}_i) \leq 0]}$. RL denotes the average loss of our regressor on the rejected test dataset.

False rejection ratio. The false rejection ratio (AR) is defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i) - y_i)^2 < c(\boldsymbol{x}_i)]\mathbb{I}[r(\boldsymbol{x}_i) \leq 0]}{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i) - y_i)^2 < c(\boldsymbol{x}_i)]}$. AR denotes the ratio of instances that should be accepted that are rejected.

False acceptance ratio. The false acceptance ratio (RA) denotes the ratio of instances that should be rejected that are accepted, and is defined as $\frac{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i) - y_i)^2 \geq c(\boldsymbol{x}_i)]\mathbb{I}[r(\boldsymbol{x}_i) > 0]}{\sum_{i=1}^{n} \mathbb{I}[(h(\boldsymbol{x}_i) - y_i)^2 \geq c(\boldsymbol{x}_i)]}$.

Table 7: Test performance (mean and std) of our surrogate loss equipped hinge loss on five UCI datasets trained with the MLP model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Supervised	RR	AL	RL	Rej	AR	RA
Abalone	3	4.44 (0.46)	2.38 (0.13)	1.89 (0.23)	9.01 (1.10)	46.59 (3.33)	37.18 (3.63)	28.54 (2.69)
	4		2.86 (0.13)	2.32 (0.21)	9.39 (1.24)	33.17 (2.87)	25.58 (3.23)	40.22 (2.75)
	5		3.21 (0.18)	2.61 (0.29)	9.78 (1.31)	26.30 (2.46)	19.52 (2.73)	45.19 (4.04)
	6		3.51 (0.32)	2.93 (0.47)	10.88 (1.39)	19.51 (3.28)	13.92 (3.11)	52.65 (6.12)
Airfoil	9	12.96 (2.60)	6.57 (0.24)	4.62 (0.53)	49.82 (5.98)	43.99 (4.98)	23.67 (4.17)	24.12 (6.29)
	12		7.73 (0.36)	5.50 (0.49)	67.45 (12.27)	34.09 (4.62)	17.15 (4.32)	29.99 (4.70)
	16		8.71 (0.54)	6.50 (0.61)	83.75 (14.26)	23.16 (3.91)	9.11 (3.41)	36.32 (5.25)
	20		9.71 (0.50)	7.21 (0.46)	85.55 (12.55)	19.44 (3.46)	7.73 (3.51)	38.29 (3.92)
	25		10.81 (0.59)	8.29 (1.15)	100.88 (15.42)	14.75 (4.51)	5.23 (3.43)	39.74 (12.43)
	30		11.49 (0.87)	8.73 (0.95)	102.39 (13.94)	12.86 (3.10)	4.74 (2.15)	38.79 (6.14)
Auto-mpg	4	8.34 (2.16)	3.67 (0.24)	2.75 (0.92)	12.89 (3.57)	64.74 (14.08)	55.24 (15.24)	23.23 (12.85)
	6		4.91 (0.82)	3.53 (1.73)	16.61 (5.28)	45.38 (20.47)	37.09 (20.75)	37.16 (22.28)
	8		7.18 (1.70)	6.57 (2.28)	26.85 (15.92)	23.72 (20.39)	20.58 (18.82)	66.51 (23.79)
	10		7.19 (1.68)	6.88 (1.85)	37.08 (24.48)	8.85 (3.41)	6.63 (2.24)	77.04 (13.17)
	13		8.11 (2.01)	7.74 (2.67)	33.34 (22.63)	6.79 (3.49)	5.44 (2.70)	80.38 (10.46)
Housing	9	12.57 (3.43)	10.05 (1.56)	9.63 (5.05)	37.68 (19.06)	61.58 (24.70)	55.68 (28.35)	30.18 (17.22)
	12		10.58 (2.54)	9.38 (3.92)	73.71 (53.32)	34.46 (25.18)	27.79 (26.17)	48.53 (21.27)
	16		10.34 (3.13)	9.56 (3.56)	118.43 (65.06)	10.56 (4.96)	7.10 (4.57)	72.29 (13.43)
	20		10.57 (3.07)	9.80 (3.46)	161.32 (122.64)	6.63 (3.91)	4.67 (3.60)	77.31 (14.72)
Concrete	20	34.44 (3.05)	18.18 (1.28)	14.89 (3.78)	136.33 (62.30)	59.13 (8.28)	40.79 (12.28)	19.17 (5.93)
	30		24.31 (1.59)	20.48 (3.04)	164.20 (54.64)	38.83 (6.48)	22.39 (5.30)	33.07 (8.57)
	40		28.46 (3.00)	24.26 (4.32)	212.30 (65.52)	26.07 (8.47)	11.65 (4.46)	43.60 (11.08)
	50		30.70 (3.49)	26.59 (4.32)	222.34 (60.20)	17.38 (6.32)	7.07 (3.46)	51.86 (8.93)
	60		35.56 (4.36)	32.32 (5.10)	215.29 (81.58)	11.70 (3.06)	5.48 (2.13)	64.22 (7.87)

D.2 Some Results for Hinge Loss

In this section, we show some experimental results of the surrogate loss function equipped with hinge loss, which can be formulated as follows:

$$\psi(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2 \max(0, 1 + r(\boldsymbol{x})) + c(\boldsymbol{x}) \max(0, 1 - r(\boldsymbol{x})).$$

Table 5, Table 6 and Table 7 show some of the experimental results on the AgeDB, BreastPathQ, and UCI datasets with MLP model equipped hinge loss, respectively. From this table, we can see that RcRloss and AL is always lower than Sup in almost all experiments, which means that our method is effective in identifying test instances should be accepted and test instances should be rejected. It is

Table 8: Test performance (mean and std) of our surrogate loss equipped logistic loss on five UCI datasets trained with the Linear model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Supervised	RR	AL	RL	Rej	AR	RA
Abalone	3	4.92 (0.51)	2.52 (0.08)	1.94 (0.21)	7.84 (1.06)	54.77 (2.66)	44.76 (3.45)	24.00 (1.93)
	4		2.99 (0.11)	2.39 (0.19)	9.83 (1.44)	36.93 (2.78)	27.94 (3.02)	36.55 (2.83)
	5		3.38 (0.18)	2.80 (0.25)	11.78 (1.86)	25.90 (2.27)	18.43 (2.08)	46.24 (3.20)
	6		3.69 (0.26)	3.19 (0.31)	13.81 (2.00)	17.80 (1.98)	11.89 (1.45)	55.08 (4.19)
Airfoil	9	23.32 (1.54)	8.83 (0.35)	7.55 (2.64)	26.44 (1.99)	85.58 (6.10)	78.07 (8.53)	7.24 (3.99)
	12		11.31 (0.49)	8.76 (2.18)	27.59 (2.06)	79.93 (3.65)	71.90 (5.57)	9.74 (1.97)
	16		14.46 (0.51)	10.90 (1.46)	30.60 (2.13)	69.47 (6.49)	60.88 (7.17)	17.14 (5.66)
	20		17.10 (0.83)	13.01 (1.79)	33.51 (4.17)	58.54 (5.07)	50.38 (5.67)	26.19 (6.54)
	25		19.62 (1.26)	15.95 (2.52)	35.26 (3.40)	39.30 (5.76)	34.28 (5.41)	47.29 (8.82)
	30		20.94 (1.84)	17.60 (3.18)	42.36 (7.32)	25.38 (8.07)	21.35 (6.97)	61.14 (12.70)
Auto-mpg	4	11.66 (2.26)	3.92 (0.26)	2.78 (1.68)	15.05 (4.12)	79.87 (11.10)	73.18 (14.13)	15.28 (7.25)
	6		5.73 (0.70)	5.25 (1.63)	17.98 (6.14)	58.97 (8.50)	52.23 (9.69)	32.06 (10.05)
	8		6.73 (0.52)	5.72 (0.92)	21.58 (7.67)	42.56 (7.84)	36.08 (8.66)	43.16 (11.24)
	10		7.37 (0.95)	5.94 (1.61)	26.05 (11.45)	31.28 (12.77)	25.72 (10.98)	53.96 (21.76)
	13		8.75 (1.64)	7.72 (1.94)	28.93 (12.64)	19.62 (4.69)	17.31 (4.60)	69.71 (13.77)
Housing	9	24.08 (5.34)	8.65 (0.75)	6.95 (3.16)	33.87 (12.62)	67.92 (11.51)	58.56 (14.04)	19.28 (7.94)
	12		10.27 (1.08)	8.19 (2.90)	40.93 (16.74)	58.32 (10.20)	48.25 (13.11)	23.32 (5.75)
	16		12.34 (1.14)	9.20 (2.03)	50.31 (19.14)	45.35 (6.04)	36.24 (6.00)	31.42 (6.77)
	20		14.19 (1.67)	10.48 (2.72)	55.08 (23.26)	38.42 (6.08)	32.22 (6.62)	42.45 (12.64)
Concrete	20	111.12 (8.01)	19.80 (0.29)	10.00 (6.44)	204.20 (63.34)	97.57 (2.04)	95.25 (4.29)	1.55 (0.86)
	30		29.51 (0.92)	24.08 (12.35)	227.13 (99.54)	91.17 (4.57)	87.60 (6.54)	6.02 (3.11)
	40		38.09 (1.38)	28.95 (7.18)	282.98 (96.63)	80.34 (6.60)	73.83 (7.46)	12.05 (6.39)
	50		46.94 (1.82)	34.22 (9.57)	242.13 (98.34)	75.34 (10.15)	68.94 (11.79)	15.98 (7.51)
	60		51.96 (2.08)	41.36 (5.76)	370.24 (113.24)	56.26 (2.61)	45.96 (2.24)	25.26 (4.63)

worth noting that in most experiments, there is a low RA, which means that there is a higher tendency to reject hard-to-predict test instances to avoid serious errors when equipping hinge loss.

D.3 Some Results for Logistic Loss

In this section, we show some experimental results of the surrogate loss function equipped with logistic loss, which can be formulated as follows:

$$\psi(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2 \log(1 + \exp(r(\boldsymbol{x}))) + c(\boldsymbol{x}) \log(1 + \exp(-r(\boldsymbol{x}))).$$

Table 10, Table 9 and Table 8 show some of the experimental results on the BreastPathQ, and UCI datasets with MLP model and Linear model equipped logistic loss, respectively. Our proposed method

Table 9: Test performance (mean and std) of our surrogate loss equipped logistic loss on five UCI datasets trained with the MLP model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Supervised	RR	AL	RL	Rej	AR	RA
Abalone	3	4.44 (0.46)	2.41 (0.10)	1.95 (0.18)	8.34 (0.87)	43.14 (2.84)	33.37 (3.34)	33.32 (2.81)
	4		2.87 (0.18)	2.33 (0.29)	9.68 (1.18)	32.29 (3.28)	24.28 (3.29)	41.96 (3.29)
	5		3.20 (0.20)	2.59 (0.29)	10.86 (1.34)	25.25 (2.15)	17.86 (2.02)	45.15 (3.61)
	6		3.48 (0.25)	2.82 (0.35)	11.54 (1.53)	20.44 (1.72)	14.74 (1.42)	51.48 (4.63)
Airfoil	9	12.96 (2.60)	6.78 (0.47)	5.05 (0.82)	51.45 (4.32)	43.68 (5.19)	20.35 (5.21)	23.75 (3.96)
	12		7.96 (0.64)	5.79 (0.82)	59.74 (3.53)	35.05 (3.02)	12.94 (2.79)	26.90 (3.03)
	16		9.04 (0.56)	7.46 (0.47)	68.20 (10.78)	18.57 (4.08)	7.36 (3.14)	48.00 (7.21)
	20		9.64 (0.74)	7.81 (0.44)	74.27 (10.19)	15.05 (7.76)	5.83 (1.91)	47.91 (11.27)
	25		10.47 (1.08)	8.50 (0.52)	80.28 (27.67)	12.03 (4.43)	3.78 (1.89)	49.00 (17.65)
	30		10.95 (1.11)	8.95 (0.78)	89.30 (31.58)	9.50 (3.86)	2.93 (1.10)	50.67 (18.37)
Auto-mpg	4	8.34 (2.16)	3.85 (0.56)	3.22 (1.51)	11.99 (3.81)	62.44 (10.72)	54.18 (9.53)	25.19 (13.29)
	6		5.33 (0.82)	4.67 (1.36)	15.08 (3.83)	43.01 (16.01)	35.19 (16.09)	41.25 (15.35)
	8		6.53 (1.18)	5.86 (1.53)	19.34 (7.60)	29.49 (13.45)	23.19 (13.60)	53.57 (14.78)
	10		7.06 (1.60)	6.42 (1.91)	21.71 (8.88)	17.95 (3.63)	14.15 (3.48)	65.59 (11.17)
	13		7.80 (1.90)	7.04 (2.09)	28.55 (14.96)	13.59 (5.35)	11.05 (5.32)	70.15 (14.83)
Housing	9	12.57 (3.43)	8.60 (2.49)	8.43 (3.15)	45.60 (27.02)	26.57 (5.78)	19.05 (5.63)	66.95 (10.39)
	12		9.50 (1.56)	8.63 (2.10)	63.52 (26.15)	25.44 (6.43)	17.38 (5.84)	52.68 (12.58)
	16		9.30 (1.37)	8.03 (1.70)	90.19 (38.08)	15.45 (4.30)	10.84 (3.87)	62.99 (9.05)
	20		9.67 (1.40)	8.33 (1.77)	103.55 (54.73)	11.18 (2.97)	8.71 (2.70)	70.06 (8.59)
Concrete	20	34.44 (3.05)	18.65 (1.41)	14.32 (4.80)	58.33 (15.88)	68.93 (13.47)	57.72 (17.08)	16.19 (9.45)
	30		25.64 (2.50)	23.16 (4.95)	80.43 (14.47)	32.85 (12.82)	19.30 (10.81)	48.23 (15.91)
	40		29.79 (2.33)	25.25 (3.55)	107.54 (22.18)	30.24 (8.58)	19.97 (7.91)	44.38 (9.87)
	50		31.63 (4.29)	25.79 (4.22)	120.02 (24.12)	24.22 (11.22)	15.29 (10.47)	47.42 (10.35)
	60		34.04 (4.48)	33.26 (4.55)	165.71 (62.18)	2.77 (5.13)	2.14 (2.93)	92.59 (12.85)

significantly outperforms the supervised regression method in almost all cases, which verifies the ability of our method to reject difficult test instances demonstrating the effectiveness of our method. In most cases, the average loss of our method in the accepted test instances (AL) is always smaller than the average loss of the supervised regression model (Sup) in all test instances. This further indicates the ability of our method to identify hard-to-predict samples and reject them. On both MLP and Linear models, our method is effective in avoiding serious errors, which verifies that our method can be adapted to different models.

Table 10: Test performance (mean and std) of our surrogate loss equipped logistic loss on BreastPathQ. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100 and Sup, RcRLoss, AL and RL are all magnified by a factor of 1000.

Cost	Sup	RcRloss	AL	RL	RR	AR	RA
5	16.77 (1.22)	4.41 (0.35)	2.92 (1.51)	35.59 (7.36)	71.56 (3.20)	47.48 (5.95)	5.71 (3.02)
10		7.99 (0.47)	4.72 (1.01)	41.34 (9.74)	61.72 (9.54)	40.69 (4.90)	10.30 (1.02)
15		11.52 (0.36)	7.98 (1.36)	42.67 (5.87)	50.48 (5.85)	35.50 (4.61)	20.18 (7.72)
20		13.69 (0.81)	8.92 (1.05)	63.51 (49.02)	43.65 (5.24)	31.11 (5.00)	22.61 (5.72)
25		16.64 (0.94)	12.96 (2.28)	35.27 (2.63)	29.84 (5.93)	23.63 (5.42)	44.77 (10.54)

Table 11: Test performance (mean and std) of our surrogate loss equipped square loss on BreastPathQ. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100 and Sup, RcRLoss, AL and RL are all magnified by a factor of 1000.

Cost	Sup	RcRloss	AL	RL	RR	AR	RA
5	16.77 (1.22)	4.67 (0.41)	3.60 (1.17)	36.70 (4.46)	69.23 (4.94)	44.09 (4.16)	6.26 (3.25)
10		8.13 (0.38)	5.30 (0.73)	43.66 (6.95)	59.69 (2.59)	37.47 (2.40)	10.42 (2.60)
15		12.02 (1.09)	8.83 (2.14)	39.83 (8.27)	51.70 (2.56)	36.78 (0.80)	17.65 (2.03)
20		14.58 (0.57)	9.66 (2.55)	43.69 (7.58)	44.72 (9.54)	33.20 (7.89)	24.21 (11.09)
25		15.75 (0.76)	11.98 (2.59)	45.65 (7.18)	27.57 (8.62)	19.94 (7.86)	43.73 (12.04)

D.4 Some Results for Square Loss

In this section, we show some experimental results of the surrogate loss function equipped with square loss, which can be formulated as follows:

$$\psi(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2(r(\boldsymbol{x}) + 1)^2 + c(\boldsymbol{x})(r(\boldsymbol{x}) - 1)^2.$$

Table 11 and Table 12 show some of the experimental results on the BreastPathQ, and UCI datasets with MLP model equipped square loss, respectively. When the rejection cost c is small, both RcRloss and AL are significantly smaller than Sup. When the rejection cost c is large, RcRloss and AL are close to Sup but always smaller, which shows the effectiveness of our method to deal with regression with cost-based rejection.

D.5 Some Results for Sigmoid

In this section, we show some experimental results of the Sigmoid function equipped with sigmoid, which can be formulated as follows:

$$\psi(h, r, c, \boldsymbol{x}, y) = (h(\boldsymbol{x}) - y)^2\mathrm{sigmoid}(r(\boldsymbol{x})) + c(\boldsymbol{x})\mathrm{sigmoid}(-r(\boldsymbol{x})).$$

Unlike other binary classification losses, sigmoid can be viewed as weight balancing prediction loss and rejection cost due to $\mathrm{sigmoid}(r(\boldsymbol{x})) + \mathrm{sigmoid}(-r(\boldsymbol{x})) = 1$. Table 13 and Table 14 show some of the experimental results on the BreastPathQ, and AgeDB equipped sigmoid, respectively. RcRloss and AL are always smaller than Sup, verifying the effectiveness of our method.

In our experiments, we used multiple binary classification losses (MAE, hinge loss, logistic loss, square loss and sigmoid) and different datasets including two deep datasets (BreastPathQ and

Table 12: Test performance (mean and std) of our surrogate loss equipped square loss on five UCI datasets trained with the MLP model. We repeat the sampling-and-training process 10 times. The metrics RR, AR, and RA are scaled to 0-100.

Datasets	Cost	Supervised	RR	AL	RL	Rej	AR	RA
Abalone	3	4.44 (0.46)	2.39 (0.10)	1.96 (0.19)	7.82 (0.63)	42.54 (2.49)	32.79 (2.69)	32.58 (2.63)
	4		2.84 (0.16)	2.33 (0.25)	8.79 (0.97)	31.82 (1.87)	23.72 (2.14)	40.81 (2.77)
	5		3.18 (0.18)	2.60 (0.27)	9.89 (1.15)	25.37 (2.13)	18.32 (2.07)	45.65 (4.21)
	6		3.50 (0.29)	2.89 (0.42)	10.40 (1.11)	20.38 (2.17)	14.37 (1.85)	49.83 (5.47)
Airfoil	9	12.96 (2.60)	6.40 (0.25)	4.36 (0.36)	51.93 (5.13)	43.65 (3.26)	22.05 (2.93)	22.22 (3.71)
	12		7.46 (0.31)	5.11 (0.38)	61.13 (5.83)	33.75 (2.90)	15.04 (3.10)	27.05 (2.74)
	16		8.57 (0.40)	5.81 (0.30)	70.20 (7.82)	26.98 (3.23)	10.54 (3.08)	28.83 (1.79)
	20		9.27 (0.42)	6.66 (0.43)	76.90 (10.02)	19.34 (2.34)	7.70 (1.54)	35.09 (4.65)
	25		9.97 (0.60)	7.23 (0.51)	87.37 (9.07)	15.35 (2.44)	5.19 (1.38)	32.96 (5.61)
	30		10.33 (0.86)	7.82 (0.79)	85.67 (18.03)	11.23 (1.95)	3.92 (1.25)	37.40 (8.27)
Auto-mpg	4	8.34 (2.16)	3.65 (0.26)	2.83 (0.90)	11.93 (3.22)	62.31 (11.46)	51.76 (12.43)	22.05 (10.30)
	6		5.19 (0.77)	4.31 (1.47)	18.00 (7.69)	39.62 (21.51)	33.51 (21.46)	47.55 (24.43)
	8		6.51 (1.35)	5.82 (1.74)	22.62 (8.96)	29.10 (14.57)	22.28 (13.86)	52.29 (16.21)
	10		6.80 (1.16)	6.08 (1.43)	23.57 (9.41)	17.82 (2.84)	13.91 (1.93)	65.62 (9.09)
	13		7.28 (1.30)	6.45 (1.36)	30.51 (15.46)	12.69 (3.74)	10.05 (3.46)	71.16 (15.70)
Housing	9	12.57 (3.43)	8.41 (1.56)	8.22 (2.10)	53.44 (20.25)	28.22 (7.81)	21.42 (7.35)	56.77 (9.38)
	12		9.03 (1.26)	8.36 (1.64)	76.10 (47.37)	17.13 (5.71)	12.16 (5.11)	66.75 (11.99)
	16		8.52 (1.35)	7.64 (1.62)	109.61 (60.72)	10.10 (3.67)	7.30 (2.71)	73.14 (13.36)
	20		9.40 (1.94)	8.56 (2.16)	148.19 (98.03)	7.03 (3.30)	5.09 (2.31)	73.76 (16.54)
Concrete	20	34.44 (3.05)	19.95 (2.56)	18.77 (5.05)	75.19 (11.68)	55.10 (13.77)	43.24 (14.20)	28.28 (11.92)
	30		25.22 (3.22)	22.44 (5.34)	103.99 (18.06)	33.45 (6.93)	22.25 (6.21)	43.75 (9.93)
	40		31.21 (1.50)	28.84 (1.84)	127.77 (19.65)	21.17 (5.11)	12.47 (4.12)	56.12 (7.44)
	50		29.55 (2.97)	25.00 (3.80)	147.99 (36.87)	18.01 (4.62)	9.87 (3.88)	52.49 (9.18)
	60		33.07 (3.51)	28.81 (3.94)	158.65 (33.36)	13.64 (3.71)	7.28 (2.91)	59.55 (8.52)

AgeDB) and five uci datasets (Abalone, Airfoil, Auto-mpg, Housing and Concrete), and our method outperformed supervised regression in most cases, which demonstrates the effective of our method.

E Limitations

In Theorem 4 and Theorem 5 we show that there is a limitation in our proposed method that requires the binary classification loss $\ell(r(\boldsymbol{x}), z)$ to be always greater than 0. This is easily satisfied by the design of the binary classification loss such as $\mathrm{logistic}(r(\boldsymbol{x}), z)$ to $\max(\alpha, \mathrm{logistic}(r(\boldsymbol{x}), z))$, where $\alpha > 0$ is the minimum value of loss. However, to avoid the modification of the binary classification loss, we further propose Theorem 6, which only requires the binary classification loss to be greater

Table 13: Test performance (mean and std) of our surrogate loss equipped sigmoid on BreastPathQ. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100 and Sup, RcRLoss, AL and RL are all magnified by a factor of 1000.

Cost	Sup	RR	AL	RL	Rej	AR	RA
5	16.77 (1.22)	4.42 (0.17)	2.40 (1.01)	43.86 (10.60)	79.22 (1.54)	56.41 (4.19)	2.81 (0.92)
10		8.44 (0.46)	5.09 (1.64)	51.35 (7.30)	69.34 (2.56)	47.12 (3.83)	6.39 (1.83)
15		11.63 (0.38)	6.30 (1.83)	58.40 (14.79)	60.23 (6.75)	41.23 (5.42)	10.88 (5.57)
20		14.31 (0.66)	8.91 (1.26)	57.83 (9.11)	49.19 (4.68)	33.84 (5.32)	17.31 (3.13)
25		16.61 (0.60)	9.09 (1.53)	95.10 (59.42)	47.38 (2.92)	29.46 (4.84)	17.78 (4.73)

Table 14: Test performance (mean and std) of our surrogate loss equipped sigmoid on AgeDB. We repeat the sampling-and-training process 5 times. The metrics RR, AR, RA are scaled to 0-100.

Cost	Sup	RR	AL	RL	Rej	AR	RA
60	100.34 (3.73)	60.20 (0.51)	60.82 (4.67)	129.91 (19.72)	88.93 (7.18)	85.88 (8.82)	7.40 (4.89)
70		69.10 (0.71)	61.91 (6.75)	136.57 (32.45)	83.85 (6.63)	79.98 (6.76)	10.74 (5.51)
80		78.33 (1.08)	64.22 (12.63)	131.88 (12.46)	80.01 (7.91)	76.58 (9.37)	12.74 (5.69)
90		84.47 (3.22)	73.65 (6.28)	134.06 (8.45)	68.11 (11.43)	63.56 (12.32)	22.91 (8.91)
100		88.52 (2.36)	75.22 (11.21)	140.65 (8.00)	61.67 (12.36)	55.55 (12.55)	25.22 (10.65)
110		94.16 (3.24)	83.70 (5.36)	156.32 (22.43)	36.81 (18.24)	32.37 (16.93)	52.43 (20.40)
120		99.69 (5.18)	90.63 (3.51)	158.91 (26.86)	28.43 (20.46)	25.67 (21.76)	64.54 (27.32)

than or equal to 0, and this is easily satisfied. Extensive experiments on various datasets demonstrate the effectiveness of our proposed method.