A. Related work

In this section, we provide more discussion on the relationship of our work with Zhang et al. (2018) and Mozannar & Sontag (2020).

A.1. Our work and Zhang et al. (2018)

Zhang et al. (2018) considers to tackle classification with rejection using angle-based classification approach.

Given $\boldsymbol{x}$ and y. Define

$$\mathbf{y}_y = \begin{cases} (K-1)^{-\frac{1}{2}}\mathbf{1}_{K-1} & y = 1, \\ -(1+K^{\frac{1}{2}})/\{(K-1)^{\frac{3}{2}}\}\mathbf{1}_{K-1} + \{K/(K-1)\}^{\frac{1}{2}}\mathbf{e}_{y-1} & 2 \le y \le K, \end{cases}$$

where $\mathbf{e}_{y-1}$ denotes the one-hot vector with one at the index $y-1$ and $\mathbf{1}_{K-1} \in R^{k-1}$ denotes a vector of all ones. Next, let $a \in \mathbb{R}$ be a positive scalar. Zhang et al. (2018) proposed the following bent hinge loss:

$$\mathcal{L}^{\text{hin}}_{\text{ANGLE}}(\boldsymbol{g}; \boldsymbol{x}, y) = \sum_{y' \neq y} \phi_{\text{hin}}\big(-\mathbf{y}_y^\top \boldsymbol{g}(\boldsymbol{x})\big),$$

where

$$\phi_{\text{hin}}(u) = \begin{cases} 1 - au & u < 0 \\ 1 - u & 0 \le u \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

and the bent distance weighted discrimination loss:

$$\mathcal{L}^{\text{dwd}}_{\text{ANGLE}}(\boldsymbol{g}; \boldsymbol{x}, y) = \sum_{y' \neq y} \phi_{\text{hin}}\big(-\mathbf{y}_y^\top \boldsymbol{g}(\boldsymbol{x})\big),$$

where

$$\phi_{\text{dwd}}(u) = \begin{cases} 1 - au & u < 0 \\ 1 - u & 0 \le u \le 0.5, \\ \frac{1}{4u} & \text{otherwise.} \end{cases}$$

Let $S_\delta(v) = \text{sign}(v)\max(|v| - \delta, 0)$. After training a classifier with their proposed loss function, the decision rule can be expressed as

$$f(\boldsymbol{x}; \boldsymbol{g}) = \begin{cases} \text{®} & \forall j \text{ s.t. } S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) = 0, \\ \arg\max_y \mathbf{y}_y^\top \boldsymbol{g}(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

In their rejection rule, an input $\boldsymbol{x}$ is rejected if all binary classifiers' outputs are close to zero. In our cost-sensitive approach, Cond. (7) rejects $\boldsymbol{x}$ as long as all $g_y(\boldsymbol{x})$'s are negative, e.g., $\boldsymbol{x}$ is also rejected if the all prediction outputs are much smaller than zero. Moreover, we do not have any hyperparameter in our rejection rule.

There are two hyperparameters that are needed to be tuned for the angle-based method, which are a bending slope a and a rejection threshold δ[2]. Zhang et al. (2018) defined the following quantities given the rejection cost c and the number of clases K:

$$a_1 = \frac{K - 1 - c}{Kc - c}, \quad a_2 = \frac{(K-1)(1-c)}{c}.$$

It was suggested that the choice of of a can be chosen by either $a = a_1$ or $a = a_2$. For δ, it needs to be tuned with the validation set with respect to the validation empirical zero-one-c risk.

[2]There is also a hyperparameter for determining the regularization strength, which is omitted here for brevity.

It can be observed that our approach and Zhang et al. (2018) are different. The modification of Zhang et al. (2018) is based on angle-based classification while it is based on the cost-sensitive one-vs-rest loss in our case. Moreover, no additional hyperparameter is introduced and we do not modify the loss by bending it to be steeper (for the hinge loss). Note that the proposed bending scheme of Zhang et al. (2018) will lead any loss to be positively unbounded, including the symmetric losses, which may cause them to lose their favorable properties. Also, it is not straightforward to design a bending scheme for any loss to the best of our knowledge. In our approach, any classification-calibrated loss can be straightforwardly applied in the surrogate loss in Definition 5. Finally, our Theorem 8 made it possible to transfer the excess risk bound from cost-sensitive classification (Steinwart, 2007; Scott, 2012) to classification with rejection. We are not aware of other works that discuss a theory that can explicitly show the excess risk bound relationship between cost-sensitive classification and cost-sensitive classification.

It is worth noting that Zhang et al. (2018) also considered a different setting called classification from refine options, where a classifier is also allowed to predict a set of labels instead of one label.

$$f(\boldsymbol{x};\boldsymbol{g}) = \begin{cases} ® & \forall j \text{ s.t. } S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) = 0, \\ \{j : S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) > 0\} & \exists j \text{ s.t. } S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) > 0, \\ \{j : S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) = 0\} & \text{otherwise.} \end{cases} \tag{14}$$

It can be observed that the second condition in Eq. (14), i.e., $\exists j$ s.t. $S_\delta(\mathbf{y}_j^\top \boldsymbol{g}(\boldsymbol{x})) > 0$ is similar to our Cond. (8), which occurs when a classifier wants to predict more than one classes. Nevertheless, in our problem, it is not allowed to predict a set and we propose to reject a data point in this scenario. It is also interesting to explore the problem of learning with refine options with our proposed cost-sensitive approach and we leave it for future work.

A.2. Our work and Mozannar & Sontag (2020)

Recently, Mozannar & Sontag (2020) has proposed a method for classification with rejection based on a reduction to cost-sensitive learning. However, the reduction scheme is different to ours since they proposed to augment a rejection class in the model and the loss choice is fixed to the cross-entropy loss. The main idea is to augment a rejection class $K+1$ in the score function $\boldsymbol{g}$. Rejection will be made if the maximum score of $\boldsymbol{g}$ is at index $K+1$. Given the rejection cost c and a score function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{K+1}$, the loss function proposed by Mozannar & Sontag (2020) can be expressed as

$$\mathcal{L}^c_{\mathrm{DEFER}}(\boldsymbol{g};\boldsymbol{x},y) = \log\left(\frac{\exp(g_y(\boldsymbol{x}))}{\sum_{j=1}^{K+1}\exp(g_j(\boldsymbol{x}))}\right) + (1-c)\log\left(\frac{\exp(g_{K+1}(\boldsymbol{x}))}{\sum_{j=1}^{K+1}\exp(g_j(\boldsymbol{x}))}\right).$$

It can be seen that our cost-sensitive approach gives a different form of loss function, that is, their loss function is not a special case of our cost-sensitive approach and vice versa. Moreover, the theoretical analysis in Mozannar & Sontag (2020) is based on analyzing this specific loss. Unlike our work, it may not be straightforward to borrow the theory of cost-sensitive classification (Steinwart, 2007; Scott, 2012) to justify the theoretical properties of a general surrogate loss function for classification with rejection in their approach. It is worth pointing out that one advantage of the loss function proposed by Mozannar & Sontag (2020) over our approach is that it is applicable to the situation where the rejection cost can be different for each $\boldsymbol{x}$ (see Mozannar & Sontag (2020) for more detail). Nevertheless, in our problem setting, the rejection cost is assumed to be a constant.

B. Proofs

In this section, we provide proofs for the theoretical results in the main body.

B.1. Proof of Proposition 4

Based on the following Chow's rule:

$$f^*(\boldsymbol{x}) = \begin{cases} ® & \max_y \eta_y(\boldsymbol{x}) \leq 1-c, \\ \arg\max_y \eta_y(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

It is straightforward to see that we can mimic the Chow's rule by only knowing whether

$$\eta_y(\boldsymbol{x}) \leq 1-c \quad \text{for all} \quad y \in \mathcal{Y}, \tag{15}$$

and

$$\arg\max_y \eta_y(\boldsymbol{x}).$$

This is because if Condition (15) is true, then a classifier refrains from making a prediction, otherwise a classifier predicts a class $\arg\max_y \eta_y(\boldsymbol{x})$, which matches Chow's rule.

To verify whether $\eta_y(\boldsymbol{x}) \leq 1-c$ for a class y, it suffices to learn a cost-sensitive binary classifier where $\alpha = 1-c$ to classify between a target class y and other classes (i.e., one-versus-rest classifier), as suggested in Definition 2. We define such optimal cost-sensitive binary classifier as $f^{*,y}_{1-c}$. As a result, we can construct K cost-sensitive binary classification problems where $\alpha = 1-c$ to verify Condition (15), that is, Condition (15) is true if and only if $f^{*,y}_{1-c} = -1$ for all $y \in \mathcal{Y}$.

Next, we show that if Condition (15) is false based on learning K cost-sensitive binary classifiers, then it is sufficient to verify $\arg\max_y \eta_y(\boldsymbol{x})$. This is because of the rejection cost c is less 0.5. Thus, if Condition (15) is false, the only possibility is that there must exists one y' such that $\eta'_y(\boldsymbol{x}) > 1-c$. The reason it can have at most one y' to have $f^{*,y'}_{1-c} = 1$ is because it indicates that $\eta'_y > 0.5$. Thus, the following rule can mimic Chow's rule.

$$f^*(\boldsymbol{x}) = \begin{cases} \text{®} & \max_y f^{*,y}_{1-c}(\boldsymbol{x}) = -1, \\ \arg\max_y f^{*,y}_{1-c}(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

This concludes the proof.

Remark: we note that if Condition (15) is true, it may not be possible to know $\arg\max_y \eta_y(\boldsymbol{x})$ given K optimal cost-sensitive binary classifiers in general. However, in classification with rejection, it is not important to know $\arg\max_y \eta_y(\boldsymbol{x})$ if Condition (15) is true since a classifier will refrain from making a prediction.

B.2. Proof of Theorem 7

Let g^* be a conditional risk minimizer that minimizes the pointwise conditional surrogate risk:

$$g^*(\boldsymbol{x}) = \arg\min_g W_{\mathcal{L}^{c,\phi}_{\text{CS}}}\big(\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x}))\big)$$

Recall that the cost-sensitive surrogate loss is defined as

$$\mathcal{L}^{c,\phi}_{\text{CS}}(\boldsymbol{g}; \boldsymbol{x}, y) = c\phi\big(g_y(\boldsymbol{x})\big) + (1-c)\sum_{y' \neq y} \phi\big(-g_{y'}(\boldsymbol{x})\big).$$

Thus,

$$\begin{aligned} W_{\mathcal{L}^{c,\phi}_{\text{CS}}}\big(\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x}))\big) &= \sum_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) \mathcal{L}^{c,\phi}_{\text{CS}}\big(\boldsymbol{g}; \boldsymbol{x}, y\big) \\ &= \sum_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) \left[c\phi\big(g_y(\boldsymbol{x})\big) + (1-c)\sum_{y' \neq y} \phi\big(-g_{y'}(\boldsymbol{x})\big) \right]. \end{aligned} \tag{16}$$

We can rewrite Eq. (16) as follows based on the perspective of g_y:

$$\begin{aligned} W_{\mathcal{L}^{c,\phi}_{\text{CS}}}\big(\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x}))\big) &= \sum_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) \left[c\phi\big(g_y(\boldsymbol{x})\big) + (1-c)\sum_{y' \neq y} \phi\big(-g_{y'}(\boldsymbol{x})\big) \right]. \\ &= \sum_{y \in \mathcal{Y}} \left[\eta_y(\boldsymbol{x}) c\phi\big(g_y(\boldsymbol{x})\big) + (1-\eta_y(\boldsymbol{x}))(1-c)\phi\big(-g_y(\boldsymbol{x})\big) \right]. \end{aligned}$$

It can be seen that $\eta_y(\boldsymbol{x})c\phi\big(g_y(\boldsymbol{x})\big) + (1-\eta_y(\boldsymbol{x}))(1-c)\phi\big(-g_y(\boldsymbol{x})\big)$ is a pointwise conditional risk of a cost-sensitive binary classifier g_y. Thus, minimizing $W_{\mathcal{L}_{\mathrm{CS}}^{c,\phi}}$ can be viewed as independently minimizing the pointwise conditional risk in cost-sensitive binary classification for each g_y. Thus g_y^* corresponds to the conditional risk minimizer of the cost-sensitive binary classification where y is a positive class and $y' \neq y$ is a negative class.

Recall the definition of $f(\boldsymbol{x}; \boldsymbol{g}^*)$:

$$f(\boldsymbol{x}; \boldsymbol{g}^*) = \begin{cases} ® & \max_y g_y^*(\boldsymbol{x}) \leq 0, \\ ® & \exists y, y' \text{ s.t. } y \neq y' \\ & \text{and } g_y^*(\boldsymbol{x}), g_{y'}^*(\boldsymbol{x}) > 0, \\ \arg\max_y g_y^*(\boldsymbol{x}) & \text{otherwise} \end{cases}$$

and $f^*(\boldsymbol{x})$:

$$f^*(\boldsymbol{x}) = \begin{cases} ® & \max_y \eta_y(\boldsymbol{x}) \leq 1-c, \\ \arg\max_y \eta_y(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

First, we prove that $f(\boldsymbol{x}; \boldsymbol{g}^*) = f^*(\boldsymbol{x})$ if ϕ is classification-calibrated.

The proof is based on the definition of α-classification calibration (α-CC) proposed by (Scott, 2012) and its relationship with ordinary classification calibration (Bartlett et al., 2006), which is equivalent to 0.5-CC. More specifically, it is known that a margin loss ϕ must also be classification-calibrated if it is α-CC, i.e., its conditional risk minimizer matches the Bayes optimal classifier of cost-sensitive binary classification when using the weighted risk minimization based on α (Scott, 2012).

For a classification-calibrated margin loss ϕ, $\mathrm{sign}(g_y^*)$ matches the Bayes optimal solution of the cost-sensitive binary classification (Scott, 2012), that is, $g_y^*(\boldsymbol{x}) > 0$ if $\eta_y(\boldsymbol{x}) > 1-c$ and $g_y^*(\boldsymbol{x}) < 0$ otherwise. Thus if g^* is obtained, the condition

$$\exists y, y' \text{ s.t. } y \neq y' \text{and } g_y^*(\boldsymbol{x}), g_{y'}^*(\boldsymbol{x}) > 0 \tag{17}$$

is impossible to occur since it is impossible to have $g_y^*(\boldsymbol{x}) > 0$ and $g_{y'}^* > 0$ simultaneously (because that can occur only if $1-c < 0.5$ which is impossible if $c < 0.5$.). Thus, it suffices to look at

$$f(\boldsymbol{x}; \boldsymbol{g}^*) = \begin{cases} ® & \max_y g_y^*(\boldsymbol{x}) \leq 0, \\ \arg\max_y g_y^*(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

$\max_y g_y^*(\boldsymbol{x}) \leq 0$ indicates that $\eta_y(\boldsymbol{x}) \leq 1-c$ for all $y \in \mathcal{Y}$ and thus coincides with the rejection criterion of Chow's rule. On the other hand, if $\max_y g_y^*(\boldsymbol{x}) > 0$, then there exists only one y such that $g_y^*(\boldsymbol{x}) > 0$ and $\eta_y(\boldsymbol{x}) > 1-c$. Thus, $f(\boldsymbol{x}; \boldsymbol{g}^*) = f^*(\boldsymbol{x})$ if ϕ is classification-calibrated.

Next, we prove the converse of our theorem, that is, if a margin loss ϕ is not classification-calibrated, then there must exist the case where $\boldsymbol{g}^{*\mathrm{w}}$ disagrees with Chow's rule, where $g^{*\mathrm{w}}$ denotes an optimal solution for ϕ that is not classification-calibrated.

Here, let us drop $\boldsymbol{x}$ and only concerns $\boldsymbol{\eta}$, which is a probability simplex for simplicity, which suffices to prove our statement. If a margin loss ϕ is not classification-calibrated, all $\mathrm{sign}(g_y^{*\mathrm{w}})$ does not match the Bayes-optimal solution of the cost-sensitive classifier with respect to the rejection cost c, which suggests that there exists $\boldsymbol{\eta}$ that makes at least one $g_y^{*\mathrm{w}}(\boldsymbol{x})$ has the wrong sign compared with $g_y^*(\boldsymbol{x})$.

We divide our analysis of $\boldsymbol{\eta}$ into two cases. First, we analyze the case where Chow's rule suggests to predict the input with the most probable class, which is the case where $\max_y \boldsymbol{\eta}_y > 1-c$. Second, we analyze the case where Chow's rule suggests to refrain from making a prediction, i.e., $\max_y \boldsymbol{\eta}_y \leq 1-c$. Note that both cases cover all possibilities of $\boldsymbol{\eta}$. Moreover, we note that if $f(\boldsymbol{x}; g^{*\mathrm{w}})$ disagrees with Chow's rule in at least one of the cases, it suffices to prove that $g^{*\mathrm{w}}$ does not achieve calibration in classification with rejection.

Case 1: $\max_y \eta_y > 1-c$

In this case, Chow's rule suggests to accept and predict the most probable class $\arg\max_y \eta_y$.

It is straightforward to see that for a decision rule $f(\boldsymbol{x}; \boldsymbol{g})$ to match Chow's rule in this case, the sign of all g_y must match the Bayes optimal classifier of binary cost-sensitive classification g_y^*, that is $g_y > 0$ only for $\eta_y > 1-c$ and $g_y' < 0$ for other

less probable classes. Based on $f(\boldsymbol{x}; \boldsymbol{g})$, this is the only possible configuration of $\boldsymbol{g}$ to have the same decision as Chow's rule in the case where $\max_y \eta_y > 1 - c$, which is predicting the most probable class $\arg\max_y \eta_y$.

Thus, if the disagreement between of $\boldsymbol{g}^{*\mathrm{w}}$ and $\boldsymbol{g}^*$ arises in the case where $\max_y \eta_y > 1 - c$, $\boldsymbol{g}^{*\mathrm{w}}$ must lead either predicting the wrong class (i.e., not the most probable class) or refrain from making a prediction, which both cases disagree with Chow's rule and lead to higher zero-one-c risk. Thus, it suffices to show that if ϕ is not classification-calibrated and the disagreement occurs when $\max_y \eta_y > 1 - c$, then $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is not calibrated.

Case 2: $\max_y \boldsymbol{\eta}_y \leq 1 - c$

In this case, Chow's rule suggests to refrain from making a prediction.

We will show that if $\boldsymbol{g}^{*\mathrm{w}}$ agrees with Chow's rule for all $\boldsymbol{\eta}$ that lie in Case 1, there must exist $\boldsymbol{\eta}$ in Case 2 such that the decision of $\boldsymbol{g}^{*w}$ disagrees with Chow's rule if no further restriction such as the number of classes is imposed.

If $f(\boldsymbol{x}; \boldsymbol{g}^{*w}) = f^*(\boldsymbol{x})$ everywhere in Case 1, then it is guaranteed that $g_y^{*w} < 0$ for $\eta_y < c$, that is, g_y^{*w} must predict negative correctly when $\max_y \eta_y > 1 - c$ and only one $g_y^{*w} > 0$ for $\eta_y > 1 - c$ for all $y \in \mathcal{Y}$. We will make use of these conditions to show that $f(\boldsymbol{x}; \boldsymbol{g}^{*w})$ will wrongly accept the data while Chow's rule suggests to refrain from making a prediction.

In this case where $\max_y \boldsymbol{\eta}_y \leq 1 - c$, there must exist $\boldsymbol{\eta}$ such that at least one y, we have $g_y > 0$ although $\eta_y \leq 1 - c$ to make the sign of g^{*w} disagrees with g^*. Let $\beta \leq 1 - c$ be a value that makes $g_y > 0$. There exists $\boldsymbol{\eta}$ such that $\eta_y = \beta$ for one y and $\eta_{y'} \leq c$ for all other classes. Since $c > 0$, it is always possible to make $\sum_{y \in \mathcal{Y}} \eta_y = 1$ with a sufficient number of classes. Thus, we have $g_y > 0$ where $\eta_y = \beta$ and $g_{y'} < 0$ for all other classes. This makes the decision of $f(\boldsymbol{x}; \boldsymbol{g}^{*w})$ to accept and predict the data while Chow's rule suggests to refrain from making a prediction, which means $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is not calibrated.

In summary, if the disagreement occurs in Case 1 , it suffices to say $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is not calibrated. If disagreement does not occur in Case 1, there exists $\boldsymbol{\eta}$ that makes $f(\boldsymbol{x}; \boldsymbol{g}^{*w})$ disagrees with Chow's rule. This concludes the proof of the converse case of our theorem.

As a result, the surrogate loss $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is calibrated for classification with rejection, that is, $f(\boldsymbol{x}; \boldsymbol{g}^*) = f^*(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, if and only if ϕ is classification-calibrated. This concludes the proof.

B.3. Proof of Theorem 8

To prove that

$$R^{\ell_{01c}}(f) - R^{\ell_{01c},*} \leq \sum_{i=1}^{K} \psi_{\phi,1-c}^{-1}(R_{1-c}^{\phi,i}(g_i) - R_{1-c}^{\phi,i,*}),$$

we divide the proof into two steps. The first step is to prove that

$$R^{\ell_{01c}}(f) - R^{\ell_{01c},*} \leq R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*} \tag{18}$$

and the second step is to prove that

$$R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*} \leq \sum_{i=1}^{K} \psi_{\phi,1-c}^{-1}(R_{1-c}^{\phi,i}(g_i) - R_{1-c}^{\phi,i,*}). \tag{19}$$

Proof of Ineq. (18):

To prove this inequality, it suffices to prove that $R^{\ell_{01c}}(f) \leq R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})$ and $R^{\ell_{01c},*} = R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}$.

To prove that $R^{\ell_{01c}}(f) \leq R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})$, it suffices to show that

$$\ell_{01c}(f(\boldsymbol{x}; \boldsymbol{g}), y) \leq \mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}; \boldsymbol{x}, y)$$

holds for any choices of $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. Thanks to the discrete nature of both the zero-one-c loss ℓ_{01c} and the zero-one loss ℓ_{01}, case analysis can be applied.

Case 1: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=0$

In this case, it suggests that $f(\boldsymbol{x};\boldsymbol{g})$ predicts a label that matches a label y. This is possible only if $g_y>0$ and $g_{y'}<0$ for $y'\neq y$. Recall the definition of the cost-sensitive surrogate loss:

$$\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=c\ell_{01}\big(g_y(\boldsymbol{x})\big)+(1-c)\sum_{y'\neq y}\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big). \tag{20}$$

It can be seen that $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}$ can only be larger or equal to zero, that is , $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)\geq 0$. Thus, $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)\leq \mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)$ if $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=0$. Nevertheless, we can show that they are in fact equal, i.e., $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=0$. This is because $c\ell_{01}\big(g_y(\boldsymbol{x})\big)=0$ and $(1-c)\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big)=0$. Thus, $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=0$ according to Eq. (20). Therefore, we have

$$\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=0.$$

Case 2: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=1$ In this case, it can be shown that

$$\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=1.$$

It can be seen that $f(\boldsymbol{x};\boldsymbol{g})$ wrongly predicts the label, which only occurs when $g_{y'}(\boldsymbol{x})>0$ where $y'\neq y$ and $g_{y''}(\boldsymbol{x})<0$ for $y''\neq y'$, which also includes the correct label. Therefore, $c\ell_{01}\big(g_y(\boldsymbol{x})\big)=c$, $(1-c)\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big)=1-c$, and $(1-c)\ell_{01}\big(-g_{y''}(\boldsymbol{x})\big)=0$ in the case where $y''\neq y$ and $y''\neq y'$. As a result, the sum of the penalty becomes $c+(1-c)=1$, which makes $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=1$.

Case 3: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=c$

Unlike the previous two cases, the bound can be loose in the case where $f(\boldsymbol{x};\boldsymbol{g})$ decides to refrain from making a prediction. $f(\boldsymbol{x};\boldsymbol{g})=c$ indicates that a decision rule decides to reject. This is possible only if $g_y^*(\boldsymbol{x})<0$ for all $y\in\mathcal{Y}$ or Condition (17) holds.

If $g_y^*(\boldsymbol{x})<0$ for all $y\in\mathcal{Y}$, then $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)=c$ because $c\ell_{01}\big(g_y(\boldsymbol{x})\big)=c$ and $(1-c)\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big)=0$.

If Condition (17) is true, we can show that $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)\geq c$. We will show by using the fact that the minimum possible value of $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)$ when having a conflict is $1-c$. This is when there exists $g_y(\boldsymbol{x})>0$ and $g_{y'}(\boldsymbol{x})>0$, where y is a correct label and y' is a wrong label. In this case, $c\ell_{01}\big(g_y(\boldsymbol{x})\big)=0$, $(1-c)\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big)=1-c$ and $(1-c)\ell_{01}\big(-g_{y''}(\boldsymbol{x})\big)=0$ for $y''\neq y'$ and $y''\neq y$. If the conflicts of only two classifiers occur and both give the wrong labels, then we the penalty is $2(1-c)+c$. More conflicts only gain the higher penalty or nothing (if the conflict comes from the correct class), thus $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)\geq 1-c\geq c$.

Therefore, $R^{\ell_{01c}}(f)\leq R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})$.

Next, to prove that $R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}=R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}$, it suffices to show that $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)$ for any choices of $\boldsymbol{x}\in\mathcal{X}$ and $y\in\mathcal{Y}$.

Case 1: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*),y)=0$

From the previous analysis, in this case we have

$$\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=0,$$

for any $\boldsymbol{g}$. Therefore, it must also hold when $\boldsymbol{g}=\boldsymbol{g}^*$.

Case 2: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*),y)=1$

From the previous analysis, in this case we have

$$\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g};\boldsymbol{x},y)=1,$$

for any $\boldsymbol{g}$. Therefore, it must also hold when $\boldsymbol{g}=\boldsymbol{g}^*$.

Case 3: $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*),y)=c$ As suggested in the proof of Theorem 7 that Condition (17) is impossible to occur for $f(\boldsymbol{x};\boldsymbol{g}^*)$. Thus, if $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*)$ rejects, it means that $g_y^*(\boldsymbol{x})<0$ for all $y\in\mathcal{Y}$. This makes $\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)=c$ because $c\ell_{01}\big(g_y(\boldsymbol{x})\big)=c$ and $(1-c)\ell_{01}\big(-g_{y'}(\boldsymbol{x})\big)=0$.

Since $\ell_{01c}(f(\boldsymbol{x};\boldsymbol{g}^*),y)=\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}(\boldsymbol{g}^*;\boldsymbol{x},y)$ always holds, $R^{\ell_{01c},*}=R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}$.

We have proven that $R^{\ell_{01c}}(f)\leq R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})$ and $R^{\ell_{01c},*}=R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}$. Thus, Ineq. (18) holds.

Proof of Ineq. (19):

In this part, the proof no longer involves with classification with rejection but the well-studied cost-sensitive classification. We borrow the existing result by Scott (2012) to prove this part.

Recall the following pointwise conditional risk:

$$\begin{aligned}W_{\mathcal{L}_{\mathrm{CS}}^{c,\phi}}\big(\boldsymbol{g}(\boldsymbol{x});\boldsymbol{\eta}(\boldsymbol{x})\big)&=\sum_{y\in\mathcal{Y}}\eta_y(\boldsymbol{x})\left[c\phi\big(g_y(\boldsymbol{x})\big)+(1-c)\sum_{y'\neq y}\phi\big(-g_{y'}(\boldsymbol{x})\big)\right].\\&=\sum_{y\in\mathcal{Y}}\left[\eta_y(\boldsymbol{x})c\phi\big(g_y(\boldsymbol{x})\big)+(1-\eta_y(\boldsymbol{x}))(1-c)\phi\big(-g_y(\boldsymbol{x})\big)\right]\end{aligned}$$

also holds when $\phi=\ell_{01}$. This suggests that the risk of the cost-sensitive surrogate equals to the sum of the pointwise surrogate risks of K cost-sensitive binary classification problem, i.e., $R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})=\sum_{i=1}^K R_{1-c}^{\ell_{01},i}(g_i)$ for any $\boldsymbol{g}$. Thus, we have

$$R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g})-R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}=\sum_{i=1}^K R_{1-c}^{\ell_{01},i}(g_i)-R_{1-c}^{\ell_{01},i,*}.$$

Next, since ϕ is classification-calibrated, Scott (2012) proved that there exists $\psi_{\phi,1-c}\colon\mathbb{R}\to\mathbb{R}$, which is a non-decreasing invertible function and $\psi_{\phi,1-c}(0)=0$ such that

$$R_{1-c}^{\ell_{01},i}(g_i)-R_{1-c}^{\ell_{01},i,*}\leq\psi_{\phi,1-c}^{-1}(R_{1-c}^{\phi,i}(g_i)-R_{1-c}^{\phi,i,*})$$

By adding the excess risk bound of K cost-sensitive binary classification problems, we have

$$R^{\ell_{01c}}(f)-R^{\ell_{01c},*}\leq\sum_{i=1}^K\psi_{\phi,1-c}^{-1}(R_{1-c}^{\phi,i}(g_i)-R_{1-c}^{\phi,i,*}).$$

Thus, Ineq. (19) holds.

Since Ineq. (18) and Ineq. (19) hold, we concludes the proof of the excess risk bound of classification with rejection based on cost-sensitive classification for a general classification-calibrated loss.

Table 2. Specification of benchmark datasets: the number of features, the number of classes, the number of data.

Name	#features	#classes	#data
Gisette	5000	2	7000
Phishing	30	2	11050
Spambase	57	2	4601
Subj	100	2	10000
Twonorm	20	2	7400
Gas-Drift	128	6	13910
HAR	561	6	10299
MNIST	28×28	10	70000
Fashion-MNIST	28×28	10	70000
KMNIST	28×28	10	70000

C. Experiment Details

In this section, we provide more information on datasets and implementation details.

C.1. Datasets

We used the train and test data for MNIST, Fashion-MNIST, and KMNIST for training and testing, respectively. For MNIST, Fashion-MNIST, and KMNIST, we used the same test data as the one provided for testing for those datasets. For hyperparameter selection, we split ten percent of the training data to use as validation data (i.e., 6000). For the other ninety percent of training data, they were used for training all methods. Note that only ANGLE and SCE used validation data.

For the datasets other than MNIST, Fashion-MNIST, and KMNIST, we randomly used half of the dataset for training all methods. For clean-labeled and noisy-labeled classification, we used ten percent of all data for validation and the other fourty percent were used for testing. For PU-classification, we used twenty percent of all data for validation and the other thirty percent were used for testing. Again, note that only ANGLE and SCE used validation data.

Table 2 shows the specification of the benchmark datasets used in this paper. Subj was preprocessed by using 100-dimensional GloVe mean word embedding (Pennington et al., 2014).

C.2. Implementation details

We used a linear-in-input model for all binary classification datasets. For MNIST, Fashion-MNIST, KMNIST, we used the same convolutional neural network (CNN) architecture. The CNN model consists of a sequence of two convolutional layers with 32 channels and two convolutional layers with 64 channels, followed by a max pooling layer and two linear layers with dimension 128. The kernel size of convolutional layers is 3, and the kernel size of max pooling layer is 2. Dropout with probability 0.5 is used between two linear layers. We used rectifier linear units (ReLU) (Nair & Hinton, 2010) as the non-linear activation function. For HAR and Gas-Drift, we used the one hidden layer multilayer perceptron as a model ($d - 64 - 1$). We also applied batch normalization (Ioffe & Szegedy, 2015) at the final layer to stabilize training. The objective functions were optimized using Adam (Kingma & Ba, 2015). The experiment code for implementing a model was written using PyTorch (Paszke et al., 2019). We ran 10 trials for each experiment.

C.2.1. Clean-labeled and noisy-labeled classification

Data generation process For noise labels, the noise rate was 0.25, i.e., 25% of labels are randomly flipped into other classes. No data augmentation was used for all experiments.

Hyperparameters For all experiments, learning rate was set to 0.001, batch size was 256. The model was trained for 10 epochs for the convolutional neural networks and 100 epochs for both the linear-in-parameter model and multilayer perceptron.

C.2.2. PU-CLASSIFICATION

Problem setting PU-classification considers situations when only positive and unlabeled data are available. We denote class conditional densities by $p_+(\boldsymbol{x}) = p(\boldsymbol{x}|y=+1)$ and $p_-(\boldsymbol{x}) = p(\boldsymbol{x}|y=-1)$ and the class prior probability by $\pi = p(y=+1)$. Then the distribution for unlabeled data can be expressed as

$$p(\boldsymbol{x}) = \pi p_+(\boldsymbol{x}) + (1-\pi)p_-(\boldsymbol{x}).$$

Let $\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, y)$ be a loss function. It is known that the expected classification risk can be expressed as (du Plessis et al., 2015):

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, y)] = \pi \mathbb{E}_{\boldsymbol{x}\sim p_+(\boldsymbol{x})}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, +1)] - \pi \mathbb{E}_{\boldsymbol{x}\sim p_+(\boldsymbol{x})}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)] + \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)].$$

Then, the unbiased risk estimator given positive examples $\{\boldsymbol{x}_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \overset{\text{i.i.d.}}{\sim} p_+(\boldsymbol{x})$ and unlabeled examples $\{\boldsymbol{x}_j^{\mathrm{u}}\}_{j=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$ can be expressed as

$$\frac{\pi}{n_{\mathrm{p}}}\sum_{i=1}^{n_{\mathrm{p}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, +1)] - \frac{\pi}{n_{\mathrm{p}}}\sum_{i=1}^{n_{\mathrm{p}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)] + \frac{1}{n_{\mathrm{u}}}\sum_{j=1}^{n_{\mathrm{u}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)]. \tag{21}$$

However, Kiryo et al. (2017) suggested that Eq. (21) is prone to overfitting because $\boldsymbol{g}$ may treat all unlabeled data as negative to minimize the empirical risk. As a result, it was suggested to instead minimize the following empirical risk:

$$\hat{R}_{\mathrm{PU}}^{\mathcal{L}}(\boldsymbol{g}) = \frac{\pi}{n_{\mathrm{p}}}\sum_{i=1}^{n_{\mathrm{p}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, +1)] + \max\left(0, \frac{1}{n_{\mathrm{u}}}\sum_{j=1}^{n_{\mathrm{u}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)] - \frac{\pi}{n_{\mathrm{p}}}\sum_{i=1}^{n_{\mathrm{p}}}[\mathcal{L}(\boldsymbol{g}; \boldsymbol{x}, -1)]\right), \tag{22}$$

which is known to be a biased but still consistent estimator. With Eq. (22), the choice of loss functions is flexible and we can easily apply any loss function to learn a classifier with rejection, e.g., our cost-sensitive approach can be easily applied in PU-classification by minimizing $\hat{R}_{\mathrm{PU}}^{\mathcal{L}_{\mathrm{CS}}^{c,\phi}}(\boldsymbol{g})$, which equals to solving two cost-sensitive PU-classification. We refer the readers to du Plessis et al. (2014; 2015); Kiryo et al. (2017) for more detail about PU-classification and Charoenphakdee & Sugiyama (2019) for more detail about cost-sensitive PU-classification.

Data generation process PU-classification needs two set of data: positive data and unlabeled data. We first decided the size of unlabeled data to be around the size of original training data, but truncated to be divisible by 200. Then, the size of positive data was set to be $\frac{1}{5}$ of the size of unlabeled data.

After deciding the size of two sets, we than sampled from the original training data. The positive data for PU is sampled from the original positive training data without replacement. Then, from the left positive training data and negative data, we sampled without replacement for unlabeled data according to the value of class prior.

Hyperparameters The class prior for the positive class is set to be 0.7 throughout all PU-classification experiments. Learning rate was set to 0.001, batch size was 64, and the number of epochs was 100.

D. Additional Experiment Results

In this section, we report full experimental results in a table format with varying rejection costs for clean-labeled classification, noisy-labeled classification, and PU-classification, respectively. We also provide more discussion on our proposed rejection conditions, i.e., Conds. (7) and (8).

D.1. Full experimental results in a table format

We report the mean and standard error over ten trials of the test empirical zero-one-c risk, rejection ratio, and test error on the non-rejected data.

TABLE INDEX:

- Tables 3, 4 and 5: Test empirical 0-1-c risk, classification error on accepted data, and rejection ratio for clean-labeled binary classification with rejection, respectively.
- Tables 6, 7 and 8: Test empirical 0-1-c risk, classification error on accepted data, and rejection ratio for noisy-labeled binary classification with rejection, respectively.
- Tables 9, 10 and 11: Test empirical 0-1-c risk, classification error on accepted data, and rejection ratio for positive-unlabeled binary classification with rejection, respectively.
- Tables 12, 13 and 14: Test empirical 0-1-c risk, classification error on accepted data, and rejection ratio for clean-labeled multiclass classification with rejection, respectively.
- Tables 15, 16 and 17: Test empirical 0-1-c risk, classification error on accepted data, and rejection ratio for noisy-labeled multiclass classification with rejection, respectively.

Table 3. Mean and standard error of 0-1-c risk of the clean-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	2.20(0.10)	5.33(0.16)	2.41(0.16)	2.47(0.14)	2.35(0.09)
	0.15	2.72(0.10)	6.41(0.34)	2.94(0.32)	2.69(0.13)	2.65(0.11)
	0.20	3.10(0.09)	7.43(0.22)	3.25(0.23)	2.82(0.13)	2.78(0.09)
	0.25	3.36(0.16)	8.53(0.25)	3.59(0.43)	3.04(0.10)	3.09(0.16)
	0.30	3.72(0.20)	9.68(0.30)	3.69(0.22)	3.19(0.16)	3.15(0.11)
	0.35	3.95(0.30)	10.46(0.19)	3.79(0.24)	3.41(0.16)	3.25(0.07)
	0.40	4.27(0.17)	11.31(0.29)	3.70(0.13)	3.52(0.15)	3.38(0.21)
Phishing	0.10	4.27(0.31)	5.49(0.02)	4.02(0.27)	3.59(0.09)	4.28(0.15)
	0.15	5.70(0.31)	7.53(0.05)	5.18(0.09)	4.84(0.04)	5.37(0.13)
	0.20	6.47(0.09)	9.26(0.06)	5.88(0.16)	5.73(0.06)	6.34(0.09)
	0.25	7.24(0.38)	10.57(0.08)	6.25(0.07)	6.12(0.07)	7.17(0.14)
	0.30	7.82(0.35)	11.32(0.03)	6.78(0.22)	6.64(0.11)	7.72(0.13)
	0.35	8.41(0.58)	11.72(0.06)	7.13(0.15)	7.07(0.07)	7.99(0.23)
	0.40	9.10(0.75)	12.01(0.06)	7.41(0.20)	7.30(0.04)	8.30(0.13)
Spambase	0.10	5.84(0.24)	7.06(0.08)	6.00(0.95)	5.61(0.14)	6.78(0.21)
	0.15	7.08(0.35)	8.93(0.08)	7.07(0.46)	6.95(0.21)	7.99(0.16)
	0.20	8.49(0.44)	10.28(0.19)	7.92(0.68)	7.79(0.21)	8.82(0.25)
	0.25	9.43(0.65)	11.32(0.24)	8.42(0.22)	8.34(0.14)	9.00(0.32)
	0.30	9.87(0.78)	12.07(0.35)	8.98(0.26)	8.55(0.14)	9.13(0.21)
	0.35	10.92(1.16)	12.43(0.40)	9.21(0.28)	8.79(0.22)	9.40(0.20)
	0.40	11.58(1.12)	12.04(0.29)	9.15(0.26)	9.01(0.17)	9.64(0.16)
Subj	0.10	7.02(0.19)	6.96(0.08)	6.50(0.33)	6.40(0.08)	8.17(0.10)
	0.15	8.39(0.26)	9.56(0.04)	8.11(0.23)	7.97(0.07)	9.21(0.13)
	0.20	9.71(0.21)	11.70(0.06)	9.25(0.07)	9.27(0.09)	9.98(0.11)
	0.25	10.96(0.20)	13.46(0.07)	10.35(0.36)	10.26(0.11)	10.60(0.06)
	0.30	12.07(0.17)	14.92(0.09)	10.96(0.15)	10.86(0.08)	11.13(0.08)
	0.35	13.09(0.19)	15.91(0.07)	11.38(0.20)	11.33(0.10)	11.48(0.07)
	0.40	13.59(0.27)	16.35(0.09)	11.61(0.11)	11.61(0.09)	11.70(0.08)
Twonorm	0.10	1.43(0.06)	5.10(0.12)	1.30(0.05)	1.34(0.07)	1.36(0.06)
	0.15	1.86(0.10)	7.03(0.18)	1.73(0.22)	1.62(0.07)	1.58(0.06)
	0.20	2.14(0.12)	8.51(0.13)	1.81(0.12)	1.82(0.05)	1.82(0.08)
	0.25	2.47(0.22)	9.93(0.18)	2.05(0.08)	1.99(0.05)	2.01(0.10)
	0.30	2.55(0.22)	10.63(0.26)	2.17(0.13)	2.08(0.06)	2.11(0.07)
	0.35	2.67(0.19)	11.13(0.11)	2.38(0.15)	2.23(0.09)	2.26(0.08)
	0.40	2.77(0.24)	11.08(0.17)	2.41(0.15)	2.26(0.06)	2.34(0.09)

Table 4. Mean and standard error of 0-1 risk for accepted data of the clean-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	1.34(0.19)	3.80(0.24)	1.34(0.23)	2.16(0.16)	2.10(0.09)
	0.15	1.60(0.33)	3.58(0.41)	1.73(0.68)	2.21(0.15)	2.32(0.13)
	0.20	1.38(0.06)	3.48(0.26)	2.16(0.67)	2.20(0.13)	2.38(0.10)
	0.25	1.67(0.33)	3.43(0.33)	2.35(0.83)	2.27(0.12)	2.66(0.18)
	0.30	1.59(0.28)	3.56(0.30)	2.48(0.63)	2.29(0.18)	2.70(0.12)
	0.35	1.81(0.37)	3.39(0.19)	3.29(0.51)	2.32(0.11)	2.81(0.07)
	0.40	1.79(0.30)	3.19(0.13)	2.48(0.47)	2.38(0.14)	2.97(0.18)
Phishing	0.10	2.75(0.45)	0.33(0.05)	2.23(0.79)	1.26(0.23)	2.73(0.29)
	0.15	3.00(0.88)	0.43(0.03)	2.57(1.14)	2.31(0.10)	3.03(0.21)
	0.20	3.40(0.56)	0.72(0.04)	4.03(0.95)	3.37(0.13)	3.66(0.15)
	0.25	3.56(0.46)	1.05(0.06)	4.26(0.58)	3.96(0.11)	4.32(0.08)
	0.30	4.40(0.81)	1.51(0.07)	5.08(1.05)	4.80(0.25)	4.89(0.11)
	0.35	3.78(0.86)	1.89(0.07)	5.61(1.14)	5.78(0.14)	5.47(0.21)
	0.40	4.72(0.84)	2.55(0.08)	5.94(1.19)	6.44(0.04)	6.44(0.18)
Spambase	0.10	3.97(0.73)	4.67(0.09)	4.96(1.54)	2.87(0.46)	5.46(0.36)
	0.15	4.48(0.55)	4.90(0.15)	5.87(1.11)	3.63(0.40)	5.65(0.27)
	0.20	4.39(0.73)	5.31(0.23)	6.02(1.92)	4.87(0.29)	6.51(0.33)
	0.25	4.41(0.82)	5.70(0.27)	7.15(1.11)	5.79(0.19)	6.56(0.25)
	0.30	5.26(0.90)	6.16(0.27)	8.34(0.83)	6.14(0.20)	6.94(0.31)
	0.35	5.31(0.73)	6.55(0.22)	8.67(0.80)	6.95(0.14)	7.53(0.25)
	0.40	5.13(0.66)	6.99(0.26)	7.93(1.07)	7.61(0.22)	8.31(0.15)
Subj	0.10	5.77(0.33)	2.72(0.17)	4.56(1.01)	4.78(0.14)	7.87(0.12)
	0.15	5.76(0.45)	3.15(0.10)	5.58(1.02)	5.76(0.11)	8.40(0.18)
	0.20	6.37(0.31)	3.43(0.07)	6.26(0.71)	6.77(0.16)	8.78(0.14)
	0.25	6.22(0.39)	3.78(0.08)	7.18(1.47)	7.65(0.15)	9.13(0.08)
	0.30	6.73(0.38)	4.25(0.11)	8.70(0.92)	8.30(0.11)	9.56(0.12)
	0.35	7.25(0.21)	4.98(0.08)	9.72(0.86)	9.05(0.14)	9.91(0.10)
	0.40	7.59(0.35)	5.58(0.17)	9.85(0.93)	9.87(0.09)	10.33(0.07)
Twonorm	0.10	0.82(0.21)	0.00(0.00)	0.68(0.20)	0.81(0.12)	0.97(0.10)
	0.15	0.94(0.23)	0.00(0.00)	0.98(0.58)	0.96(0.15)	1.03(0.14)
	0.20	1.03(0.32)	0.01(0.02)	0.98(0.30)	1.04(0.12)	1.13(0.09)
	0.25	1.11(0.22)	0.02(0.02)	1.35(0.42)	1.20(0.12)	1.16(0.16)
	0.30	1.00(0.14)	0.03(0.02)	1.73(0.52)	1.31(0.07)	1.25(0.13)
	0.35	1.24(0.24)	0.04(0.01)	1.53(0.57)	1.57(0.13)	1.46(0.11)
	0.40	1.33(0.21)	0.05(0.02)	1.74(0.45)	1.70(0.09)	1.67(0.14)

Table 5. Mean and standard error of rejection ratio of the clean-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	9.88(1.63)	24.67(0.66)	12.25(3.40)	3.97(0.34)	3.22(0.15)
	0.15	8.32(2.52)	24.78(0.70)	8.93(3.72)	3.73(0.16)	2.63(0.16)
	0.20	9.20(0.28)	23.91(0.59)	5.97(3.66)	3.51(0.28)	2.26(0.14)
	0.25	7.21(1.87)	23.63(0.69)	5.37(3.74)	3.39(0.24)	1.90(0.21)
	0.30	7.50(1.36)	23.12(0.39)	4.37(2.37)	3.24(0.17)	1.65(0.16)
	0.35	6.45(1.81)	22.36(0.71)	1.57(1.34)	3.34(0.36)	1.37(0.13)
	0.40	6.48(1.02)	22.08(0.88)	3.24(1.24)	3.03(0.18)	1.13(0.13)
Phishing	0.10	20.96(2.29)	53.33(0.27)	22.53(5.75)	26.60(1.52)	21.26(1.37)
	0.15	22.15(4.73)	48.74(0.33)	20.37(7.23)	19.94(0.54)	19.57(0.63)
	0.20	18.36(2.79)	44.30(0.32)	11.30(5.28)	14.21(0.58)	16.43(0.77)
	0.25	17.14(3.03)	39.75(0.33)	9.54(2.57)	10.29(0.34)	13.82(0.73)
	0.30	13.28(2.69)	34.44(0.20)	6.71(3.36)	7.26(0.52)	11.28(0.41)
	0.35	14.71(4.00)	29.69(0.22)	5.05(3.38)	4.42(0.31)	8.51(0.45)
	0.40	12.33(3.81)	25.26(0.22)	4.19(3.66)	2.57(0.14)	5.56(0.30)
Spambase	0.10	30.11(8.00)	44.83(1.27)	17.92(9.93)	38.36(2.72)	29.17(1.39)
	0.15	24.47(5.80)	39.94(0.93)	12.34(6.33)	29.12(1.98)	25.08(1.61)
	0.20	26.01(6.07)	33.83(1.26)	12.27(10.00)	19.31(1.51)	17.12(1.13)
	0.25	24.18(5.81)	29.15(1.22)	6.82(4.84)	13.25(0.64)	13.19(0.94)
	0.30	18.44(5.74)	24.81(1.14)	2.86(2.56)	10.09(0.71)	9.49(0.62)
	0.35	18.75(5.78)	20.68(1.40)	2.00(2.02)	6.56(0.47)	6.81(0.62)
	0.40	18.41(4.40)	15.29(0.71)	3.69(3.53)	4.33(0.43)	4.18(0.30)
Subj	0.10	29.46(1.47)	58.17(0.59)	34.30(8.36)	31.08(0.73)	14.22(0.33)
	0.15	28.39(1.45)	54.10(0.37)	26.19(6.17)	23.90(0.44)	12.21(0.47)
	0.20	24.50(1.23)	49.89(0.24)	21.56(3.69)	18.89(0.53)	10.67(0.38)
	0.25	25.24(1.41)	45.62(0.31)	17.27(6.22)	15.04(0.20)	9.26(0.25)
	0.30	22.92(1.36)	41.41(0.25)	10.45(3.31)	11.81(0.21)	7.67(0.26)
	0.35	21.03(0.61)	36.42(0.32)	6.48(2.96)	8.79(0.27)	6.28(0.26)
	0.40	18.51(1.44)	31.28(0.35)	5.75(2.82)	5.79(0.19)	4.62(0.12)
Twonorm	0.10	6.57(1.62)	50.96(1.17)	6.59(2.19)	5.73(0.71)	4.28(0.61)
	0.15	6.48(1.56)	46.87(1.18)	5.21(2.96)	4.68(0.62)	3.94(0.64)
	0.20	5.82(1.49)	42.55(0.62)	4.37(1.74)	4.12(0.48)	3.66(0.52)
	0.25	5.69(1.52)	39.66(0.75)	2.92(1.60)	3.30(0.33)	3.59(0.40)
	0.30	5.34(0.95)	35.36(0.85)	1.50(1.54)	2.68(0.23)	3.01(0.42)
	0.35	4.23(1.20)	31.72(0.33)	2.50(1.82)	1.97(0.25)	2.38(0.26)
	0.40	3.72(1.08)	27.62(0.43)	1.73(1.41)	1.48(0.14)	1.74(0.20)

Table 6. Mean and standard error of 0-1-c risk of the noisy-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	10.05(0.28)	21.29(0.70)	10.03(0.14)	23.98(0.54)	7.64(0.64)
	0.15	14.40(0.45)	22.97(0.42)	14.89(0.61)	26.03(0.54)	9.09(0.60)
	0.20	18.24(0.46)	24.36(0.95)	19.03(0.33)	27.23(0.69)	11.26(0.64)
	0.25	22.30(0.81)	26.03(0.74)	22.60(0.82)	27.76(0.93)	12.76(0.76)
	0.30	25.39(0.50)	27.75(0.62)	25.80(0.81)	28.47(0.80)	13.52(1.06)
	0.35	28.26(1.12)	29.18(0.63)	27.69(0.84)	29.00(0.79)	14.91(0.70)
	0.40	31.01(1.00)	31.39(0.66)	28.78(1.07)	29.72(0.81)	15.88(0.68)
Phishing	0.10	6.54(0.41)	9.12(0.17)	9.29(1.18)	10.00(0.00)	8.55(3.22)
	0.15	8.61(0.63)	12.77(0.21)	13.74(1.37)	15.00(0.00)	11.12(3.68)
	0.20	10.05(1.09)	15.46(0.36)	13.80(2.40)	19.21(1.21)	10.52(3.41)
	0.25	11.66(1.29)	17.05(0.33)	9.89(1.07)	21.69(0.05)	8.91(0.57)
	0.30	13.03(1.37)	17.22(0.35)	8.17(0.39)	20.41(3.14)	9.47(0.84)
	0.35	14.54(1.13)	15.81(0.44)	8.13(0.37)	10.38(1.23)	9.06(0.29)
	0.40	16.03(2.26)	13.35(0.56)	8.94(0.75)	9.93(0.34)	9.02(0.30)
Spambase	0.10	9.25(0.93)	9.70(0.15)	9.95(0.52)	8.62(0.39)	7.44(0.24)
	0.15	11.55(0.82)	14.07(0.24)	13.61(1.28)	12.27(0.45)	9.43(0.32)
	0.20	13.84(0.85)	17.93(0.35)	14.95(1.41)	15.45(0.61)	11.70(0.62)
	0.25	15.91(1.26)	20.32(0.59)	14.21(1.39)	18.50(0.76)	12.62(0.74)
	0.30	17.07(1.52)	21.43(0.83)	12.87(1.36)	20.16(0.82)	13.11(0.88)
	0.35	17.92(1.93)	19.56(0.97)	12.95(0.90)	18.74(1.57)	12.06(0.60)
	0.40	19.67(2.44)	17.24(0.82)	13.37(0.73)	14.29(0.87)	10.75(0.49)
Subj	0.10	10.05(0.41)	9.14(0.07)	10.26(0.43)	10.00(0.00)	7.58(0.13)
	0.15	12.38(0.43)	12.99(0.15)	14.19(1.22)	15.00(0.00)	9.00(0.20)
	0.20	14.32(0.71)	16.28(0.29)	16.15(1.70)	20.00(0.00)	10.11(0.25)
	0.25	16.03(0.89)	18.70(0.32)	13.63(0.98)	25.00(0.00)	11.04(0.27)
	0.30	17.68(0.62)	19.95(0.46)	13.74(0.80)	23.61(2.42)	11.73(0.21)
	0.35	19.74(0.84)	19.97(0.40)	13.81(0.67)	17.25(0.52)	12.12(0.17)
	0.40	20.97(1.10)	19.12(0.42)	13.94(0.55)	14.97(0.44)	12.51(0.17)
Twonorm	0.10	4.26(0.52)	8.61(0.19)	8.99(1.59)	10.00(0.00)	1.58(0.15)
	0.15	5.57(1.07)	12.38(0.54)	8.13(3.57)	14.97(0.05)	2.06(0.19)
	0.20	6.13(0.90)	14.97(0.85)	7.69(3.16)	18.28(1.42)	2.12(0.16)
	0.25	7.62(1.51)	16.31(0.60)	5.49(0.53)	15.55(4.13)	2.25(0.09)
	0.30	8.44(2.13)	15.84(0.61)	3.61(0.76)	7.39(1.03)	2.48(0.22)
	0.35	9.16(1.54)	14.33(1.26)	2.94(0.37)	4.46(0.37)	2.45(0.11)
	0.40	10.67(2.60)	11.53(0.27)	2.80(0.27)	3.26(0.16)	2.51(0.13)

Table 7. Mean and standard error of 0-1 risk for accepted data of the noisy-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	10.46(2.00)	27.33(1.04)	19.21(21.10)	26.18(0.67)	6.99(0.81)
	0.15	12.35(1.86)	27.30(0.74)	12.58(3.53)	27.47(0.62)	7.70(0.69)
	0.20	11.51(1.58)	26.77(1.47)	16.13(2.20)	28.05(0.76)	9.33(0.78)
	0.25	12.87(1.78)	26.61(1.15)	17.27(3.11)	28.04(1.03)	10.28(0.81)
	0.30	13.11(1.15)	26.50(0.96)	20.47(4.24)	28.32(0.88)	10.71(1.16)
	0.35	13.96(2.14)	25.88(1.04)	24.33(2.90)	28.47(0.86)	11.79(0.75)
	0.40	15.22(1.47)	26.31(1.17)	25.84(3.05)	28.90(0.85)	12.80(0.66)
Phishing	0.10	4.67(0.77)	0.07(0.20)	12.13(18.57)	0.00(0.00)	7.60(9.63)
	0.15	5.61(1.03)	0.08(0.13)	8.77(5.76)	0.00(0.00)	7.18(7.92)
	0.20	5.74(1.04)	0.13(0.13)	9.44(3.28)	0.00(0.00)	3.70(4.76)
	0.25	5.26(0.97)	0.25(0.12)	6.40(1.46)	0.00(0.00)	2.15(0.28)
	0.30	6.24(0.91)	0.45(0.17)	6.06(1.41)	0.10(0.24)	2.83(0.64)
	0.35	5.97(1.35)	0.97(0.15)	6.28(1.10)	5.30(0.65)	3.72(0.34)
	0.40	5.52(1.32)	2.23(0.37)	7.26(0.96)	7.58(0.79)	4.63(0.34)
Spambase	0.10	8.30(2.06)	5.86(1.93)	13.46(6.41)	3.10(1.03)	5.02(0.49)
	0.15	8.40(1.75)	5.83(1.79)	10.67(2.96)	2.27(0.46)	4.65(0.57)
	0.20	8.92(1.87)	5.26(1.10)	12.25(1.67)	2.63(0.69)	5.40(0.62)
	0.25	8.11(1.20)	4.56(1.23)	11.81(2.43)	2.49(0.43)	5.49(0.60)
	0.30	9.48(1.24)	5.09(0.92)	10.18(2.08)	2.27(0.57)	5.73(0.43)
	0.35	9.05(1.60)	6.12(0.76)	11.31(1.69)	3.26(0.76)	6.61(0.50)
	0.40	9.79(1.19)	7.26(0.69)	10.79(2.01)	6.55(0.99)	7.10(0.45)
Subj	0.10	10.08(0.74)	1.76(0.53)	23.99(34.15)	0.00(0.00)	6.36(0.23)
	0.15	10.38(0.55)	2.13(0.41)	11.35(4.78)	0.00(0.00)	6.75(0.29)
	0.20	10.70(0.77)	2.39(0.39)	12.65(2.87)	0.00(0.00)	7.11(0.40)
	0.25	10.26(1.22)	2.89(0.34)	9.78(2.03)	0.00(0.00)	7.81(0.48)
	0.30	10.49(0.91)	3.51(0.21)	10.86(1.57)	2.40(0.76)	8.54(0.38)
	0.35	10.79(0.89)	4.57(0.34)	11.10(1.39)	5.06(0.45)	9.10(0.34)
	0.40	10.86(0.68)	5.88(0.28)	12.96(0.91)	7.53(0.49)	9.96(0.22)
Twonorm	0.10	2.08(0.53)	0.00(0.00)	21.63(34.02)	0.00(0.00)	0.49(0.09)
	0.15	2.03(0.50)	0.00(0.00)	5.03(3.63)	0.00(0.00)	0.54(0.12)
	0.20	2.33(0.46)	0.00(0.00)	4.65(3.06)	0.00(0.00)	0.72(0.16)
	0.25	2.06(0.40)	0.00(0.00)	3.95(0.86)	0.03(0.04)	0.78(0.13)
	0.30	2.35(0.44)	0.02(0.03)	2.58(0.90)	0.13(0.06)	0.92(0.20)
	0.35	2.11(0.39)	0.04(0.04)	2.10(0.76)	0.44(0.07)	1.10(0.10)
	0.40	2.18(0.37)	0.11(0.05)	2.03(0.65)	0.82(0.09)	1.35(0.16)

Table 8. Mean and standard error of rejection ratio of the noisy-labeled binary classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	84.89(5.15)	34.86(0.94)	96.31(3.59)	13.58(0.60)	21.90(1.23)
	0.15	76.78(7.37)	35.21(0.63)	82.31(10.98)	11.49(0.59)	19.09(1.09)
	0.20	78.23(7.98)	35.62(0.71)	70.45(8.83)	10.17(0.39)	18.17(1.26)
	0.25	77.22(7.48)	35.79(0.95)	62.96(16.44)	9.41(0.31)	16.90(1.17)
	0.30	72.54(4.04)	35.76(0.39)	47.33(19.82)	8.84(0.58)	14.59(1.32)
	0.35	67.36(7.82)	36.13(1.42)	28.41(12.22)	8.21(0.58)	13.41(0.71)
	0.40	63.43(5.69)	37.05(1.03)	18.55(11.06)	7.41(0.33)	11.32(0.83)
Phishing	0.10	34.53(6.97)	91.13(1.71)	77.34(19.65)	100.00(0.00)	62.86(6.85)
	0.15	31.53(6.13)	85.04(1.37)	72.12(14.58)	100.00(0.00)	57.96(8.23)
	0.20	30.15(6.10)	77.18(1.76)	42.01(12.15)	96.04(6.04)	44.05(8.43)
	0.25	32.27(7.31)	67.89(1.33)	18.49(5.68)	86.75(0.20)	29.58(2.91)
	0.30	28.31(7.95)	56.75(1.35)	8.49(5.42)	67.89(10.59)	24.46(2.68)
	0.35	29.32(5.64)	43.62(1.29)	6.31(3.94)	17.01(5.61)	17.06(1.36)
	0.40	30.20(8.85)	29.43(1.87)	5.07(2.86)	7.22(1.67)	12.40(0.87)
Spambase	0.10	46.18(10.07)	92.89(1.21)	78.50(24.15)	80.13(4.62)	48.50(2.35)
	0.15	45.53(11.85)	89.90(1.58)	58.81(20.39)	78.61(3.30)	46.15(1.64)
	0.20	43.21(9.73)	85.95(2.02)	33.30(18.65)	73.85(3.05)	43.18(2.86)
	0.25	45.70(9.65)	77.11(2.48)	17.14(7.24)	71.12(3.24)	36.59(2.68)
	0.30	36.53(9.99)	65.59(3.27)	12.74(9.60)	64.49(3.36)	30.41(3.54)
	0.35	34.19(6.08)	46.56(2.80)	6.60(5.41)	48.64(6.02)	19.17(2.32)
	0.40	32.59(8.55)	30.48(1.99)	8.51(5.24)	23.07(3.57)	11.07(1.90)
Subj	0.10	44.78(4.04)	89.50(0.77)	93.12(10.68)	100.00(0.00)	33.50(1.54)
	0.15	43.40(5.46)	84.34(1.30)	75.00(16.19)	100.00(0.00)	27.27(1.00)
	0.20	39.16(3.43)	78.89(1.60)	46.94(13.23)	100.00(0.00)	23.24(1.50)
	0.25	39.05(4.64)	71.48(1.55)	24.31(8.76)	100.00(0.01)	18.74(1.35)
	0.30	36.76(3.25)	62.07(1.67)	14.68(4.86)	76.61(9.20)	14.82(1.29)
	0.35	36.90(3.59)	50.59(1.50)	11.06(5.79)	40.70(2.19)	11.65(0.81)
	0.40	34.71(3.48)	38.82(0.96)	3.52(3.75)	22.89(1.39)	8.47(0.64)
Twonorm	0.10	27.32(6.19)	86.07(1.90)	79.71(25.94)	100.00(0.00)	11.44(2.23)
	0.15	27.14(8.93)	82.55(3.58)	37.35(22.79)	99.81(0.36)	10.49(1.72)
	0.20	21.40(5.79)	74.87(4.25)	20.48(7.83)	91.38(7.10)	7.27(1.32)
	0.25	24.16(7.34)	65.24(2.39)	7.18(3.93)	62.14(16.54)	6.08(0.70)
	0.30	21.98(8.06)	52.76(2.03)	3.69(3.04)	24.29(3.53)	5.34(1.16)
	0.35	21.42(4.80)	40.88(3.61)	2.52(1.24)	11.61(1.07)	3.98(0.46)
	0.40	22.46(6.83)	28.62(0.76)	2.02(1.63)	6.23(0.45)	3.01(0.46)

Table 9. Mean and standard error of 0-1-c risk of the positive-unlabeled classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	11.18(0.21)	17.90(1.22)	24.85(2.86)	10.63(0.62)	9.28(0.79)
	0.15	16.01(0.14)	20.96(1.36)	30.54(1.93)	15.31(1.20)	11.17(0.62)
	0.20	20.87(0.18)	21.63(2.20)	30.83(2.65)	19.87(1.37)	12.11(0.69)
	0.25	25.59(0.26)	22.98(1.84)	34.17(2.15)	22.73(1.32)	14.23(0.98)
	0.30	30.16(0.81)	24.69(2.80)	33.99(2.17)	26.46(1.38)	15.41(0.93)
	0.35	34.67(0.91)	26.10(1.91)	36.58(2.46)	29.17(1.90)	18.10(1.17)
	0.40	40.04(0.17)	26.76(2.57)	36.25(2.42)	31.08(1.01)	19.69(0.93)
Phishing	0.10	11.13(0.95)	5.98(0.16)	9.59(1.16)	6.46(1.14)	4.91(0.44)
	0.15	12.45(0.61)	8.40(0.19)	12.62(1.22)	7.30(0.84)	5.96(0.25)
	0.20	13.43(0.74)	10.07(0.12)	13.59(0.79)	9.31(0.51)	7.12(0.33)
	0.25	14.64(0.62)	11.36(0.25)	13.56(0.88)	10.22(0.41)	8.26(0.38)
	0.30	15.31(0.89)	12.14(0.33)	13.10(0.59)	11.11(0.43)	9.15(0.32)
	0.35	16.42(1.15)	12.81(0.26)	12.56(0.38)	11.74(0.37)	9.86(0.39)
	0.40	16.72(1.22)	13.22(0.25)	12.17(0.39)	11.82(0.24)	10.40(0.34)
Spambase	0.10	21.59(2.37)	12.03(0.75)	18.98(2.07)	11.26(0.56)	9.91(0.72)
	0.15	23.52(2.13)	15.40(0.97)	22.83(2.34)	15.00(2.10)	12.63(0.65)
	0.20	24.97(1.50)	18.13(0.35)	25.04(1.74)	20.22(0.93)	14.82(1.03)
	0.25	26.10(1.11)	19.49(0.56)	25.53(2.02)	22.66(0.95)	17.04(0.97)
	0.30	26.97(1.52)	21.55(0.36)	25.73(1.10)	24.60(0.71)	19.60(1.29)
	0.35	28.35(1.38)	22.32(0.83)	26.30(1.04)	25.93(0.84)	21.23(1.22)
	0.40	29.26(1.11)	24.17(0.58)	25.47(0.98)	26.43(0.66)	21.71(1.18)
Subj	0.10	18.61(0.81)	7.84(0.13)	16.76(2.40)	7.88(0.16)	8.87(0.35)
	0.15	20.82(2.10)	10.83(0.18)	20.67(2.17)	10.86(0.43)	10.05(0.21)
	0.20	22.33(0.88)	13.10(0.31)	25.37(1.48)	14.74(0.50)	11.19(0.45)
	0.25	23.43(1.09)	15.31(0.22)	26.45(1.42)	17.34(0.50)	12.21(0.39)
	0.30	24.91(0.95)	17.34(0.33)	26.50(1.14)	19.27(0.34)	13.21(0.40)
	0.35	25.95(0.69)	18.42(0.39)	25.70(1.39)	20.17(0.42)	13.90(0.36)
	0.40	27.11(0.59)	19.57(0.36)	25.27(0.80)	21.20(0.30)	14.42(0.31)
Twonorm	0.10	8.59(1.44)	5.29(0.11)	8.27(2.20)	2.38(0.53)	1.51(0.05)
	0.15	8.66(1.04)	7.13(0.31)	12.58(2.24)	2.46(0.34)	1.80(0.19)
	0.20	9.76(0.84)	8.77(0.35)	11.88(2.21)	3.37(0.66)	2.19(0.22)
	0.25	10.88(1.29)	9.78(0.31)	10.59(2.05)	4.02(0.50)	2.46(0.19)
	0.30	11.09(1.03)	10.37(0.37)	9.20(0.78)	4.89(0.46)	2.63(0.18)
	0.35	12.64(2.13)	10.29(0.39)	8.59(0.90)	5.76(0.59)	3.07(0.18)
	0.40	12.90(1.51)	9.99(0.42)	7.55(0.59)	6.63(0.58)	3.21(0.15)

Table 10. Mean and standard error of 0-1 risk for accepted data of the positive-unlabeled classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	45.39(9.08)	21.52(1.57)	29.98(2.93)	11.04(0.98)	9.03(1.06)
	0.15	49.00(7.41)	23.43(2.00)	32.99(1.91)	15.54(1.87)	9.69(0.89)
	0.20	46.40(7.76)	22.30(3.05)	32.12(2.69)	19.82(2.05)	9.35(1.01)
	0.25	44.87(9.27)	22.28(2.42)	34.83(2.03)	21.63(2.03)	10.40(0.98)
	0.30	39.79(10.70)	23.05(3.51)	34.13(2.23)	24.77(2.12)	10.47(1.07)
	0.35	34.16(10.19)	23.07(2.43)	36.60(2.56)	26.73(2.59)	11.25(1.76)
	0.40	41.78(4.27)	23.30(2.76)	36.17(2.47)	27.11(1.75)	11.53(0.98)
Phishing	0.10	11.36(1.12)	1.06(0.30)	9.39(1.40)	0.32(0.37)	3.07(0.78)
	0.15	11.52(1.16)	1.60(0.33)	11.82(1.77)	3.08(1.05)	3.79(0.48)
	0.20	11.55(1.15)	1.86(0.17)	12.54(1.36)	5.88(0.29)	4.73(0.58)
	0.25	11.82(1.17)	2.29(0.25)	12.05(1.72)	7.24(0.28)	5.99(0.56)
	0.30	12.25(1.53)	2.68(0.34)	11.05(1.74)	7.97(0.33)	7.30(0.34)
	0.35	11.65(1.38)	3.61(0.45)	10.92(1.15)	9.12(0.36)	8.41(0.36)
	0.40	11.68(1.60)	4.67(0.30)	10.78(1.08)	9.89(0.32)	9.17(0.29)
Spambase	0.10	24.75(2.41)	13.79(1.21)	20.75(1.98)	12.12(0.87)	9.84(1.09)
	0.15	25.37(2.05)	15.66(1.61)	23.71(2.08)	14.68(3.17)	11.72(0.87)
	0.20	26.10(1.56)	17.10(0.59)	25.37(1.73)	20.25(1.13)	13.23(1.38)
	0.25	26.30(1.31)	17.17(0.72)	25.45(2.21)	22.21(1.16)	15.13(1.24)
	0.30	26.32(2.01)	18.63(0.59)	25.43(1.26)	23.84(0.96)	17.92(1.68)
	0.35	27.06(1.82)	18.65(0.84)	25.89(1.48)	24.97(0.90)	19.72(1.56)
	0.40	26.77(1.55)	20.34(0.51)	25.04(1.25)	25.43(0.70)	20.15(1.44)
Subj	0.10	22.08(0.97)	4.37(0.31)	20.04(2.78)	4.04(0.98)	8.62(0.45)
	0.15	22.57(2.20)	4.86(0.39)	22.67(2.59)	7.16(1.00)	9.19(0.27)
	0.20	23.10(1.12)	5.26(0.49)	26.43(1.64)	11.75(0.77)	9.88(0.59)
	0.25	22.86(1.46)	6.31(0.41)	26.63(1.62)	14.23(0.50)	10.61(0.46)
	0.30	23.43(1.41)	7.83(0.52)	26.09(1.37)	16.00(0.49)	11.56(0.48)
	0.35	23.36(1.25)	8.85(0.63)	24.53(2.07)	17.08(0.58)	12.34(0.32)
	0.40	23.12(1.28)	10.87(0.52)	24.36(1.35)	18.70(0.33)	13.04(0.38)
Twonorm	0.10	8.08(1.92)	0.03(0.05)	7.91(2.49)	0.52(0.22)	0.95(0.17)
	0.15	6.88(1.11)	0.01(0.03)	12.24(2.59)	1.18(0.42)	1.03(0.27)
	0.20	7.50(1.47)	0.03(0.08)	11.06(2.82)	1.84(0.47)	1.37(0.22)
	0.25	7.89(1.26)	0.10(0.12)	9.85(2.78)	2.60(0.35)	1.62(0.26)
	0.30	7.46(1.96)	0.11(0.14)	7.87(1.71)	3.33(0.48)	1.82(0.20)
	0.35	6.76(1.38)	0.13(0.13)	7.73(1.12)	3.85(0.44)	2.27(0.24)
	0.40	7.84(1.06)	0.33(0.09)	6.61(1.40)	4.50(0.47)	2.52(0.18)

Table 11. Mean and standard error of rejection ratio of the positive-unlabeled classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gisette	0.10	96.36(1.36)	31.50(4.21)	26.12(6.31)	39.56(3.48)	26.76(2.69)
	0.15	96.92(0.66)	29.06(3.77)	13.71(4.00)	35.44(3.75)	27.78(1.75)
	0.20	96.50(0.93)	25.94(3.29)	11.09(3.93)	33.52(2.22)	25.88(1.35)
	0.25	96.39(1.54)	26.53(3.65)	7.14(2.78)	31.95(2.10)	26.34(3.03)
	0.30	95.30(2.98)	24.51(3.87)	4.46(2.48)	31.83(2.13)	25.29(2.97)
	0.35	93.86(3.04)	25.63(3.14)	3.55(1.98)	29.38(2.52)	28.79(2.32)
	0.40	96.08(0.65)	21.09(2.85)	2.11(2.06)	30.58(3.38)	28.63(2.62)
Phishing	0.10	21.32(6.66)	55.03(0.67)	21.98(6.20)	63.11(12.67)	26.21(4.62)
	0.15	23.95(8.43)	50.74(0.78)	21.17(7.93)	35.15(7.50)	19.23(1.63)
	0.20	21.74(5.80)	45.28(0.70)	13.06(6.49)	24.29(2.75)	15.62(1.73)
	0.25	21.02(5.39)	39.91(1.18)	10.90(5.60)	16.79(1.32)	11.92(1.83)
	0.30	16.96(4.66)	34.65(1.25)	10.33(5.73)	14.22(1.79)	8.14(0.87)
	0.35	20.09(7.37)	29.33(0.58)	6.62(3.63)	10.10(1.32)	5.44(0.78)
	0.40	17.59(5.26)	24.21(0.51)	4.63(3.13)	6.40(0.89)	3.99(0.77)
Spambase	0.10	21.96(4.71)	47.17(2.78)	17.02(5.16)	41.88(4.09)	33.29(3.08)
	0.15	18.67(5.77)	40.36(2.50)	11.51(6.73)	30.92(7.42)	27.89(1.87)
	0.20	19.56(5.30)	35.27(1.80)	7.29(5.40)	18.39(2.35)	23.28(1.81)
	0.25	17.96(3.45)	29.73(2.03)	7.14(5.66)	15.55(1.89)	19.24(1.67)
	0.30	15.63(4.62)	25.58(1.71)	5.98(3.57)	12.02(2.58)	13.66(1.98)
	0.35	15.77(3.62)	22.48(1.83)	3.79(4.59)	9.60(1.21)	9.78(1.28)
	0.40	18.58(3.12)	19.45(2.00)	2.76(2.74)	6.85(1.37)	7.79(1.35)
Subj	0.10	28.78(3.10)	61.62(0.40)	33.86(5.22)	63.66(5.24)	17.23(1.89)
	0.15	25.02(7.34)	58.82(1.06)	27.58(4.88)	46.83(4.67)	14.64(1.29)
	0.20	25.83(3.27)	53.20(1.48)	17.18(3.65)	36.16(2.67)	12.92(1.03)
	0.25	24.43(4.32)	48.17(0.85)	13.30(4.20)	28.92(2.30)	11.14(0.86)
	0.30	21.99(3.93)	42.89(0.61)	9.49(4.04)	23.34(1.20)	8.95(0.99)
	0.35	21.84(4.79)	36.59(0.65)	10.20(5.59)	17.25(1.58)	6.88(0.78)
	0.40	23.42(4.08)	29.86(0.39)	5.50(4.21)	11.72(0.99)	5.13(0.52)
Twonorm	0.10	19.36(8.49)	52.73(1.01)	13.25(4.94)	19.65(5.11)	6.15(1.69)
	0.15	21.71(8.79)	47.51(2.06)	6.63(6.32)	9.24(2.27)	5.48(0.92)
	0.20	17.54(5.54)	43.74(1.74)	7.56(5.89)	8.44(2.82)	4.40(0.74)
	0.25	17.18(8.19)	38.88(1.23)	4.22(4.28)	6.31(1.57)	3.56(0.73)
	0.30	15.69(6.22)	34.33(1.17)	5.68(5.18)	5.84(1.33)	2.87(0.58)
	0.35	20.81(6.86)	29.14(1.01)	3.13(1.76)	6.13(1.14)	2.46(0.42)
	0.40	15.58(6.44)	24.36(1.15)	2.69(3.28)	6.01(1.14)	1.84(0.49)

Table 12. Mean and standard error of 0-1-c risk of the clean-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	0.88(0.20)	2.53(0.39)	1.00(0.12)	1.22(0.08)	0.70(0.06)
	0.15	1.07(0.40)	2.84(0.50)	1.13(0.19)	1.39(0.13)	0.70(0.07)
	0.20	0.83(0.14)	3.21(0.40)	1.09(0.15)	1.34(0.17)	0.74(0.04)
	0.25	0.89(0.25)	3.49(0.51)	1.12(0.15)	1.27(0.22)	0.85(0.10)
	0.30	0.86(0.24)	2.90(1.08)	1.12(0.33)	1.33(0.27)	0.87(0.09)
	0.35	1.13(0.76)	2.64(0.59)	1.10(0.27)	1.24(0.27)	0.91(0.12)
	0.40	0.82(0.19)	2.91(0.66)	1.12(0.23)	1.17(0.22)	0.90(0.07)
HAR	0.10	1.30(0.21)	2.57(0.31)	1.17(0.26)	1.23(0.26)	1.52(0.80)
	0.15	1.42(0.44)	3.45(0.68)	1.32(0.20)	1.50(0.19)	1.25(0.14)
	0.20	1.95(1.83)	3.86(0.66)	1.53(0.22)	1.59(0.09)	1.25(0.09)
	0.25	1.29(0.21)	3.87(0.42)	1.53(0.11)	1.78(0.27)	1.42(0.39)
	0.30	1.22(0.12)	4.97(1.11)	1.55(0.21)	1.93(0.45)	1.72(1.24)
	0.35	1.79(0.77)	4.49(0.77)	1.56(0.12)	1.94(0.19)	1.44(0.19)
	0.40	1.35(0.19)	4.65(1.00)	1.64(0.17)	2.04(0.36)	1.81(1.09)
MNIST	0.10	0.63(0.04)	1.03(0.09)	90.09(0.40)	0.46(0.03)	0.47(0.05)
	0.15	0.62(0.07)	1.26(0.14)	90.02(0.58)	0.56(0.03)	0.49(0.03)
	0.20	0.63(0.06)	1.44(0.09)	89.86(0.46)	0.62(0.05)	0.56(0.04)
	0.25	0.65(0.04)	1.50(0.09)	90.03(0.49)	0.66(0.05)	0.63(0.05)
	0.30	0.61(0.05)	1.64(0.09)	89.82(0.43)	0.74(0.04)	0.67(0.04)
	0.35	0.66(0.07)	1.61(0.13)	89.68(0.77)	0.82(0.05)	0.77(0.06)
	0.40	0.65(0.04)	1.65(0.15)	89.82(0.46)	0.86(0.04)	0.82(0.05)
Fashion-MNIST	0.10	7.82(0.20)	3.79(0.21)	90.00(0.00)	3.82(0.21)	5.47(0.17)
	0.15	7.91(0.22)	4.76(0.13)	90.00(0.00)	4.83(0.15)	5.94(0.15)
	0.20	7.95(0.20)	5.77(0.16)	90.00(0.00)	5.82(0.17)	6.42(0.18)
	0.25	7.85(0.27)	6.54(0.19)	90.00(0.00)	6.52(0.13)	6.73(0.10)
	0.30	7.96(0.20)	7.16(0.27)	90.00(0.00)	7.19(0.21)	7.25(0.22)
	0.35	7.83(0.20)	7.67(0.20)	90.00(0.00)	7.85(0.19)	7.61(0.19)
	0.40	7.86(0.16)	8.02(0.23)	90.00(0.00)	8.29(0.26)	7.97(0.18)
KMNIST	0.10	4.31(0.16)	3.53(0.14)	90.00(0.00)	2.43(0.12)	2.39(0.13)
	0.15	4.26(0.09)	4.81(0.24)	90.00(0.00)	3.03(0.15)	2.86(0.14)
	0.20	4.14(0.20)	5.70(0.33)	90.00(0.00)	3.36(0.17)	3.21(0.17)
	0.25	4.32(0.20)	6.55(0.21)	90.00(0.00)	3.86(0.20)	3.67(0.13)
	0.30	4.34(0.30)	6.94(0.19)	90.00(0.00)	4.25(0.23)	3.96(0.18)
	0.35	4.43(0.26)	7.24(0.37)	90.00(0.00)	4.69(0.24)	4.29(0.20)
	0.40	4.27(0.32)	7.36(0.20)	90.00(0.00)	4.92(0.26)	4.67(0.34)

Table 13. Mean and standard error of 0-1 risk for accepted data of the clean-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	0.87(0.19)	0.33(0.08)	0.82(0.18)	0.39(0.04)	0.51(0.09)
	0.15	1.07(0.40)	0.39(0.09)	0.88(0.16)	0.40(0.05)	0.46(0.06)
	0.20	0.83(0.14)	0.38(0.12)	0.98(0.12)	0.40(0.03)	0.47(0.04)
	0.25	0.88(0.25)	0.42(0.13)	1.01(0.17)	0.44(0.05)	0.56(0.08)
	0.30	0.86(0.24)	0.40(0.09)	0.84(0.20)	0.53(0.20)	0.58(0.07)
	0.35	1.13(0.76)	0.41(0.09)	0.95(0.13)	0.61(0.17)	0.56(0.07)
	0.40	0.82(0.19)	0.45(0.08)	1.04(0.24)	0.51(0.07)	0.56(0.07)
HAR	0.10	1.30(0.21)	0.09(0.08)	0.80(0.46)	0.70(0.33)	1.28(0.68)
	0.15	1.42(0.44)	0.10(0.12)	0.84(0.31)	0.83(0.26)	1.07(0.15)
	0.20	1.95(1.83)	0.08(0.05)	1.23(0.20)	0.84(0.15)	1.04(0.10)
	0.25	1.29(0.21)	0.25(0.18)	1.32(0.18)	1.00(0.28)	1.10(0.16)
	0.30	1.22(0.12)	0.15(0.11)	1.42(0.25)	1.09(0.44)	1.23(0.68)
	0.35	1.79(0.77)	0.14(0.07)	1.56(0.10)	1.14(0.22)	1.09(0.20)
	0.40	1.35(0.19)	0.18(0.07)	1.46(0.28)	1.16(0.36)	1.04(0.16)
MNIST	0.10	0.63(0.04)	0.01(0.01)	90.09(0.40)	0.14(0.05)	0.35(0.05)
	0.15	0.62(0.07)	0.01(0.01)	90.02(0.58)	0.18(0.04)	0.30(0.03)
	0.20	0.63(0.06)	0.02(0.01)	89.86(0.46)	0.20(0.05)	0.31(0.03)
	0.25	0.65(0.04)	0.04(0.01)	90.03(0.49)	0.20(0.03)	0.32(0.04)
	0.30	0.61(0.05)	0.04(0.01)	89.82(0.43)	0.22(0.04)	0.30(0.05)
	0.35	0.66(0.07)	0.05(0.02)	89.68(0.77)	0.26(0.05)	0.33(0.04)
	0.40	0.65(0.04)	0.07(0.02)	89.82(0.46)	0.25(0.03)	0.32(0.03)
Fashion-MNIST	0.10	7.82(0.20)	1.29(0.15)	90.00(0.00)	1.69(0.36)	5.03(0.19)
	0.15	7.91(0.22)	1.54(0.13)	90.00(0.00)	1.73(0.21)	5.07(0.16)
	0.20	7.95(0.20)	1.84(0.15)	90.00(0.00)	2.24(0.23)	5.17(0.21)
	0.25	7.85(0.27)	2.21(0.19)	90.00(0.00)	2.53(0.23)	5.13(0.09)
	0.30	7.96(0.20)	2.65(0.26)	90.00(0.00)	2.95(0.26)	5.26(0.24)
	0.35	7.83(0.21)	3.15(0.20)	90.00(0.00)	3.46(0.31)	5.25(0.14)
	0.40	7.86(0.16)	3.81(0.28)	90.00(0.00)	3.75(0.29)	5.09(0.19)
KMNIST	0.10	4.31(0.16)	0.11(0.03)	90.00(0.00)	0.91(0.16)	1.75(0.13)
	0.15	4.25(0.09)	0.16(0.04)	90.00(0.00)	0.92(0.12)	1.84(0.16)
	0.20	4.13(0.20)	0.26(0.04)	90.00(0.00)	1.02(0.16)	1.80(0.22)
	0.25	4.30(0.20)	0.30(0.04)	90.00(0.00)	1.14(0.13)	1.98(0.16)
	0.30	4.33(0.30)	0.36(0.07)	90.00(0.00)	1.35(0.14)	1.92(0.24)
	0.35	4.42(0.26)	0.45(0.05)	90.00(0.00)	1.39(0.11)	2.07(0.16)
	0.40	4.26(0.32)	0.50(0.09)	90.00(0.00)	1.44(0.12)	2.06(0.17)

Table 14. Mean and standard error of rejection ratio of the clean-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	0.03(0.09)	22.78(3.96)	1.91(1.34)	8.61(0.95)	2.03(0.72)
	0.15	0.00(0.00)	16.79(3.17)	1.77(0.99)	6.77(0.94)	1.68(0.46)
	0.20	0.00(0.00)	14.43(1.96)	0.57(0.77)	4.80(0.86)	1.36(0.21)
	0.25	0.01(0.02)	12.47(2.14)	0.47(0.35)	3.35(0.86)	1.18(0.30)
	0.30	0.00(0.00)	8.46(3.63)	0.93(1.21)	2.71(0.79)	0.99(0.25)
	0.35	0.00(0.01)	6.44(1.56)	0.43(0.72)	1.83(0.57)	1.01(0.26)
	0.40	0.00(0.01)	6.22(1.77)	0.18(0.33)	1.67(0.48)	0.86(0.16)
HAR	0.10	0.00(0.00)	24.95(2.93)	3.91(2.31)	5.68(0.72)	2.93(2.03)
	0.15	0.00(0.00)	22.49(4.38)	3.35(1.33)	4.71(0.97)	1.34(0.30)
	0.20	0.00(0.00)	18.97(3.46)	1.62(0.93)	3.92(0.57)	1.13(0.36)
	0.25	0.00(0.00)	14.62(1.97)	0.87(0.84)	3.23(0.58)	1.35(1.03)
	0.30	0.00(0.00)	16.13(3.70)	0.44(0.45)	2.90(0.57)	1.77(2.16)
	0.35	0.00(0.00)	12.46(2.25)	0.02(0.07)	2.37(0.30)	1.04(0.21)
	0.40	0.00(0.00)	11.22(2.46)	0.45(1.03)	2.28(0.37)	1.98(2.52)
MNIST	0.10	0.01(0.01)	10.19(0.91)	0.00(0.00)	3.24(0.24)	1.24(0.06)
	0.15	0.00(0.00)	8.32(0.94)	0.00(0.00)	2.56(0.18)	1.24(0.09)
	0.20	0.00(0.00)	7.09(0.44)	0.00(0.00)	2.12(0.11)	1.30(0.09)
	0.25	0.00(0.00)	5.84(0.35)	0.00(0.00)	1.87(0.19)	1.29(0.13)
	0.30	0.00(0.00)	5.36(0.32)	0.00(0.00)	1.76(0.10)	1.24(0.08)
	0.35	0.00(0.00)	4.47(0.36)	0.00(0.00)	1.61(0.07)	1.28(0.15)
	0.40	0.00(0.00)	3.95(0.33)	0.00(0.00)	1.53(0.09)	1.26(0.09)
Fashion-MNIST	0.10	0.01(0.01)	28.67(2.42)	0.00(0.00)	25.64(1.80)	8.78(0.45)
	0.15	0.01(0.01)	23.87(1.21)	0.00(0.00)	23.39(1.18)	8.78(0.23)
	0.20	0.00(0.01)	21.65(0.89)	0.00(0.00)	20.16(1.45)	8.43(0.34)
	0.25	0.00(0.00)	18.98(0.98)	0.00(0.00)	17.75(0.97)	8.07(0.40)
	0.30	0.00(0.00)	16.48(1.25)	0.00(0.00)	15.67(1.00)	8.04(0.35)
	0.35	0.00(0.00)	14.21(0.75)	0.00(0.00)	13.89(0.93)	7.94(0.29)
	0.40	0.00(0.00)	11.64(0.86)	0.00(0.00)	12.52(1.22)	8.26(0.33)
KMNIST	0.10	0.12(0.05)	34.53(1.47)	0.00(0.00)	16.72(1.31)	7.81(0.49)
	0.15	0.10(0.05)	31.36(1.63)	0.00(0.00)	14.99(0.87)	7.81(0.25)
	0.20	0.08(0.03)	27.53(1.73)	0.00(0.00)	12.32(0.60)	7.77(0.53)
	0.25	0.06(0.03)	25.29(0.87)	0.00(0.00)	11.41(0.69)	7.34(0.41)
	0.30	0.04(0.01)	22.21(0.64)	0.00(0.00)	10.14(0.64)	7.25(0.47)
	0.35	0.03(0.02)	19.66(1.06)	0.00(0.00)	9.82(0.62)	6.75(0.43)
	0.40	0.02(0.01)	17.38(0.61)	0.00(0.00)	9.03(0.61)	6.88(0.60)

Table 15. Mean and standard error of 0-1-c risk of the noisy-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	2.44(0.66)	8.29(0.35)	9.66(0.32)	9.16(0.19)	1.88(0.18)
	0.15	2.80(1.01)	11.14(0.66)	13.73(2.12)	12.88(0.43)	2.10(0.19)
	0.20	3.13(0.83)	13.54(1.26)	16.95(4.65)	14.49(0.98)	2.08(0.18)
	0.25	2.57(0.73)	13.75(1.62)	22.02(5.15)	11.38(1.13)	1.94(0.28)
	0.30	3.34(1.46)	14.43(1.13)	24.32(7.80)	7.33(0.74)	1.67(0.37)
	0.35	2.44(0.89)	12.43(1.84)	31.49(4.50)	5.33(0.46)	1.59(0.27)
	0.40	2.48(0.46)	12.92(2.28)	29.77(11.38)	4.42(0.43)	1.52(0.27)
HAR	0.10	24.36(3.18)	9.05(0.74)	9.25(0.82)	7.21(0.50)	2.68(0.37)
	0.15	23.37(3.51)	11.50(0.63)	11.98(2.28)	9.52(0.84)	3.50(0.43)
	0.20	25.47(3.55)	14.26(1.00)	12.07(3.00)	11.14(0.73)	3.58(0.34)
	0.25	23.94(2.69)	17.37(1.18)	13.80(4.15)	12.17(0.64)	3.86(0.28)
	0.30	23.85(3.93)	19.39(1.30)	13.67(5.36)	12.29(1.81)	5.23(0.96)
	0.35	24.56(3.33)	22.23(1.92)	10.76(0.98)	11.84(1.70)	5.53(1.37)
	0.40	26.99(3.28)	24.87(1.36)	10.31(2.12)	11.06(1.68)	5.96(0.86)
MNIST	0.10	1.30(0.10)	7.88(0.58)	90.18(0.48)	7.85(0.81)	0.51(0.05)
	0.15	1.28(0.10)	10.41(0.87)	90.11(0.59)	11.43(0.50)	0.64(0.05)
	0.20	1.32(0.05)	11.91(0.87)	89.83(0.64)	14.30(1.20)	0.71(0.03)
	0.25	1.36(0.09)	13.13(1.32)	89.98(0.79)	15.14(1.47)	0.78(0.07)
	0.30	1.33(0.12)	11.26(1.57)	89.69(0.72)	12.46(0.85)	0.85(0.08)
	0.35	1.36(0.06)	10.72(1.59)	89.92(0.58)	9.36(0.76)	0.90(0.06)
	0.40	1.37(0.13)	9.01(0.88)	90.18(0.48)	6.47(0.38)	0.97(0.07)
Fashion-MNIST	0.10	9.78(0.29)	8.44(0.30)	90.00(0.00)	9.44(0.85)	5.33(0.25)
	0.15	9.84(0.24)	12.11(0.63)	90.00(0.00)	13.77(1.33)	6.20(0.21)
	0.20	9.69(0.33)	14.44(0.88)	90.00(0.00)	16.59(2.25)	6.78(0.16)
	0.25	9.91(0.25)	16.19(0.96)	90.00(0.00)	18.42(0.72)	7.51(0.24)
	0.30	9.91(0.23)	16.49(1.21)	90.00(0.00)	18.17(0.91)	7.82(0.23)
	0.35	9.88(0.21)	17.00(1.82)	90.00(0.00)	17.45(1.00)	8.33(0.18)
	0.40	9.92(0.24)	16.76(1.12)	90.00(0.00)	16.35(0.80)	8.97(0.25)
KMNIST	0.10	7.88(0.37)	8.35(0.38)	90.00(0.00)	8.69(0.24)	2.84(0.21)
	0.15	8.14(0.49)	11.70(0.58)	90.00(0.00)	12.77(0.81)	3.55(0.25)
	0.20	8.35(0.40)	14.52(0.77)	90.00(0.00)	15.44(0.56)	4.18(0.19)
	0.25	8.50(0.31)	16.79(1.01)	90.00(0.00)	18.01(0.65)	4.54(0.20)
	0.30	8.37(0.32)	18.61(0.46)	90.00(0.00)	18.74(0.67)	4.99(0.29)
	0.35	8.55(0.42)	19.79(1.04)	90.00(0.00)	18.18(0.81)	5.53(0.28)
	0.40	8.61(0.37)	20.01(0.63)	90.00(0.00)	16.76(0.72)	5.84(0.22)

Table 16. Mean and standard error of 0-1 risk for accepted data of the noisy-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	2.43(0.66)	1.74(0.91)	11.70(29.45)	0.93(0.25)	0.58(0.08)
	0.15	2.80(1.00)	1.42(0.67)	3.51(3.95)	0.71(0.37)	0.58(0.09)
	0.20	3.11(0.82)	1.29(0.52)	2.88(2.30)	0.51(0.17)	0.55(0.08)
	0.25	2.56(0.73)	0.85(0.28)	2.38(1.30)	0.38(0.14)	0.59(0.12)
	0.30	3.33(1.45)	0.78(0.21)	2.73(1.74)	0.42(0.09)	0.54(0.04)
	0.35	2.43(0.89)	0.70(0.17)	4.61(2.84)	0.46(0.05)	0.60(0.12)
	0.40	2.46(0.45)	0.85(0.37)	3.33(1.40)	0.48(0.07)	0.63(0.13)
HAR	0.10	24.45(3.22)	7.89(1.68)	5.68(3.15)	3.11(0.86)	1.54(0.42)
	0.15	23.41(3.53)	7.33(0.78)	7.98(1.45)	2.69(0.76)	1.63(0.31)
	0.20	25.50(3.58)	7.86(1.25)	9.65(2.44)	2.54(0.73)	1.67(0.26)
	0.25	23.94(2.70)	9.10(1.30)	10.54(3.32)	2.34(0.39)	1.66(0.21)
	0.30	23.84(3.94)	9.31(1.29)	9.60(2.68)	2.17(0.54)	2.47(0.64)
	0.35	24.54(3.33)	10.91(2.21)	9.58(0.86)	2.10(0.34)	2.39(0.70)
	0.40	26.97(3.26)	12.03(1.71)	9.16(2.66)	2.00(0.34)	2.62(0.52)
MNIST	0.10	1.28(0.10)	0.00(0.00)	90.18(0.48)	0.70(0.41)	0.29(0.05)
	0.15	1.27(0.11)	0.00(0.01)	90.11(0.59)	0.83(0.63)	0.33(0.05)
	0.20	1.31(0.05)	0.00(0.01)	89.83(0.64)	0.70(0.30)	0.33(0.05)
	0.25	1.35(0.09)	0.00(0.01)	89.98(0.79)	0.64(0.35)	0.35(0.06)
	0.30	1.32(0.11)	0.00(0.01)	89.69(0.72)	0.35(0.16)	0.36(0.05)
	0.35	1.35(0.06)	0.01(0.01)	89.92(0.58)	0.33(0.09)	0.35(0.03)
	0.40	1.37(0.13)	0.01(0.01)	90.18(0.48)	0.24(0.08)	0.36(0.04)
Fashion-MNIST	0.10	9.78(0.29)	0.50(0.22)	90.00(0.00)	0.95(1.45)	4.56(0.33)
	0.15	9.82(0.24)	0.57(0.27)	90.00(0.00)	1.78(1.92)	4.87(0.30)
	0.20	9.67(0.33)	0.62(0.12)	90.00(0.00)	1.62(1.22)	4.91(0.31)
	0.25	9.89(0.25)	0.75(0.12)	90.00(0.00)	2.23(0.85)	5.32(0.30)
	0.30	9.89(0.23)	0.94(0.16)	90.00(0.00)	1.82(0.26)	5.29(0.30)
	0.35	9.87(0.21)	1.27(0.20)	90.00(0.00)	2.13(0.40)	5.46(0.22)
	0.40	9.91(0.24)	1.32(0.20)	90.00(0.00)	2.12(0.25)	5.66(0.18)
KMNIST	0.10	7.84(0.38)	0.09(0.06)	90.00(0.00)	3.01(0.79)	1.68(0.25)
	0.15	8.05(0.50)	0.14(0.07)	90.00(0.00)	2.57(1.24)	1.74(0.25)
	0.20	8.25(0.42)	0.11(0.05)	90.00(0.00)	2.89(0.84)	1.85(0.11)
	0.25	8.39(0.30)	0.16(0.06)	90.00(0.00)	2.26(0.74)	1.81(0.21)
	0.30	8.29(0.32)	0.17(0.06)	90.00(0.00)	1.78(0.42)	1.94(0.24)
	0.35	8.46(0.44)	0.18(0.06)	90.00(0.00)	1.75(0.33)	2.09(0.28)
	0.40	8.55(0.37)	0.24(0.07)	90.00(0.00)	1.64(0.31)	2.09(0.21)

Table 17. Mean and standard error of rejection ratio of the noisy-labeled multiclass classification setting (rescaled to 0-100).

Dataset	c	SCE	DEFER	ANGLE	CS-hinge	CS-sigmoid
Gas-Drift	0.10	0.08(0.05)	79.35(3.21)	95.94(3.27)	90.74(1.95)	13.80(1.67)
	0.15	0.04(0.07)	71.61(4.47)	87.60(21.06)	85.17(2.82)	10.56(1.05)
	0.20	0.10(0.19)	65.57(6.27)	79.39(32.80)	71.74(4.99)	7.87(1.10)
	0.25	0.03(0.05)	53.48(6.20)	85.90(26.03)	44.69(4.55)	5.52(1.14)
	0.30	0.04(0.04)	46.74(3.64)	77.63(32.41)	23.37(2.37)	3.85(1.21)
	0.35	0.02(0.02)	34.20(5.23)	87.36(18.73)	14.11(1.34)	2.88(0.71)
	0.40	0.04(0.07)	30.83(5.82)	71.95(31.27)	9.97(1.08)	2.25(0.62)
HAR	0.10	0.61(0.19)	58.17(5.07)	70.44(30.79)	59.38(5.42)	13.46(2.02)
	0.15	0.38(0.11)	54.58(5.37)	54.29(35.03)	55.73(4.66)	14.01(2.05)
	0.20	0.45(0.23)	53.02(3.74)	21.57(23.57)	49.32(3.07)	10.42(1.45)
	0.25	0.30(0.18)	52.25(3.96)	18.63(26.08)	43.37(2.41)	9.42(1.45)
	0.30	0.25(0.08)	48.83(4.36)	17.54(26.15)	36.41(5.83)	10.04(1.54)
	0.35	0.14(0.04)	47.23(4.40)	4.60(3.37)	29.61(5.03)	9.66(2.52)
	0.40	0.26(0.45)	45.96(2.67)	3.53(3.23)	23.87(3.99)	8.96(1.23)
MNIST	0.10	0.18(0.06)	78.81(5.78)	0.00(0.00)	76.61(9.03)	2.30(0.25)
	0.15	0.13(0.04)	69.42(5.79)	0.00(0.00)	74.89(3.04)	2.10(0.16)
	0.20	0.08(0.04)	59.56(4.36)	0.00(0.00)	70.45(6.35)	1.95(0.19)
	0.25	0.05(0.02)	52.52(5.28)	0.00(0.00)	59.45(6.71)	1.78(0.14)
	0.30	0.05(0.02)	37.52(5.24)	0.00(0.00)	40.82(2.83)	1.64(0.18)
	0.35	0.02(0.02)	30.61(4.55)	0.00(0.00)	26.07(2.16)	1.60(0.13)
	0.40	0.01(0.01)	22.51(2.21)	0.00(0.00)	15.67(0.92)	1.55(0.16)
Fashion-MNIST	0.10	0.36(0.07)	83.60(3.09)	0.00(0.00)	91.82(12.52)	14.12(0.92)
	0.15	0.28(0.11)	79.91(4.44)	0.00(0.00)	89.29(11.38)	13.07(0.80)
	0.20	0.18(0.07)	71.30(4.59)	0.00(0.00)	80.71(12.75)	12.41(1.06)
	0.25	0.14(0.06)	63.68(4.00)	0.00(0.00)	71.09(3.21)	11.14(0.80)
	0.30	0.08(0.02)	53.51(4.18)	0.00(0.00)	58.00(3.47)	10.21(0.63)
	0.35	0.05(0.02)	46.65(5.38)	0.00(0.00)	46.59(3.05)	9.72(0.49)
	0.40	0.02(0.01)	39.91(2.93)	0.00(0.00)	37.56(2.28)	9.62(0.42)
KMNIST	0.10	1.45(0.26)	83.36(3.81)	0.00(0.00)	81.34(2.64)	13.92(1.34)
	0.15	1.23(0.23)	77.79(3.91)	0.00(0.00)	81.57(6.82)	13.68(0.73)
	0.20	0.84(0.15)	72.46(3.89)	0.00(0.00)	73.39(2.59)	12.83(0.96)
	0.25	0.65(0.15)	66.96(4.09)	0.00(0.00)	69.20(3.08)	11.78(0.72)
	0.30	0.39(0.11)	61.83(1.54)	0.00(0.00)	60.11(2.23)	10.87(0.78)
	0.35	0.32(0.11)	56.30(3.03)	0.00(0.00)	49.41(2.51)	10.46(0.91)
	0.40	0.17(0.04)	49.72(1.58)	0.00(0.00)	39.41(2.04)	9.87(0.43)

D.2. More discussion on the proposed rejection conditions

In this section, we provide more discussion on our proposed rejection conditions. We provide additional experimental results without using Cond. (8). We found that using Cond. (8) slightly affects the performance of our cost-sensitive approach. This is because there is only a small portion of data that satisfies Cond. (8). Note that if we succeed to obtain the optimal classifier $\boldsymbol{g}^*$, this condition is *impossible* to be satisfied. Recall that in Section 3.4, for $\boldsymbol{g}^*$, at most one $g_y^*(\boldsymbol{x})$ can be more than zero since it implies $\eta_y > 1 - c > 0.5$. Nevertheless, Cond. (8) may hold in practice due to empirical estimation.

Figures 6 and 7 illustrate the performance of the proposed approach that uses only Cond. (7) as a rejection condition and other baselines in the binary and multiclass settings, respectively. Figures 8, 9, and 10 illustrate rejected test data based on different conditions for MNIST, Fashion-MNIST, and KMNIST, respectively.

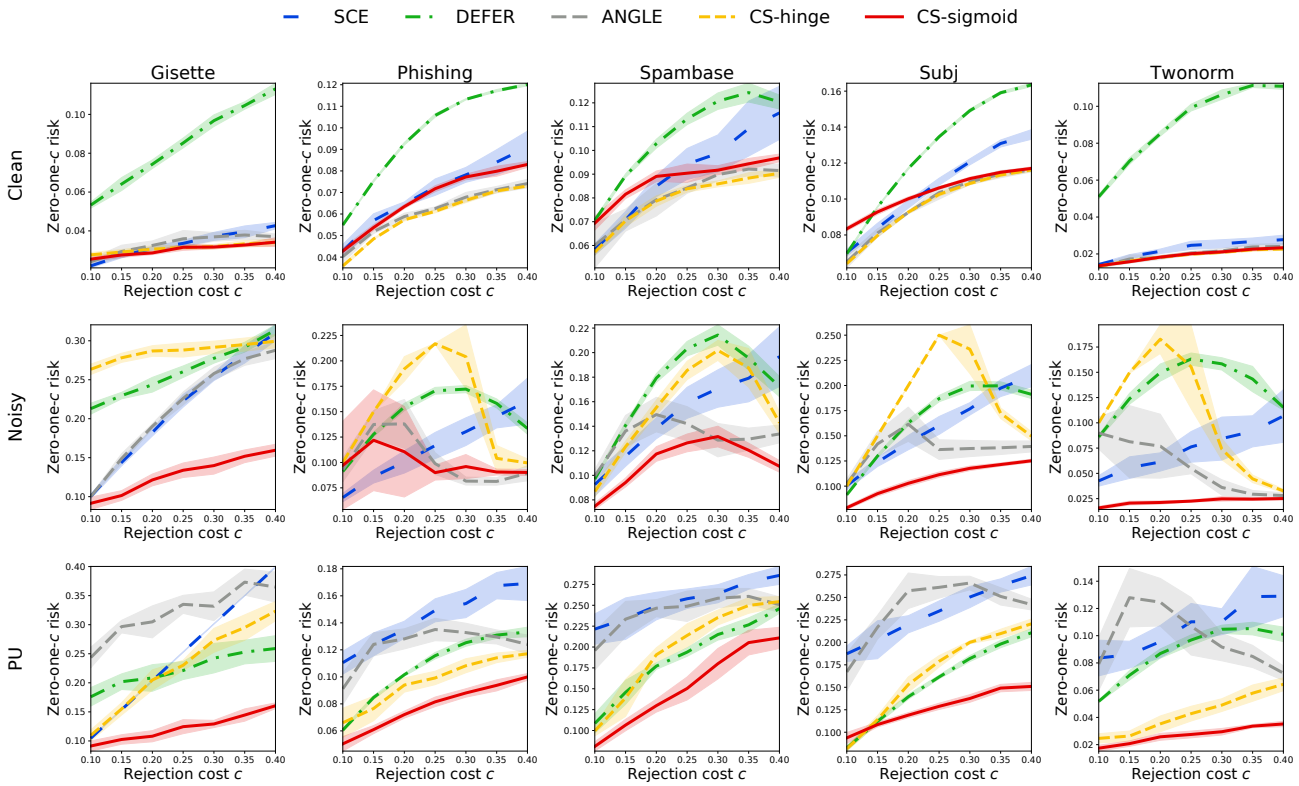

Figure 6. Mean and standard error of the test empirical zero-one-c risk over ten trials with varying rejection costs (Binary classification). Our approach in this figure only used Cond. (7) as a rejection condition. Each column indicates the performance with respect to one dataset. (Top) clean-labeled classification with rejection. (Middle) noisy-labeled classification with rejection. (Bottom) PU-classification with rejection. It can be seen that a similar trend as Figure 4 can be observed.

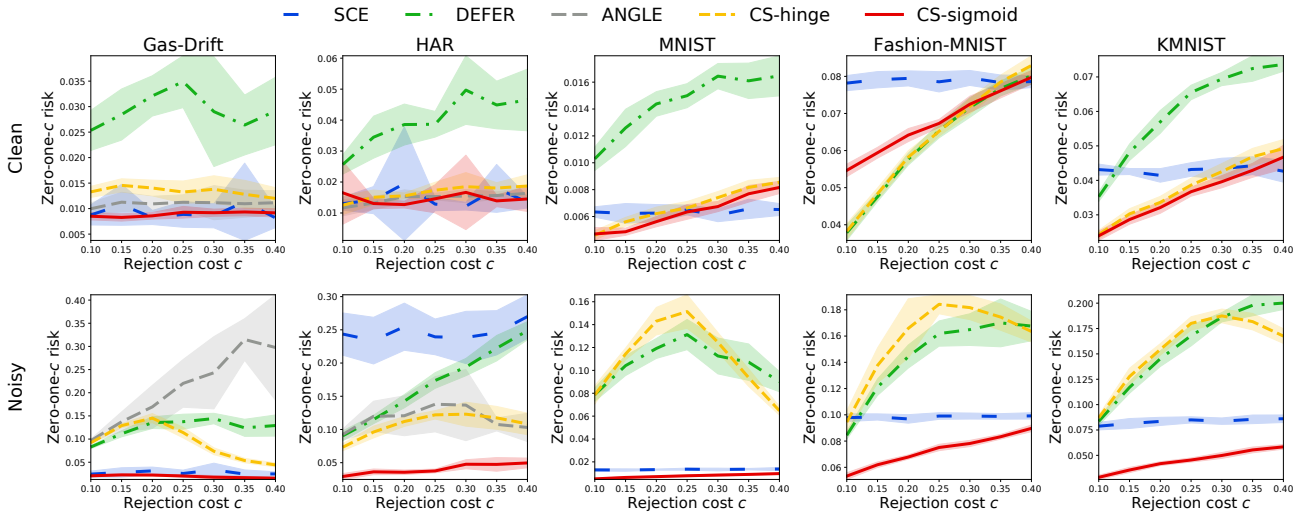

Figure 7. Mean and standard error of the test empirical zero-one-c risk over ten trials with varying rejection cost (Multiclass classification). Our approach in this figure only used Cond. (7) as a rejection condition. Each column indicates the performance with respect to one dataset. (Top) clean-labeled classification with rejection. (Bottom) noisy-labeled classification with rejection. For MNIST, Fashion-MNIST, and KMNIST, we found that ANGLE failed miserably and has zero-one-c risk more than 0.5 and thus it is excluded from the figure for readability. It can be seen that a similar trend as Figure 5 can be observed.

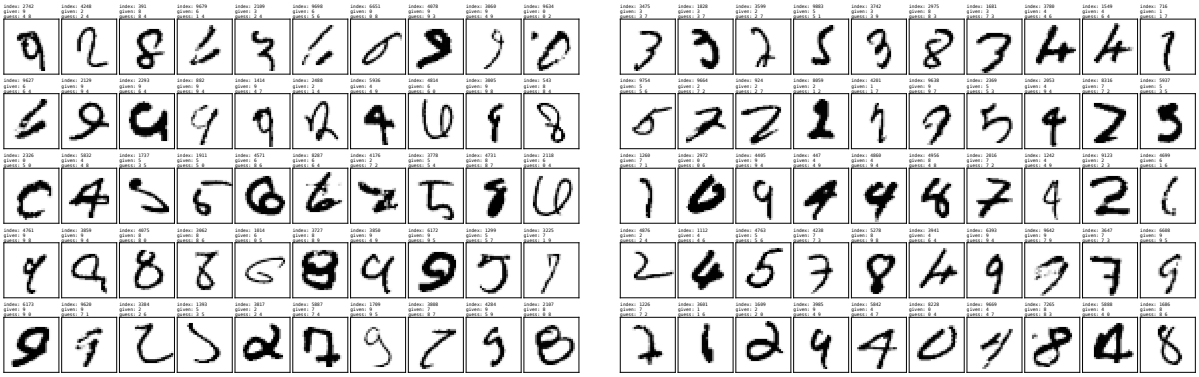

Figure 8. Examples of rejected test data in MNIST. (Left) rejections that were made by Cond. (7) (distance rejection). (Right) rejections that were made by Cond. (8) (ambiguity rejection). The data rejected by Cond. (7) appeared to look more chaotic than the one rejected by Cond. (8). On the right figure, several images can be seen to be able to associated to more than one classes, e.g., 1 vs 7, 4 vs 9, and 3 vs 5.

Figure 9. Examples of rejected test data in Fashion-MNIST. (Left) rejections that were made by Cond. (7) (distance rejection). (Right) rejections that were made by Cond. (8) (ambiguity rejection). The data rejected by Cond. (7) appeared to have more texture information than the one rejected by Cond. (8).

Figure 10. Examples of rejected test data in KMNIST. (Left) rejections that were made by Cond. (7) (distance rejection). (Right) rejections that were made by Cond. (8) (ambiguity rejection). The data rejected by Cond. (7) appeared to look more chaotic than the one rejected by Cond. (8).