Generalizing Consistent Multi-Class Classification with Rejection to be Compatible with Arbitrary Losses

Anonymous Author(s)
Affiliation
Address
email

Abstract

Classification with rejection (CwR) refrains from making a prediction to avoid critical misclassification when encountering test samples that are difficult to classify. Though previous methods for CwR have been provided with theoretical guarantees, they are only compatible with certain loss functions, making them not flexible enough when the loss needs to be changed with the dataset in practice. In this paper, we derive a novel formulation for CwR that can be equipped with arbitrary loss functions while maintaining the theoretical guarantees. First, we show that K-class CwR is equivalent to a $(K+1)$-class classification problem on the original data distribution with an augmented class, and propose an empirical risk minimization formulation to solve this problem with an estimation error bound. Then, we find a necessary and sufficient condition for the learning *consistency* of the surrogates constructed on our proposed formulation equipped with any classification-calibrated multi-class losses, where consistency means the surrogate risk minimization implies the target risk minimization for CwR. Finally, experiments on benchmark datasets validate the effectiveness of our proposed method.

1 Introduction

In risk-sensitive multi-class classification applications (e.g., medical diagnosis, healthcare, autonomous driving, and product inspections [12, 21, 43]), misclassification can cause serious or even fatal consequences. To alleviate this issue, many studies have been conducted on *classification with rejection* (CwR) [10, 6, 61, 12, 13, 15, 21, 51, 47, 43, 8], which can abstain from making an unsure prediction to prevent such critical misclassification.

Most of the previous studies follow the framework that provides the reject option with a pre-defined cost c which is lower than the misclassification cost 1. Given cost c, the problem is further formulated as a risk minimization problem that aims to minimize the expectation of the zero-one-c loss, i.e., the zero-one-c risk. With the risk minimization process, the obtained classifier can balance the cost of rejection and prediction by choosing to incur a rejection cost c if the misclassification risk is high.

Due to the discontinuous nature of the zero-one-c loss, recent works focused on finding its continuous surrogates to make the optimization problem tractable. A basic requirement for surrogate losses is the *consistency* [63, 7, 53, 46], i.e., the surrogate risk minimization implies the zero-one-c risk minimization. Moreover, compared with the traditional K-class classification task where decisions are normally made from the index of the maximum coordinate of a K-dimensional scoring function, the design of decision criteria in the CwR task is more elusive due to the existence of a reject option.

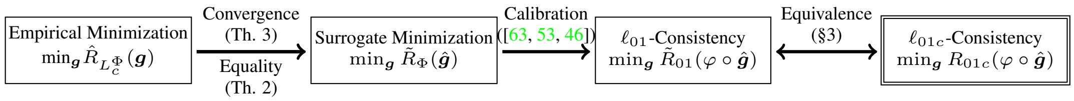

Figure 1: Overview of the construction of consistent surrogates for classification with rejection in this work.

By adopting different classification and rejection criteria, various surrogates of the zero-one-c loss have been proposed with consistency analyses [6, 61, 12, 13, 47, 43, 8].

Classical studies focused on developing *confidence-based methods*[6, 61, 47, 43], which use the outputs of classifiers as confidence values and set a real-valued threshold as the rejection rule. Representative methods [61, 43] used surrogates that depend on *class-posterior possibility estimation* (CPE) [49, 56], which is challenging when using deep models [24]. Though some of them [6, 47, 34, 23] could avoid CPE, most of them applied the modification of non-differentiable hinge/ramp-like surrogates, and their performance was only validated with linear models.

To avoid the use of the confidence threshold, Cortes et al. [12] provided an upper bound of the zero-one-c loss as the surrogate that allows the use of a separated rejector and can be trained simultaneously with the classifier, which is regarded as *classifier-rejector methods*. Though these methods achieved state-of-the-art performance in binary classification scenarios, they only provided a consistency guarantee for hinge-like and exponential losses and cannot be directly generalized to the multi-class scenario as shown in Ni et al. [43]. Charoenphakdee et al. [8] showed that K-class CwR can be decomposed into K binary cost-sensitive classification problems [16, 50, 11] and proposed a family of surrogates are the ensembles of arbitrary binary classification losses, which can avoid CPE and the use of confidence threshold with properly chosen losses when the cost function is constant. Mozannar and Sontag [40] provided a modified version of the cross entropy loss as the surrogate for the task of learning to defer [40, 41] that can also be used in CwR, while its optimal solution still relies on CPE. In summary, previous works only took limited types of losses into consideration, and there lacks a theoretically grounded framework that can cover all the surrogates used in multi-class classification.

In this paper, we propose a novel framework for CwR that allows the use of arbitrary surrogate losses used in traditional multi-class classification as long as they are classification-calibrated, including but not limited to the well-known cross entropy loss, mean absolute error, focal loss [32, 9], and the pairwise/one-versus-all generalizations of binary margin losses [63]. Thanks to the flexible choices of losses, we be free of the restricted analyses on the consistency of certain surrogates. An overview of our framework is shown in Figure 1. We summarize the main contributions of this work as follows:

- We disclose the equivalence between K-class CwR and a $(K+1)$-class classification problem on the original data distribution with an augmented class, by showing the equality between their classification risks.
- We propose a formulation of surrogates for ℓ_{01c} that can recover the surrogate risk of a $(K+1)$-class classification task only with the K-class training distribution, and derive an estimation error bound for its empirical risk minimization.
- We find a necessary and sufficient condition for the consistency of the proposed family of surrogates *w.r.t.* the zero-one-c loss that allows the use of any calibrated multi-class surrogates.
- We for the first time provide an analysis on the calibration of the *generalized cross entropy* loss [64] that benefits from both the cross entropy loss and mean absolute error, and experimentally demonstrate that it is suitable for our proposed framework.

2 Preliminaries

In this section, we provide preliminary knowledge of CwR and calibrated surrogate losses, and discuss the consistency in CwR.

2.1 Classification with Rejection

The problem setting of CwR is based on the cost-based framework [10]. Let us denote by $\mathcal{X}$ the feature space, $\mathcal{Y} = \{1, 2, \ldots, K\}$ the label space , and $\mathcal{Y}^{®} = \{1, 2, \ldots, K, ®\}$ the label space with a reject option. We are given instance-label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ independently and identically drawn from an underlying distribution with probability density $p(\boldsymbol{x}, y)$. The goal of CwR is to train a classifier $f : \mathcal{X} \to \mathcal{Y}^{®}$ that can abstain from making a decision, where ® denotes the reject option. The evaluation metric of this task is the zero-one-c loss ℓ_{01c}, which can be expressed as a variant of the traditional zero-one loss $\ell_{01}(f(\boldsymbol{x}), y) = [\![f(\boldsymbol{x}) \neq y]\!]$:

$$\ell_{01c}(f(\boldsymbol{x}), y) = \begin{cases} c, & f(\boldsymbol{x}) = ®, \\ [\![f(\boldsymbol{x}) \neq y]\!], & f(\boldsymbol{x}) \in \{1, 2, \ldots, k\}, \end{cases}$$

where $[\![\cdot]\!]$ is the Iverson bracket notation as suggested by Knuth [28] and the cost c can be further extended to an instance-dependent function $c(\boldsymbol{x})$. Our goal is to train a classifier that can minimize the expectation of ℓ_{01c} over the data distribution:

$$R_{01c}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell_{01c}(f(\boldsymbol{x}), y)]. \tag{1}$$

Let us denote by $f^* = \operatorname{argmin}_f R_{01c}(f)$ the Bayes optimal solution and $\boldsymbol{\eta}(\boldsymbol{x}) = \{p(y|\boldsymbol{x})\}_{y=1}^K$ the posterior probabilities. When evaluated by ℓ_{01c}, a classifier receives a standard classification error in $\{0, 1\}$ if it makes a prediction and a cost of c if it does not make a prediction (i.e., chooses the reject option). Intuitively, an optimal solution f^* should balance the possibility of misclassification and the rejection cost c. This explanation is theoretically justified by Chow's rule [10]:

Definition 1. (Chow's Rule) A classifier $f : \mathcal{X} \to \mathcal{Y}^{®}$ is the optimal solution of (1) if and only if it meets the following condition almost surely:

$$f(\boldsymbol{x}) = \begin{cases} ®, & \max_y \eta_y(\boldsymbol{x}) \leq 1 - c, \\ \operatorname{argmax}_y \eta_y(\boldsymbol{x}), & \text{else.} \end{cases}$$

Chow's rule shows that the optimal solution should refrain from making a decision if the most competent prediction of an example is still not confident enough given a rejection cost c.

2.2 Calibrated Surrogate Losses

Most classification problems can be formalized as the minimization of the target risk, which is the expectation of a target loss. Then *empirical risk minimization* (ERM) is conducted to obtain models with performance guarantees. However, most of the target losses are discontinuous, e.g., the zero-one loss in multi-class classification and the Hamming/ranking loss in multi-label classification [19]. Therefore, directly optimizing them is usually difficult and even NP-hard [17].

In order to optimize the target risk efficiently, surrogate risk minimization is preferred that minimizing the expectation of a continuous surrogate loss instead, e.g., the hinge loss in binary classification and the cross entropy loss in multi-class classification. For the statistical consistency of learning, *calibration* [52] is considered as a basic requirement for surrogate losses, which is a pointwise version of consistency and means that the minimization of the surrogate loss yields that of the target loss for each possible sample. A commonly adopted definition of the calibration of surrogates in multi-class classification is given as follows:

Definition 2. (ℓ_{01}-Calibration [7, 53, 46]) For a K-class classification problem with target loss ℓ_{01}, we say $\Phi : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$ is ℓ_{01}-calibrated if for any $\boldsymbol{p} \in \Delta^K$:

$$\inf_{\boldsymbol{u} \in \mathbb{R}^K, \boldsymbol{u} \notin \operatorname{argmin}_{\boldsymbol{u}} \boldsymbol{p}^T \boldsymbol{L}_{01}(\boldsymbol{u})} \boldsymbol{p}^T \boldsymbol{\Phi}(\boldsymbol{u}) > \inf_{\boldsymbol{u} \in \mathbb{R}^K} \boldsymbol{p}^T \boldsymbol{\Phi}(\boldsymbol{u}),$$

where $\boldsymbol{\Phi}(\boldsymbol{u}) = \{\Phi(\boldsymbol{u}, y)\}_{y=1}^K$, $\boldsymbol{L}_{01}(\boldsymbol{u}) = \{\ell_{01}(\operatorname{argmax}_{y' \in \mathcal{Y}} \boldsymbol{u}_{y'}, y)\}_{y=1}^K$.

The definition of ℓ_{01}-calibration requires that a surrogate loss should be able to distinguish between optimal solutions and non-optimal ones *w.r.t.* any potential posterior distribution $\boldsymbol{p}$. This property

Table 1: Comparisons between our proposed method and previous works of multi-class classification with rejection. Since our method is induced from a $(K+1)$-class classification problem, we can render a consistent learning guarantee with arbitrary surrogate losses that are calibrated *w.r.t.* the zero-one loss. Thanks to the abundant choices of losses, our proposed method can avoid CPE and the use of confidence thresholds.

Method	CPE-Free	Instance-Dependent Cost	Confidence Threshold-Free	Arbitrary Losses
[47]	✓	✓	×	×
[43]	×	✓	×	×
[40]	×	✓	✓	×
[8]	✓	×	✓	×
Proposed	✓	✓	✓	✓

is shown to be a necessary and sufficient condition for the statistical consistency of surrogate risk minimization, and fruitful research on the verification of ℓ_{01}-calibrated surrogates has been conducted [7, 63, 53, 46, 45, 18].

Besides multi-class classification, the calibration of surrogate losses also has been studied in various aspects of statistical learning, including but not limited to, multi-label classification [19, 62, 29, 57], AUC optimization [20, 38], general linear-fractional utility maximization [3], cost-sensitive learning [11, 50], top-K classification [31, 60], and adversarially robust classification [4, 2, 1].

2.3 Consistency in Classification with Rejection

In the field of CwR, we are also interested in the consistency of surrogate losses. Let $\mathcal{C} \subset \mathbb{R}^d$ where $d \in \mathbb{N}$ and $\Phi : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}_+$ is a surrogate loss, the consistency is defined as follows:

Definition 3. (ℓ_{01c}-Consistency) A surrogate loss $\Phi : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}_+$ is ℓ_{01c}-consistent if there exists a function $\varphi : \mathcal{C} \to \mathcal{Y}^{®}$ for all probability distributions and all the sequences of functions $\{\boldsymbol{g}_i\}_{i\in\mathbb{N}} : \mathcal{X} \to \mathcal{C}$:

$$R_\Phi(\boldsymbol{g}_i) \to R_\Phi^* \Rightarrow R_{01c}(\varphi \circ \boldsymbol{g}_i) \to R_{01c}^*, \tag{2}$$

where $R_\phi(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)}[\Phi(\boldsymbol{g}(\boldsymbol{x}), y)]$, $R_\Phi^* = \inf_{\boldsymbol{g}:\mathcal{X}\to\mathcal{C}} R_\Phi(\boldsymbol{g})$, and $R_{01c}^* = \inf_{f:\mathcal{X}\to\mathcal{Y}^{®}} R_{01c}(f)$.

This definition is inspired by the problem of general multi-class classification [46]. For an ℓ_{01c}-consistent surrogate loss Φ, we can safely minimize the surrogate risk R_Φ instead while remaining the consistency guarantee of R_{01c}.

To ensure the consistency of Φ, it is routine to discuss the calibration of surrogate losses. However, unlike the classical multi-class classification problem, where φ is usually an argmax operator, the design of φ in the field of CwR can be quite complicated and hard to be unified, which makes it difficult to directly conduct calibration analysis on Φ. The flexibility of φ also limits the discussions to specific types of surrogate losses. In Ramaswamy et al. [47], the authors considered the multi-class extensions of the hinge-loss with a confidence threshold. Ni et al. [43] indicated that the confidence-based method is indispensable and only focuses on class probability estimation via surrogate risk minimization. Both of Mozannar and Sontag [40] and Charoenphakdee et al. [8] gave surrogate losses for the zero-one-c loss that does not depend on the accurate estimation of the class probability, while Mozannar and Sontag [40] focused on a variant of the cross entropy loss and Charoenphakdee et al. [8] constructed calibrated surrogate losses with the ensemble of K calibrated losses for binary classification.

In this paper, instead of directly discussing the calibration of surrogate Φ, we show that there is an equivalence between classical multi-class classification and CwR. Based on this equivalence, we show that it is sufficient for Φ to be ℓ_{01c}-consistent by letting it be a simple variant of **any** calibrated surrogate loss *w.r.t.* the traditional zero-one loss ℓ_{01}. The comparison of the proposed method and related works is shown in Table 1.

3 Equivalence between Classification with Rejection and Ordinary Classification

In this section, we first show that the risk $R_{01c}(f)$ can be formalized as a $(K+1)$-class classification problem, and show that we can obtain ℓ_{01c}-consistent surrogates with a variant of any calibrated surrogate *w.r.t.* ℓ_{01}, which enables the use of $\mathcal{C} \subset \mathbb{R}^{K+1}$ and $\varphi(\cdot) = \mathrm{argmax}(\cdot)$ as in the traditional multi-class classification tasks. We also show that such equivalence also holds when the cost c depends on sample $\boldsymbol{x}$. The proof of the conclusions in this section can be found in Appendix A.

We start by considering the following distribution $\mathcal{D}_c^{®}$ over $\mathcal{X} \times \mathcal{Y}^{®}$ with probability density $\tilde{p}(\boldsymbol{x}, \tilde{y})$:

Definition 4. (Self-Augmented Distribution) A distribution $\mathcal{D}_c^{®}$ is called a c-self-augmented distribution $w.r.t.$ $\mathcal{D}$ if its probability density meets the following conditions:

$$\tilde{p}(\boldsymbol{x}, \tilde{y}) = \begin{cases} \frac{p(\boldsymbol{x}, y)}{2-c}, & \tilde{y} \in \{1, 2, \ldots, K\}, \\ \frac{(1-c)p(\boldsymbol{x})}{2-c}, & \tilde{y} = ®. \end{cases}$$

It can be seen that distribution $\mathcal{D}_c^{®}$ shares the same marginal density of $\boldsymbol{x}$ as the original distribution $\mathcal{D}$ while $\mathcal{D}_c^{®}$ has an augmented class ® with class possibility determined by the rejection cost c. Based on the connection between $\mathcal{D}_c^{®}$ and $\mathcal{D}$, we can further explore the relation between the two tasks: classification on $\mathcal{D}_c^{®}$ and CwR on $\mathcal{D}$.

Theorem 1. For any classifier $f : \mathcal{X} \to \mathcal{Y}^{®}$, the following equation holds:

$$R_{01c}(f) - R_{01c}^{*} = (2-c)\left(\tilde{R}_{01}(f) - \tilde{R}_{01}^{*}\right),$$

where $\tilde{R}_{01}(f) = \mathbb{E}_{\tilde{p}(\boldsymbol{x},\tilde{y})}[\ell_{01}(f(\boldsymbol{x}), \tilde{y})]$ and $\tilde{R}_{01}^{*} = \inf_{f:\mathcal{X}\to\mathcal{Y}^{®}} \tilde{R}_{01}(f)$.

This equation reveals the equivalence between the two tasks in a straightforward manner. Since the multiplication of the classification risk on $\mathcal{D}_c^{®}$ with a positive constant is equal to $R_{01c}(f)$, the minimization of $\tilde{R}_{01}(f)$ immediately yields the minimization of $R_{01c}(f)$ and vice versa. Furthermore, according to the linear correlation between $\tilde{R}_{01}(f)$ and $R_{01c}(f)$, we can directly quantify the excess error $R_{01c}(f) - R_{01c}^{*}$ by bounding $\tilde{R}_{01}(f) - \tilde{R}_{01}^{c}$, which is an easier work thanks to the existing research of multi-class classification. In conclusion, risk minimization with $\tilde{R}_{01}(f)$ can also give a classifier with a rejection option with the optimality guarantee, and then we can consider a surrogate risk minimization problem for multi-class classification instead of CwR.

When the cost $c(\boldsymbol{x})$ is an instance-dependent function, we show that such equivalence still holds with a minor modification. Considering the reweighted zero-one loss: $\bar{\ell}_{01}(f(\boldsymbol{x}), y) = (2-c(\boldsymbol{x}))[\![f(\boldsymbol{x}) \neq y]\!]$ and its expectation $\bar{R}_{01}(f)$ on $\mathcal{D}_c^{®}$, we have the following conclusion:

Corollary 1. For any classifier $f : \mathcal{X} \to \mathcal{Y}^{®}$, the following inequalities holds:

$$\bar{R}_{01}(f) - \bar{R}_{01}^{*} = R_{01c}(f) - R_{01c}^{*}.$$

It is obvious that Lemma 1 is a special case of Lemma 1 with constant cost functions. Though here we consider a reweighted classification task, the calibration result of multi-class surrogate losses can still be applied without any modification since the minimization of $\bar{R}_{01}(f)$ can be seen as ordinary classification risk minimization with a slightly different marginal density $p'(\boldsymbol{x})$, which does not affect the calibration result since the class-posterior possibilities remain unchanged. All the conclusions in the rest of this paper can be extended to the scenario of instance-dependent cost and we provide them in Appendix G.

4 ℓ_{01c}-Consistent Surrogates with Arbitrary ℓ_{01}-Calibrated Losses

According to the discussions in Section 3, CwR can be safely replaced by multi-class classification on a special distribution $\mathcal{D}_c^{®}$. Following the practice of surrogate risk minimization in multi-class classification, we can replace the zero-one loss ℓ_{01} with a surrogate risk $\Phi : \mathbb{R}^{K+1} \times \mathcal{Y} \cup \{K+1\} \to$

$\mathbb{R}_+$ and minimizing the surrogate risk with a score-based classifier $\boldsymbol{g}: \mathcal{X} \to \mathbb{R}^{K+1}$ instead, which is defined as follows:

$$\tilde{R}_\Phi(\boldsymbol{g}) = \mathbb{E}_{\tilde{p}(\boldsymbol{x},\tilde{y})}[\Phi(\boldsymbol{g}(\boldsymbol{x}), t(\tilde{y}))], \tag{3}$$

where $t(\tilde{y}) = K+1$ if $\tilde{y} = ®$ and $t(\tilde{y}) = \tilde{y}$ otherwise. (3) is a typical formulation of the multi-class classification risk and we can asymptotically minimize it following the ERM framework [54]. After the risk minimization process, the prediction is generated with the following link function $\varphi: \mathbb{R}^{K+1} \to \mathcal{Y}^{®}$:

$$\varphi(\boldsymbol{u}) = \begin{cases} ®, & \underset{y \in \mathcal{Y} \cup \{K+1\}}{\operatorname{argmax}} \quad \boldsymbol{u}_y(\boldsymbol{x}) = K+1, \\ \underset{y \in \mathcal{Y} \cup \{K+1\}}{\operatorname{argmax}} \quad \boldsymbol{u}_y(\boldsymbol{x}), & \text{else.} \end{cases}$$

With a properly chosen surrogate Φ, the minimization of $\tilde{R}_\Phi(\boldsymbol{g})$ can lead to that of $\tilde{R}_{01}(\varphi(\boldsymbol{g}))$, which indicates the minimization of $R_{01c}(\varphi\boldsymbol{g})$ according to Lemmas 1 and 1. The theory of how to find such surrogates has been thoroughly studied in the field of the classification-calibration of multi-class surrogates [63, 53, 46].

However, we do not have direct access toward $\mathcal{D}_c^{®}$ though it is closely related to the available data distribution $\mathcal{D}$. In this section, we propose a family of surrogate losses based on the conclusions in the previous section, which allows the use of any multi-class classification surrogates. With this formulation of surrogates, we can recover the classification risk of $\tilde{R}_\Phi(\boldsymbol{g})$ without access to $\mathcal{D}_c^{®}$ by taking its expectation in $\mathcal{D}$. Based on the loss formulation, we also provide the estimation error bound to show the validity of ERM.

4.1 Formulation of Surrogates

Here, we begin with the definition of a family of surrogates for the zero-one-c loss, and then show how it can relate $\mathcal{D}$ and $\mathcal{D}_c^{®}$. With any multi-class classification loss Φ, we have the following formulation of surrogate losses for ℓ_{01c}:

Definition 5. Given a pre-defined rejection cost c, we have the following formulation of surrogate $L_c^\Phi: \mathbb{R}^{K+1} \times \mathcal{Y} \to \mathbb{R}_+$ for CwR:

$$L_c^\Phi(\boldsymbol{u}, y) = \Phi(\boldsymbol{u}, y) + (1-c)\Phi(\boldsymbol{u}, K+1), \tag{4}$$

where $\Phi: \mathbb{R}^{K+1} \times \mathcal{Y} \cup \{K+1\} \to \mathbb{R}_+$ and $\boldsymbol{u} \in \mathbb{R}^{K+1}$.

The proposed surrogate loss is the linear combination of a $(K+1)$ dimensional multi-class classification loss with coefficient determined by the predefined cost c. It can also be learned from Appendix A.2 of Charoenphakdee et al. [8] that when Φ is the softmax cross entropy loss, (4) is equivalent to Mozannar and Sontag [40]. The following theorem reveals the connection between $\tilde{R}_\Phi(\boldsymbol{g})$ and the expectation of L_c^Φ on $\mathcal{D}$:

Theorem 2. For any $\boldsymbol{g}: \mathcal{X} \to \mathbb{R}^{K+1}$ and $R_{L_c^\Phi}(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)}[L_c^\Phi(\boldsymbol{g}(\boldsymbol{x}), y)]$:

$$R_{L_c^\Phi}(\boldsymbol{g}) = (2-c)\tilde{R}_\Phi(\boldsymbol{g}).$$

The proof is provided in Appendix B. From Theorem 2, we can obtain the risk $\tilde{R}_\Phi(\boldsymbol{g})$ without access to $\mathcal{D}_c^{®}$ with the use of the proposed surrogate (4). Following the common practice, we can finally conduct ERM [54] that minimizes the unbiased estimator of $R_{L_c^\Phi}(\boldsymbol{g})$, which is also that of $\tilde{R}_\Phi(\boldsymbol{g})$ according to Theorem 2:

$$\hat{R}_{L_c^\Phi}(\boldsymbol{g}) = \frac{1}{n}\sum\nolimits_{i=1}^{n} L_c^\Phi(\boldsymbol{g}(\boldsymbol{x}_i), y_i) \tag{5}$$

After minimizing $\hat{R}_{L_c^\Phi}(\boldsymbol{g})$ and obtaining the empirically optimal $\hat{\boldsymbol{g}}$, we can use it for predicting with the link function $\varphi \circ \hat{\boldsymbol{g}}$, where $\varphi \circ \hat{\boldsymbol{g}}(\boldsymbol{x}) = \varphi(\hat{\boldsymbol{g}}(\boldsymbol{x}))$.

According to the unbiasedness of (3), it is promising that the induced prediction rule $\varphi \circ \hat{\boldsymbol{g}}$ can approximate Chow's rule (Definition 1). To quantify such approximation, there remains two questions: what is the relation between the minimization of empirical risk $\hat{R}_{L_c^\Phi}(\boldsymbol{g})$ and $R_{L_c^\Phi}(\boldsymbol{g})$, and whether the minimization of $R_{L_c^\Phi}(\boldsymbol{g})$ yields that of $R_{01c}(\varphi \circ \boldsymbol{g})$. We will answer the two problems in Section 4.2 and Section 5, respectively.

4.2 Estimation Error Bound

In Section 4.1, we proposed a family of surrogates that can recover the surrogate risk on $\mathcal{D}_c^{®}$ with only $\mathcal{D}$ and provided an ERM framework to learn the empirically optimal $\hat{\boldsymbol{g}}$. Here we further justify the use of ERM by showing that the minimization of $\hat{R}_{L_c^\Phi}$ can also result in that of $R_{L_c^\Phi}$ with the following estimation error bound.

Theorem 3. For any $\delta \in (0, 1)$, suppose the model class of g_y is $\mathcal{G}_y$ and $\boldsymbol{g} \in \mathcal{G}$, where $\mathcal{G}_y \subset \mathcal{X} \to \mathbb{R}$ and $\mathcal{G} \subset \mathcal{X} \to \mathbb{R}^{K+1}$ is composed of $\{\mathcal{G}_y\}_{y=1}^{K+1}$. $\Phi(\cdot, y)$ is ρ-Lipschitz continuous and is bounded by $C_\Phi > 0$. Assume that the identifiable condition holds, i.e., $\min_{\boldsymbol{g}\in\mathcal{G}} R_{L_c^\Phi}(\boldsymbol{g}) = R^*_{L_c^\Phi}$, then the following inequality holds with probability at least $1 - \delta$:

$$R_{L_c^\Phi}(\hat{\boldsymbol{g}}) - R^*_{L_c^\Phi} \leq 4\sqrt{2}(2-c)\rho \sum\nolimits_{y=1}^{K+1} \mathfrak{R}_n(\mathcal{G}_y) + (2-c)C_\Phi \sqrt{\frac{2\log 2/\delta}{n}}, \tag{6}$$

where $\mathfrak{R}_n(\mathcal{G}_y)$ is the Rademacher complexity [5] *w.r.t.* $\mathcal{G}_y$ on the distribution with density $p(\boldsymbol{x})$ that often decays in the rate of $O(\frac{1}{\sqrt{n}})$.

We prove this conclusion in Appendix C. From the theorem above, we can learn that with the identifiable condition which is a common assumption with the use of complex models [4, 26, 33], $R_{L_c^\Phi}(\hat{\boldsymbol{g}})$ converges to $R^*_{L_c^\Phi}$ in $O_p(1/\sqrt{n})$, which is the optimal parametric convergence rate without additional assumptions [37]. According to Theorem 2, it is straightforward that $\tilde{R}_\Phi(\boldsymbol{g}) \xrightarrow{P} \tilde{R}^*_\Phi$ also holds. Nevertheless, the relation between the minimization of surrogate risk $\tilde{R}_\Phi(\boldsymbol{g})$ and that of the target risk $\tilde{R}_{01}(\varphi \circ \boldsymbol{g})$ is still unknown. According to Lemma 1, the minimization of $\tilde{R}_{01}(\varphi \circ \boldsymbol{g})$ is equivalent to zero-one-c risk minimization, which is the goal of CwR. We answer this question in the next section by giving a necessary and sufficient condition for the ℓ_{01c}-consistency for L_c^Φ.

5 Theoretical Analysis

In this section, we first point out the necessary and sufficient condition for L_c^Φ to be ℓ_{01c}-calibrated. Then we further specify the regret transfer bounds for a family of CPE-free surrogates [64], which has not been provided with theoretical analysis before.

5.1 Necessary and Sufficient Condition for ℓ_{01c}-Consistency

Given the loss formulation (4), a natural idea is to construct surrogate L_c^Φ with commonly used multi-class loss functions. However, the ℓ_{01c}-consistency of such surrogates still remains unchecked. Here, we show that we can borrow the calibration analyses of multi-class surrogates and set Φ to any $(K+1)$-class ℓ_{01}-calibrated surrogates according to the following necessary and sufficient condition:

Theorem 4. L_c^Φ is ℓ_{01c}-consistent for any $c \in [0, 1]$ if and only if Φ is an ℓ_{01}-calibrated surrogate loss.

The complete proof is shown in Appendix D and here we provide its sketch. The equivalence between CwR and multi-class classification on $\mathcal{D}_c$ shown in Lemmas 1, 1, and Theorem 2 directly yields the sufficiency of this condition. Though the equivalent classification problem is limited on $\mathcal{D}_c$, $\tilde{p}(y|\boldsymbol{x})$ can be any valid class-posterior probabilities due to the arbitrariness of c and thus the calibration of Φ is necessary.

As a result, we can use any Φ in an off-the-shelf manner, i.e., to the consistency of different L_c^Φ, we only have to check if Φ is ℓ_{01}-calibrated, which has been studied thoroughly [7, 53, 46], instead of tedious case-based discussions. Furthermore, there is also no need for the consideration of any other potential Φ since ℓ_{01}-calibration is also necessary.

5.2 Calibration Result for Generalized Cross Entropy Loss

Given the necessary and sufficient condition for ℓ_{01c}-consistency, we can construct L_c^{Φ} with any ℓ_{01}-calibrated surrogates. However, it has been shown in Charoenphakdee et al. [8] that it can lead to a model that rejects more data than necessary if the cross entropy (CE) loss is used as Φ, which is a popular choice as a surrogate. Another common surrogate is the mean absolute error (MAE). Though it can avoid CPE and only focus on the crucial class with the maximum posterior possibility, it usually takes more training epochs before convergence [64], which can be costly in practical use.

Here, we consider the *generalized cross entropy* (GCE) loss [64] that can take the advantages of the CE loss and MAE, which is defined as below:

Definition 6. (Generalized cross entropy losses) For any $\gamma \in (0, 1]$, the GCE loss is defined as below:

$$\Phi_\gamma(\boldsymbol{g}(\boldsymbol{x}), y) = (1 - S(\boldsymbol{g})_y^\gamma)/\gamma,$$

where $S(\cdot)$ is the softmax-transformation.

It can be seen that the loss formulation is equivalent to MAE if $\gamma = 1$ and it is also reported in Zhang and Sabuncu [64] that the GCE loss can approximate the CE loss if $\gamma \to 0$. Though the GCE loss has proved to be effective in practical use, to the best of our knowledge, its calibration results remain unknown, and thus it is unsafe directly combining it with L_c^{Φ}.

Theorem 5. The GCE loss Φ^γ is ℓ_{01}-calibrated for any $\gamma \in (0, 1]$. For the optimal model $\boldsymbol{g}^*$, $S(\boldsymbol{g}^*)_y = \eta_y^{\frac{1}{1-\gamma}} / \sum_{y'=1}^{K} \eta_{y'}^{\frac{1}{1-\gamma}}$ for all the $\boldsymbol{x} \in \mathcal{X}$ almost surely if $\gamma \in (0, 1)$. If $\gamma = 1$, $S(\boldsymbol{g}^*)_{\mathrm{argmax}_y \eta_y} = 1$.

The proof can be found in Appendix E. After verifying the calibration result of the GCE loss, we can combine it with the loss formulation L_c^{Φ} and obtain an ℓ_{01c}-consistent surrogate. We will experimentally demonstrate its effectiveness in the next section.

6 Experiments

In this section, we provide the experiment results of CwR with deep models, which are evaluated by the zero-one-c loss following the common practice [43, 8]. We also show the misclassification rate of the accepted data and the ratio of the rejected data. Details of the setup and the experiments for instance-dependent cost can be found in Appendix F and G, respectively.

Datasets and Models. In the experiments, we evaluate the proposed methods and baselines on three widely-used benchmarks Fashion-MNIST [58], SVHN [42], CIFAR-10 [30] with cost c selected from $\{0.05, 0.06, 0.07, 0.08, 0.09, 0.10\}$ for Fashion-MNIST and $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ the other two. We conduct data augmentation for CIFAR-10 and use the original datasets of Fashion-MNIST and SVHN in the experiments. For Fashion-MNIST, we use a CNN defined in Charoenphakdee et al. [8], and ResNet-18 and ResNet-34 [25] are used for SVHN and CIFAR-10, respectively.

Baselines. We compare our method with state-of-the-art methods in CwR, including confidence-based cross entropy loss (CE) [43], learning to defer (DEFER) [40], and cost-sensitive learning-based method with sigmoid loss (CS) [8], in which DEFER is a special case of our method that use cross entropy loss as Φ. For CE, we also conduct the temperature scaling [24] to alleviate overconfidence . For the proposed method, we use GCE with default parameter $\gamma = 0.7$ as suggested in Zhang and Sabuncu [64] and pairwise-sigmoid (Sigmoid) loss [63] to construct the surrogate L_c^{Φ}.

We implemented all the methods by Pytorch [44], and conducted all the experiments on NVIDIA GeForce 3090 GPUs.

Table 2: The mean and standard error of the zero-one-c losses (rescaled to 0-100), rejection ratio, and missclassification rates of the accepted data for 5 trails. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Method	Cost	CE			CS			DEFER			GCE			Sigmoid		
		01c	Rej	01	01c	Rej	01	01c	Rej	01	01c	Rej	01	01c	Rej	01
FMNIST	0.05	2.30 (0.07)	25.17 (3.17)	1.39 (0.11)	2.93 (0.25)	34.95 (1.94)	1.81 (0.48)	3.79 (0.28)	50.461 (2.51)	2.58 (0.46)	3.22 (0.07)	50.47 (2.49)	1.39 (0.30)	**2.23 (0.01)**	30.98 (0.62)	0.99 (0.05)
	0.06	**2.58 (0.07)**	22.92 (1.45)	1.56 (0.09)	3.37 (0.15)	33.13 (1.27)	2.07 (0.28)	4.63 (0.10)	56.45 (3.69)	2.84 (0.42)	3.78 (0.17)	50.46 (1.24)	1.53 (0.30)	**2.62 (0.08)**	26.76 (3.37)	1.37 (0.21)
	0.07	**2.73 (0.14)**	21.17 (2.23)	1.58 (0.31)	3.45 (0.17)	35.77 (2.62)	1.47 (0.04)	5.18 (0.47)	56.46 (6.85)	2.86 (0.41)	4.23 (0.21)	48.05 (5.24)	1.66 (0.25)	2.94 (0.07)	29.87 (0.85)	1.21 (0.17)
	0.08	**3.12 (0.11)**	20.71 (1.68)	1.85 (0.07)	4.13 (0.36)	33.68 (0.32)	2.17 (0.52)	5.86 (0.30)	54.08 (3.47)	3.36 (0.29)	4.50 (0.06)	45.66 (2.36)	1.55 (0.25)	**3.14 (0.17)**	26.10 (0.23)	1.43 (0.25)
	0.09	**3.55 (0.21)**	23.64 (1.82)	1.86 (0.18)	4.20 (0.21)	31.90 (1.74)	1.96 (0.15)	6.31 (0.40)	54.62 (4.33)	3.09 (0.49)	4.95 (0.06)	44.05 (1.74)	1.77 (0.23)	**3.50 (0.05)**	23.71 (.28)	1.79 (0.18)
	0.10	**3.59 (0.16)**	18.32 (1.56)	2.15 (0.32)	4.45 (0.20)	28.96 (0.13)	2.18 (0.41)	6.72 (0.07)	52.69 (0.74)	3.08 (0.18)	5.06 (0.23)	39.01 (4.87)	1.89 (0.26)	**3.73 (0.05)**	23.96 (1.90)	1.76 (0.20)
SVHN	0.05	3.33 (0.14)	14.37 (0.94)	3.05 (0.13)	4.42 (0.13)	12.81 (0.14)	4.33 (0.12)	4.19 (0.29)	33.05 (1.59)	3.80 (0.37)	**2.68 (0.17)**	19.79 (0.72)	2.10 (0.24)	**2.70 (0.14)**	29.56 (1.16)	1.73 (0.17)
	0.10	4.66 (0.20)	10.91 (0.57)	4.01 (0.21)	4.48 (0.14)	12.85 (0.42)	3.67 (0.11)	5.55 (0.56)	30.72 (2.64)	3.58 (0.52)	**4.13 (0.11)**	14.83 (0.54)	3.10 (0.10)	**4.13 (0.39)**	19.16 (1.94)	2.74 (0.43)
	0.15	5.40 (0.09)	8.52 (0.15)	4.50 (0.07)	5.14 (0.10)	13.21 (0.62)	3.64 (0.19)	6.37 (0.21)	21.19 (0.94)	4.05 (0.25)	**4.66 (0.06)**	11.47 (0.41)	3.31 (0.07)	**4.83 (0.44)**	18.38 (1.37)	2.54 (0.61)
	0.20	6.16 (0.13)	7.74 (0.26)	4.99 (0.09)	**5.51 (0.20)**	12.78 (1.03)	3.19 (0.24)	5.99 (0.17)	12.33 (0.51)	4.02 (0.16)	**5.44 (0.04)**	10.02 (0.25)	3.82 (0.03)	6.39 (0.45)	15.86 (0.72)	3.82 (0.48)
	0.25	7.08 (0.32)	6.51 (1.06)	5.83 (0.36)	6.77 (0.16)	12.96 (0.97)	4.06 (0.18)	6.69 (0.16)	9.18 (0.35)	4.33 (0.16)	**5.75 (0.14)**	8.64 (0.20)	3.93 (0.12)	6.74 (0.13)	13.79 (0.33)	3.82 (0.14)
	0.30	7.12 (0.16)	5.31 (0.36)	5.83 (0.18)	7.26 (0.33)	13.21 (1.20)	3.80 (0.41)	7.07 (0.31)	12.35 (2.31)	4.55 (0.34)	**6.30 (0.09)**	8.72 (0.11)	4.04 (0.09)	7.69 (0.22)	10.79 (0.76)	5.00 (0.13)
CIFAR-10	0.05	4.43 (0.23)	29.93 (1.85)	4.18 (0.33)	6.59 (0.27)	20.20 (0.51)	7.00 (0.35)	4.62 (0.47)	44.97 (5.24)	4.30 (0.88)	3.80 (0.20)	34.52 (2.77)	3.16 (0.35)	**3.67 (0.03)**	42.69 (8.74)	2.63 (0.49)
	0.10	7.13 (0.11)	21.13 (0.81)	6.35 (0.18)	7.68 (0.32)	20.31 (0.66)	7.08 (0.42)	6.56 (0.26)	26.21 (1.12)	5.34 (0.39)	**5.84 (0.12)**	25.47 (0.98)	4.41 (0.15)	**6.11 (0.13)**	31.66 (2.17)	4.30 (0.30)
	0.15	9.03 (0.32)	7.76 (0.39)	7.74 (0.37)	8.35 (0.29)	21.83 (0.92)	6.49 (0.45)	8.39 (0.19)	20.39 (1.59)	6.69 (0.35)	**7.56 (0.14)**	20.43 (0.60)	5.65 (0.23)	8.18 (0.10)	23.39 (0.82)	6.10 (0.18)
	0.20	10.45 (0.29)	14.53 (0.47)	8.82 (0.38)	**9.32 (0.21)**	21.86 (0.46)	6.33 (0.33)	9.65 (0.14)	17.16 (1.04)	7.50 (0.11)	**9.09 (0.14)**	18.45 (1.93)	6.62 (0.42)	9.69 (0.15)	19.54 (1.55)	7.20 (0.07)
	0.25	11.64 (0.26)	11.20 (0.30)	9.96 (0.32)	**10.46 (0.24)**	22.02 (0.40)	6.35 (0.35)	10.85 (0.08)	14.22 (1.35)	8.50 (0.30)	**10.31 (0.23)**	15.39 (1.47)	7.64 (0.38)	10.96 (0.11)	14.99 (1.71)	8.48 (0.40)
	0.30	12.20 (0.18)	10.02 (0.53)	10.89 (0.15)	**11.43 (0.23)**	22.23 (0.81)	6.13 (0.24)	11.90 (0.17)	11.48 (0.75)	9.55 (0.31)	**11.23 (0.16)**	12.52 (0.122)	8.55 (0.14)	12.14 (0.12)	11.08 (0.60)	9.91 (0.25)

Experimental Results. As can be seen from the experimental results reported in Table 2, our proposed method (i.e., either GCE or Sigmoid) significantly outperforms other compared methods in most cases. Obviously, for all the datasets and cost c, our GCE method outperforms the baseline DEFER method, which indicates that CwR cannot be simply solved by the methods used for learning to defer. It can be also seen that confidence-based CE is only comparable to the proposed method on FMNIST with a simple CNN. When complex models are used, the effect of overconfidence is inevitable even with the use of temperature scaling, which can be induced from the fact that CE often rejects less data than GCE on SVHN and CIFAR-10. Though CS is comparable to GCE on CIFAR-10 when the rejection cost is high, its performance degrades drastically when the classification cost decreases, which shows that it is not the best choice in highly error-critical tasks. When ResNet-18 and ResNet-34 are used on SVHN and CIFAR-10 respectively, our GCE method outperforms or is comparable to all the baselines, which shows that GCE is more stable on complex models. Our proposed Sigmoid method performs better than most baselines and is comparable to CE with the use of a simple CNN model, which aligns with the existing observations that pairwise losses are often effective with simple models [55, 14]. These results show that our method can benefit from the flexibility of the choices of loss functions.

7 Conclusion

In this paper, we studied the problem of classification with rejection, which can refrain from making a prediction to avoid critical misclassification. We derived a novel formulation for CwR that can be equipped with arbitrary loss functions while maintaining the theoretical guarantees, making them highly adaptive to the dataset in practical use. First, we showed the equivalence between K-class CwR and a $(K+1)$-class classification problem, and proposed an empirical risk minimization formulation to solve this problem with an estimation error bound. Then, we pointed out a necessary and sufficient condition for the learning consistency of the surrogates constructed on our proposed formulation equipped with any classification-calibrated multi-class losses. Finally, experimental results demonstrated the effectiveness of our proposed method.

References

[1] Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *CoRR*, abs/2104.09658, 2021. URL https://arxiv.org/abs/2104.09658.

[2] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. A finer calibration analysis for adversarial robustness. *CoRR*, abs/2105.01550, 2021. URL https://arxiv.org/abs/2105.01550.

[3] Han Bao and Masashi Sugiyama. Calibrated surrogate maximization of linear-fractional utility in binary classification. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 2337–2347. PMLR, 2020.

[4] Han Bao, Clayton Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 408–451. PMLR, 2020.

[5] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002. URL http://jmlr.org/papers/v3/bartlett02a.html.

[6] Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, 2008.

[7] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[8] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1507–1517. PMLR, 2021.

[9] Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *CVPR*, pages 5202–5211, 2021.

[10] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

[11] Scott Clayton. Calibrated asymmetric surrogate losses. *Electronic Journal of Stats*, 6:238–238, 2012.

[12] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *NeurIPS*, pages 1660–1668, 2016.

[13] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *ALT*, volume 9925, pages 67–82, 2016.

[14] Ürün Dogan, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *J. Mach. Learn. Res.*, 17:45:1–45:32, 2016.

[15] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010.

[16] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, pages 973–978. Morgan Kaufmann, 2001.

[17] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM J. Comput.*, 41(6):1558–1590, 2012.

[18] Jessica Finocchiaro, Rafael M. Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *NeurIPS*, pages 10780–10790, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/9ec51f6eb240fb631a35864e13737bca-Abstract.html.

[19] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. *Artif. Intell.*, 199-200:22–44, 2013.

[20] Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945. AAAI Press, 2015.

[21] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, pages 4878–4887, 2017.

[22] Gustavo L. Gilardoni. On pinsker's and vajda's type inequalities for csiszár's f-divergences. *IEEE Trans. Inf. Theory*, 56(11):5377–5386, 2010. doi: 10.1109/TIT.2010.2068710.

[23] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support vector machines with a reject option. In *NeurIPS*, pages 537–544, 2008.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 463–469, 2016.

[26] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pages 448–456. JMLR.org, 2015.

[28] D. E. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, 1992.

[29] Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent multilabel classification. In *NeurIPS*, pages 3321–3329, 2015.

[30] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[31] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40 (7):1533–1554, 2018.

[32] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017.

[33] Nan Lu, Shida Lei, Gang Niu, Issei Sato, and Masashi Sugiyama. Binary classification from multiple unlabeled datasets via surrogate set classification. In *ICML*, 2021.

[34] Naresh Manwani, Kalpit Desai, Sanand Sasidharan, and Ramasubramanian Sundararajan. Double ramp loss based reject option classifier. In *PAKDD*, volume 9077, pages 151–163, 2015.

[35] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *ALT*, pages 3–17, 2016.

[36] C. Mcdiarmid. *On the method of bounded differences*. Surveys in Combinatorics, 1989.

[37] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inf. Theory*, 54(8):3797–3803, 2008. doi: 10.1109/TIT.2008.926323. URL https://doi.org/10.1109/TIT.2008.926323.

[38] Aditya Krishna Menon and Robert C. Williamson. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 68–106. JMLR.org, 2014. URL http://proceedings.mlr.press/v35/menon14.html.

[39] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 978-0-262-01825-8.

[40] Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 2020.

[41] Hussein Mozannar, Arvind Satyanarayan, and David A. Sontag. Teaching humans when to defer to a classifier via examplars. *CoRR*, abs/2111.11297, 2021. URL https://arxiv.org/abs/2111.11297.

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 1 2011.

[43] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *NeurIPS*, pages 2582–2592, 2019.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[45] Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *CoRR*, abs/1609.06385, 2016. URL http://arxiv.org/abs/1609.06385.

[46] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *J. Mach. Learn. Res.*, 17:14:1–14:45, 2016. URL http://jmlr.org/papers/v17/14-316.html.

[47] Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554, 2018.

[48] Alexey E. Rastegin. Bounds of the pinsker and fannes types on the tsallis relative entropy. *Mathematical Physics Analysis & Geometry*, 16(3):213–228, 2013.

[49] Mark D. Reid and Robert C. Williamson. Composite binary losses. *J. Mach. Learn. Res.*, 11: 2387–2422, 2010.

[50] Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 153–160. Omnipress, 2011. URL https://icml.cc/2011/papers/138_icmlpaper.pdf.

[51] Song-Qing Shen, Bin-Bin Yang, and Wei Gao. AUC optimization with a reject option. In *AAAI*, pages 5684–5691, 2020.

[52] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

[53] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *J. Mach. Learn. Res.*, 8:1007–1025, 2007.

[54] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

[55] Yutong Wang and Clayton Scott. Weston-watkins hinge loss and ordered partitions. In *NeurIPS*, 2020.

[56] Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *J. Mach. Learn. Res.*, 17:223:1–223:52, 2016.

[57] Guoqiang Wu, Chongxuan Li, Kun Xu, and Jun Zhu. Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization. *CoRR*, abs/2105.05026, 2021. URL https://arxiv.org/abs/2105.05026.

[58] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

[59] F. Y. Yang. Maximum lq-likelihood estimation. *Annals of Statistics*, 2010.

[60] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 10727–10735. PMLR, 2020.

[61] Ming Yuan and Marten H. Wegkamp. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.*, 11:111–130, 2010.

[62] Mingyuan Zhang, Harish Guruprasad Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 11246–11255. PMLR, 2020.

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

[64] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8792–8802, 2018.

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
    (b) Did you describe the limitations of your work? [Yes] See Appendix H.
    (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix H.
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [Yes]
    (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6 and Appendix.
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 2.
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.
    (b) Did you mention the license of the assets? [N/A] The used datasets are open benchmarks.
    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please refer to the supplemental materials.
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 1 and Corollary 1

We begin with the proof of Corollary 1 and show that Theorem 1 is its special case.

Proof. First of all, we prove that the Bayes optimal solution on $\tilde{p}(\boldsymbol{x}, \tilde{y})$ coincide with the Chow's rule of $p(\boldsymbol{x}, y)$ with cost c. According to the optimality condition of multi-class classification, the optimal classifier $f^*(\boldsymbol{x})$ on $\tilde{p}(\boldsymbol{x}, \tilde{y})$ should fulfill the following condition almost surely:

$$f^*(\boldsymbol{x}) = \operatorname{argmax}_{\tilde{y}} \ \tilde{p}(\tilde{y}|\boldsymbol{x}), \tilde{y} \in \{1, \cdots, K, \circledR\}.$$

According to the definition of $\tilde{p}$, we can further rewrite it as:

$$f^*(\boldsymbol{x}) = \begin{cases} \circledR, & \max_{\tilde{y} \in \{1,\cdots,K\}} \frac{p(\tilde{y}|\boldsymbol{x})}{2-c(\boldsymbol{x})} \leq \frac{1-c(\boldsymbol{x})}{2-c(\boldsymbol{x})}, \\ \operatorname{argmax}_{\tilde{y} \in \{1,\cdots,K\}} \frac{p(\tilde{y}|\boldsymbol{x})}{2-c(\boldsymbol{x})}, & \text{else}, \end{cases}$$

which coincides with the Chow's rule. Then we have the following conclusions:

$$\begin{aligned} \bar{R}_{01}(f) &= \mathbb{E}_{\tilde{p}(\boldsymbol{x},\tilde{y})}[(2 - c(\boldsymbol{x}))\ell_{01}(f(\boldsymbol{x}), \tilde{y})] \\ &= \int_{\boldsymbol{x}} \sum_{\tilde{y}=1}^{K} (2 - c(\boldsymbol{x}))\ell_{01}(f(\boldsymbol{x}), \tilde{y}) \frac{p(\boldsymbol{x}, y)}{2 - c(\boldsymbol{x})} d\boldsymbol{x} + \int_{\boldsymbol{x}} (2 - c(\boldsymbol{x}))\ell_{01}(f(\boldsymbol{x}), K + 1) \frac{p(\boldsymbol{x})}{2 - c(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}} \sum_{\tilde{y}=1}^{K} (2 - c(\boldsymbol{x}))\ell_{01}(f(\boldsymbol{x}), \tilde{y}) \frac{p(\boldsymbol{x}, y)}{2 - c(\boldsymbol{x})} d\boldsymbol{x} + \int_{\boldsymbol{x}} (2 - c(\boldsymbol{x}))\ell_{01}(f(\boldsymbol{x}), \circledR) \frac{(1 - c(\boldsymbol{x}))p(\boldsymbol{x})}{2 - c(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}} \sum_{\tilde{y}=1}^{K} \ell_{01}(f(\boldsymbol{x}), \tilde{y})p(\boldsymbol{x}, y)d\boldsymbol{x} + \int_{\boldsymbol{x}} \ell_{01}(f(\boldsymbol{x}), \circledR)(1 - c(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \end{aligned}$$

Suppose $C(f(\boldsymbol{x})) = \sum_{\tilde{y}=1}^{K} \ell_{01}(f(\boldsymbol{x}), \tilde{y})p(y|\boldsymbol{x}) + \ell_{01}(f(\boldsymbol{x}), \circledR)(1 - c(\boldsymbol{x}))$ is the inner risk and f^* is the Chow's rule, we have that

- If $f^*(\boldsymbol{x}) = \circledR$ and $f(\boldsymbol{x}) \neq f^*(\boldsymbol{x})$:
$$C(f(\boldsymbol{x})) - C(f^*(\boldsymbol{x})) = 1 - c(\boldsymbol{x}) - p(f(\boldsymbol{x})|\boldsymbol{x}).$$
- If $f^*(\boldsymbol{x}) \in \{1, \cdots, K\}$ and $f(\boldsymbol{x}) = \circledR$:
$$C(f(\boldsymbol{x})) - C(f^*(\boldsymbol{x})) = p(f^*(\boldsymbol{x})|\boldsymbol{x}).$$
- If $f^*(\boldsymbol{x}), f(\boldsymbol{x}) \in \{1, \cdots, K\}$ and $f(\boldsymbol{x}) \neq f^*(\boldsymbol{x})$:
$$C(f(\boldsymbol{x})) - C(f^*(\boldsymbol{x})) = p(f^*(\boldsymbol{x})|\boldsymbol{x})) - p(f(\boldsymbol{x})|\boldsymbol{x}).$$

These conclusion shows that

$$C(f(\boldsymbol{x})) - C(f^*(\boldsymbol{x})) = \mathbb{E}_{p(y|\boldsymbol{x})}[\ell_{01c}(f(\boldsymbol{x}), y)] - \mathbb{E}_{p(y|\boldsymbol{x})}[\ell_{01c}(f^*(\boldsymbol{x}), y)].$$

We can conclude the proof by taking the expectation over $p(\boldsymbol{x})$ on both sides of the equation. □

It can be seen that when $c(\boldsymbol{x})$ is constant, we can divide each side of Corollary 1 to get the proof of Theorem 1.

B Proof of Theorem 2

Proof.

$$\begin{aligned} R_{L_c^{\Phi}}(\boldsymbol{g}) &= \mathbb{E}_{p(\boldsymbol{x},y)}[L_c^{\Phi}(\boldsymbol{g}(\boldsymbol{x}), y)] \\ &= \mathbb{E}_{p(\boldsymbol{x},y)}[\Phi(\boldsymbol{g}(\boldsymbol{x}), y)] + (1 - c)\mathbb{E}_{p(\boldsymbol{x})}[\Phi(\boldsymbol{g}(\boldsymbol{x}), K + 1)] \\ &= (2 - c)\tilde{R}_{\Phi}(\boldsymbol{g}) \end{aligned}$$

□

C Proof of Theorem 3

We first give the definition of Rademacher complexity:

Definition 7. *(Rademacher complexity [5])* 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:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{Z_1,\cdots,Z_n\sim\mu}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right], \tag{7}$$

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

Then we can begin proving Theorem 3.

Proof. According to the conditions in Theorem 3, we can learn that L_c^Φ is $(2-c)\rho$-Lipschitz continuous and is bounded by $(2-c)C_\Phi$. By applying the McDiarmid's inequality [36], it is routine [39] to show that the following inequalities holds with probability at least $1 - \frac{\delta}{2}$, respectively:

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left(R_{L_c^\Phi(\boldsymbol{g})} - \hat{R}_{L_c^\Phi(\boldsymbol{g})}\right) \le \mathbb{E}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{\boldsymbol{g}\in\mathcal{G}}\left(R_{L_c^\Phi(\boldsymbol{g})} - \hat{R}_{L_c^\Phi(\boldsymbol{g})}\right)\right] + (2-c)C_\Phi\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left(\hat{R}_{L_c^\Phi(\boldsymbol{g})} - R_{L_c^\Phi(\boldsymbol{g})}\right) \le \mathbb{E}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{\boldsymbol{g}\in\mathcal{G}}\left(\hat{R}_{L_c^\Phi(\boldsymbol{g})} - R_{L_c^\Phi(\boldsymbol{g})}\right)\right] + (2-c)C_\Phi\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

By applying Talagrand's contraction lemma [35], we can learn that:

$$\mathbb{E}_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n}\left[\sup_{\boldsymbol{g}\in\mathcal{G}}\left(R_{L_c^\Phi(\boldsymbol{g})} - \hat{R}_{L_c^\Phi(\boldsymbol{g})}\right)\right] \le \sqrt{2}(2-c)\rho\sum_{y=1}^{K+1}\mathfrak{R}_n(\mathcal{G}_y)$$

and this conclusion also holds for another direction. Plugging this conclusion into the former inequalities and using the union bound, we can learn:

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left|R_{L_c^\Phi(\boldsymbol{g})} - \hat{R}_{L_c^\Phi(\boldsymbol{g})}\right| \le \sqrt{2}(2-c)\rho\sum_{y=1}^{K+1}\mathfrak{R}_n(\mathcal{G}_y) + (2-c)C_\Phi\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

According to the definition of empirical risk minimization and identifiable condition, we can get the following conclusion, where $\boldsymbol{g}^*$ is the optimal solution among all the measurable functions:

$$\begin{aligned}
R_{L_c^\Phi}(\hat{\boldsymbol{g}}) - R_{L_c^\Phi}^* &= \left(R_{L_c^\Phi}(\hat{\boldsymbol{g}}) - \hat{R}_{L_c^\Phi}(\hat{\boldsymbol{g}})\right) + \left(\hat{R}_{L_c^\Phi}(\hat{\boldsymbol{g}}) - \hat{R}_{L_c^\Phi}(\boldsymbol{g}^*)\right) + \left(\hat{R}_{L_c^\Phi}(\boldsymbol{g}^*) - R_{L_c^\Phi}^*\right) \\
&\le \left(R_{L_c^\Phi}(\hat{\boldsymbol{g}}) - \hat{R}_{L_c^\Phi}(\hat{\boldsymbol{g}})\right) + \left(\hat{R}_{L_c^\Phi}(\boldsymbol{g}^*) - R_{L_c^\Phi}^*\right) \\
&\le 2\sup_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}}\left|R_{L_c^\Phi}(\boldsymbol{g}) - \hat{R}_{L_c^\Phi}(\boldsymbol{g})\right|
\end{aligned}$$

which concludes the proof. □

D Proof of Theorem 4

Proof. According to Theorem 1, Theorem 2, and Theorem 3 in Ramaswamy and Agarwal [46], we can immediately learn the sufficiency of this condition.

We complete the proof of the necessity of the calibration of Φ by contradiction. Suppose there are some $\boldsymbol{u} \in \Delta^{K+1}$ that:

$$\inf_{\boldsymbol{u}\in\mathbb{R}^K, \boldsymbol{u}\notin\operatorname{argmin}_{\boldsymbol{u}} \boldsymbol{p}^T\boldsymbol{L}_{01}(\boldsymbol{u})} \boldsymbol{p}^T\boldsymbol{\Phi}(\boldsymbol{u}) = \inf_{\boldsymbol{u}\in\mathbb{R}^K} \boldsymbol{p}^T\boldsymbol{\Phi}(\boldsymbol{u}).$$

It is easy to learn that any re-permutation of $\boldsymbol{u}$ also fulfill the equation above, and we define the collection of these vectors as $\mathcal{U}$. Then we can construct a distribution over $\mathcal{X} \times \{1, \cdots, K+1\}$ whose posterior possibility is $\boldsymbol{u}' \in \mathcal{U}$ for all $\boldsymbol{x}$, on which Φ is not $\ell_{01}-$consistent. However, in our scenario, we only focus on a special distribution with density $\tilde{p}(\boldsymbol{x}, \tilde{y})$ over $\mathcal{X} \times \{1, \cdots, K+1\}$, where $\tilde{p}(K+1|\boldsymbol{x}) = \frac{1-c}{2-c}$ and $\tilde{p}(\tilde{y}|\boldsymbol{x}) = p(\tilde{y}|\boldsymbol{x})/(2-c)$ if $\tilde{y} \neq K+1$. A natural idea is that according to the particularity of $\tilde{p}$, there may not be overlap between $\mathcal{U}$ and all the potential $\{\tilde{p}(\tilde{y}|\boldsymbol{x})\}_{\tilde{y}=1}^{K+1}$. However, according to the arbitrariness of c, this idea is not true, i.e., there always exists a distribution $\{p(y|\boldsymbol{x})\}_{y=1}^{K}$ and c that $\{\tilde{p}(\tilde{y}|\boldsymbol{x})\}_{\tilde{y}=1}^{K+1} \in \mathcal{U}$. Then we can easily define a distribution based on $\{\tilde{p}(\tilde{y}|\boldsymbol{x})\}_{\tilde{y}=1}^{K+1}$, on which Φ is not $\ell_{01}-$consistent. According to the equivalence shown in Theorem 1 and 2, this observation indicates that L_c^{Φ} is not $\ell_{01c}-$consistent *w.r.t.* to this distribution, which shows the necessity of the $\ell_{01}-$calibration of Φ.

$\square$

E Proof of Theorem 5

Proof. According to [59], we can directly get the formulation of the optimal solution of GCE. Based on this formulation, we prove the classification-calibration of GCE constructively by giving an regret transfer bound.

First of all, we show that the excess error of GCE loss for any $\boldsymbol{x}$ is a reweighted version of the Tsallis relative entropy [22, 48] in actual. Denote by $S(\boldsymbol{g}^*)_y = \boldsymbol{q}_y^*$, $S(\boldsymbol{g})_y = \boldsymbol{q}_y$ for any $\boldsymbol{g}$, and $p(y|\boldsymbol{x}) = \eta_y$. We substitute γ with r in the proof for simplicity:

$$
\begin{aligned}
Ex(\boldsymbol{q}, \boldsymbol{x}) &= \sum_{y=1}^{y} \eta_y \frac{(1-\boldsymbol{q}_y^r)}{r} - \sum_{y=1}^{y} \eta_y \frac{(1-\boldsymbol{q}_y^{*r})}{r} \\
&= \frac{\sum_{y=1}^{K} \eta_y (\boldsymbol{q}_y^{*r} - \boldsymbol{q}_y^r)}{r} \\
&= \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{1-r} \frac{\left(1 - \sum_{y=1}^{K} q_y^{*(1-r)} q_y^r\right)}{r}
\end{aligned}
$$

It can be seen that the second term of the last equation is the Tsallis relative entropy between discrete possibilities $\boldsymbol{q}^*$ and $\boldsymbol{q}$. According to the Corollary 9 of [22] and (4.13) of [48], we can lower bound the excess error with the total variation distance between $\boldsymbol{q}^*$ and $\boldsymbol{q}$ and get a Pinsker's type inequality:

$$
Ex(\boldsymbol{q}, \boldsymbol{x}) \geq \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{1-r} \frac{1-r}{2} \|\boldsymbol{q}^* - \boldsymbol{q}\|_1^2
$$

Then we have to connect the *r.h.s.* of the inequality to the excess error w.r.t. 0-1 loss. When $\operatorname{argmax}_y q_y(\boldsymbol{x}) \neq \operatorname{argmax}_y \eta_y$, denote by $\operatorname{argmax}_y q_y(\boldsymbol{x}) = pred$ and $\operatorname{argmax}_y \eta_y = max$:

$$
\begin{aligned}
\|\boldsymbol{q}^* - \boldsymbol{q}\|_1 &= \sum_{y=1}^{K} |q_y^* - q_y| \\
&\geq |q_{max}^* - q_{max}| + |q_{pred}^* - q_{pred}| \\
&\geq |q_{max}^* - q_{pred}^* + q_{pred} - q_{max}|
\end{aligned}
$$

According to the formulation of the optimal solution of GCE, we can learn that $q^*_{max} \geq q^*_{pred}$. Since $\text{argmax}_y\, q_y(\boldsymbol{x}) \neq \text{argmax}_y\, \eta_y$, we can learn that $q_{pred} \geq q_{max}$. Then we can further learn that:

$$\begin{aligned}
\|\boldsymbol{q}^* - \boldsymbol{q}\|_1 \geq |q^*_{max} - q^*_{pred}| \\
&= \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1} |\eta_{max}^{\frac{1}{1-r}} - \eta_{pred}^{\frac{1}{1-r}}| \\
&= \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1} (\eta_{max}^{\frac{1}{1-r}} - \eta_{pred}^{\frac{1}{1-r}}) \\
&= \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1} (\eta_{max} * \eta_{max}^{\frac{r}{1-r}} - \eta_{pred} * \eta_{pred}^{\frac{r}{1-r}}) \\
&\geq \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1} (\eta_{max} * \eta_{max}^{\frac{r}{1-r}} - \eta_{pred} * \eta_{max}^{\frac{r}{1-r}}) \\
&= \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1} \eta_{\max}^{\frac{r}{1-r}} (\eta_{max} - \eta_{pred})
\end{aligned}$$

Then we can learn that:

$$\begin{aligned}
Ex(\boldsymbol{q}, \boldsymbol{x}) &\geq \left(\sum_{y=1}^{K} \eta_y^{\frac{1}{1-r}}\right)^{-1-r} \eta_{\max}^{\frac{2r}{1-r}} * \frac{1-r}{2} (\eta_{max} - \eta_{pred})^2 \\
&\geq \frac{1-r}{2K^{\frac{2r}{1-r}+r+r^2}} (\eta_{max} - \eta_{pred})^2
\end{aligned}$$

Then we have the following regret transfer bound:

$$R_{01}(\operatorname*{argmax}_y \boldsymbol{g}_y) - R^*_{01} \leq \sqrt{C(R_G(\boldsymbol{g}) - R^*_G)},$$

where $C = \frac{2K^{\frac{2r}{1-r}+r+r^2}}{1-r}$, R_G is the expected version of GCE loss, and R^*_G and R^*_{01} are the optimal value of the expected version of GCE loss and 0-1 loss, respectively. From this bound, we constructively prove the classification-calibration of GCE loss with $r \in (0, 1)$. $\square$

It is noticeable that the bound does not hold for $r = 1$, e.g., the case of MAE loss, and the regret transfer bound becomes less compact when r increases. We prove the classification-calibration of MAE loss by showing its regret transfer bound.

Corollary 2. Suppose the expected version of MAE loss is $R_M(\boldsymbol{g})$ and its minimal value is R^*_M. Then we have:

$$R_{01}(\operatorname*{argmax}_y \boldsymbol{g}_y) - R^*_{01} \leq K(R_M(\boldsymbol{g}) - R^*_M).$$

Proof. Given the formulation of the optimal solution $\boldsymbol{q}^*$ of expected MAE loss in Theorem 5, for any $\boldsymbol{x}$, the excess error can be written as:

$$\begin{aligned}
Ex(\boldsymbol{q}, \boldsymbol{x}) &= \sum_{y=1}^{K} \eta_y(1 - q_y) - \sum_{y=1}^{K} \eta_y(1 - q^*_y) \\
&= \sum_{y=1}^{K} \eta_y(q^*_y - q_y) \\
&= \eta_{max} - \sum_{y=1}^{K} \eta_y q_y
\end{aligned}$$

When When $\text{argmax}_y\, q_y(\boldsymbol{x}) \neq \text{argmax}_y\, \eta_y$:

$$\begin{aligned}\eta_{max} - \sum_{y=1}^{K} \eta_y q_y &= \eta_{max} - \eta_{pred} q_{pred} - \sum_{y \neq pred}^{K} \eta_y q_y \\ &\geq \eta_{max} - \eta_{pred} q_{pred} - \eta_{max}(1 - q_{pred}) \\ &= q_{pred}(\eta_{max} - \eta_{pred}) \\ &\geq \frac{1}{K}(\eta_{max} - \eta_{pred}),\end{aligned}$$

which concludes the proof by taking the expectation on both sides. □

Combine the conclusions above and we can conclude the proof. Though the bound for GCE becomes less tight when r increases, the MAE loss has a better regret transfer bound, which indicates that the regret transfer bound of GCE for $r \in (0, 1)$ may not be good enough. A potential reason is that [22, 48] considered the general case of Tsallis relative entropy while we only need the case that $\boldsymbol{q}$ is a probability distribution. It is promising to further tighten this bound by modifying the conclusions in [22, 48] and limiting $\boldsymbol{q}$ to a $K - 1$-dimensional probability simplex.

F Details of the Experiment Setup

F.1 Detailed Information of Benchmark Datasets

In the experiments, we used 3 widely-used benchmark datasets. Here, we report the sources of these datasets and the way we split them.

- Fashion-MNIST [58]. It is a 10-class dataset of fashion items. Each instance is a 28*28 grayscale image. Source: https://github.com/zalandoresearch/fashion-mnist.
- SVHN [42] It is a 10-class dataset for 10 different digits and each instance is a 32*32*3 colored image in RGB format. Source: http://ufldl.stanford.edu/housenumbers/.
- CIFAR-10 [30]. It is a 10-class dataset for 10 different objects and each instance is a 32*32*3 colored image in RGB format. Source: https://www.cs.toronto.edu/~kriz/cifar.html.

For Fashion-MNIST and SVHN, we trained models on the whole training dataset. For CIFAR-10, we splited 10% of the training dataset as the validation set and conducted random crop and flips for data augmentation. The cost c is less than 0.5 as suggested in [47] and further decreased on Fashion-MNIST since it is a less difficult dataset.

F.2 Detailed Information of the Models and Optimization Algorithm

For Fashion-MNIST, we used the model defined in [8] for the experiments. For SVHN and CIFAR-10, ResNet-18 and ResNet-34 is used, respectively. For the cost-sensitive method [8], we use batch normalization [27] at the output layer as suggested in [8] since it fails to work without this modification.

Adam with default momentum was used for optimization in this paper. For Fashion-MNIST, the epoch number, batch size, learning rate, and weight decay are set to 20, 256, 1e-3, and 1e-4. For SVHN, the epoch number, batch size, learning rate, and weight decay are set to 20, 1024, 1e-3, and 1e-4. For CIFAR-10, the epoch number, batch size, learning rate, and weight decay are set to or selected from 200, 1024, {1e-3, 2e-3, 3e-3}, and 1e-4. For Fashion-MNIST and SVHN, we use the model after the 20th epoch for performance evaluation. For CIFAR-10, we report the performance of the model with the best performance on the validation dataset. Temperature scaling is further conducted for CE on CIFAR-10.

G Details of Instance-dependent Rejection Cost

In practical applications, it can be beneficial letting the rejection cost $c(\boldsymbol{x})$ vary among different samples. For example, when constructing a system to automatically prescribe for users, a wrong prescription can be fatal for users of advanced ages or with underlying diseases. To prevent such wrong prescriptions, the cost for this type of users can be decreased to encourage rejection. However, it is not suitable encouraging rejection for all the users, which makes the system meaningless. An acceptable choice is to increase the cost for rejection instead for users of low risk.

In this appendix, we expand the Theorem 2 and propose a surrogate for instance dependent cost based on Corollary 1, whose estimation error bound and calibration analysis can be derived almost symmetrically thanks to the equivalence shown in Corollary 3. Then we further evaluate its performance on SVHN dataset.

G.1 Expansion of Theorem 2

Theorem 2 tells the equivalence between surrogate risk minimization of L_c^Φ on $p(\boldsymbol{x}, y)$ and surrogate risk minimization of Φ on $\tilde{p}(\boldsymbol{x}, \tilde{y})$. Here we expand it to the case of instance-dependent cost.

Given the cost function $c(\boldsymbol{x})$ and any function $\Phi(\cdot) : \mathbb{R}^{K+1} \times \{1, \cdots, K+1\} \to \mathbb{R}^+$:

$$L_{c(\boldsymbol{x})}^\Phi(\boldsymbol{u}, y) = (\Phi(\boldsymbol{u}, y) + (1 - c(\boldsymbol{x}))\Phi(\boldsymbol{u}, K+1))/(2 - c(\boldsymbol{x})).$$

Then we have the following conclusion:

Corollary 3. For any $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^{K+1}$ and $R_{L_{c(\boldsymbol{x})}^\Phi}(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)}[L_{c(\boldsymbol{x})}^\Phi(\boldsymbol{g}(\boldsymbol{x}), y)]$:

$$R_{L_{c(\boldsymbol{x})}^\Phi}(\boldsymbol{g}) = \tilde{R}_\Phi(\boldsymbol{g})$$

Proof.

$$\begin{aligned} R_{L_{c(\boldsymbol{x})}^\Phi}(\boldsymbol{g}) &= \mathbb{E}_{p(\boldsymbol{x},y)}[L_{c(\boldsymbol{x})}^\Phi(\boldsymbol{g}(\boldsymbol{x}), y)] \\ &= \mathbb{E}_{p(\boldsymbol{x},y)}[(\Phi(\boldsymbol{g}(\boldsymbol{x}), y) + (1 - c(\boldsymbol{x}))\Phi(\boldsymbol{g}(\boldsymbol{x}), K+1))/(2 - c(\boldsymbol{x}))] \\ &= \int_{\boldsymbol{x}} \sum_{y=1}^{K} \Phi(\boldsymbol{g}(\boldsymbol{x}), y) \frac{p(\boldsymbol{x}, y)}{2 - c(\boldsymbol{x})} d\boldsymbol{x} + \int_{\boldsymbol{x}} \frac{(1 - c(\boldsymbol{x}))p(\boldsymbol{x})}{2 - c(\boldsymbol{x})} \Phi(\boldsymbol{g}(\boldsymbol{x}), K+1) d\boldsymbol{x} \\ &= \tilde{R}_\Phi(\boldsymbol{g}) \end{aligned}$$

□

The derivation of its estimation error bound is similar to that of Theorem 3 by modifying the upper bound and Lipschitz constant, and the necessity and sufficiency of the ℓ_{01}-calibration of Φ can also be proved by utilizing the arbitrariness of $\tilde{p}(\boldsymbol{x}, y)$ as in Appendix D.

G.2 Experiments on SVHN

In this section, we compare our proposed surrogate $L_{c(\boldsymbol{x})}^\Phi$ with CE and DEFER on SVHN. The cost-sensitive learning-based method [8] is not compared since it cannot tackle the case of instance-dependent cost.

In the experiments, we use SVHN [42] to demonstrate the effectiveness of $L_{c(\boldsymbol{x})}^\Phi$. To generate instance-dependent costs, we split 10% of the training dataset and manually corrupt it into to a binary dataset by aggregating the 10 classes into [‘0’, ‘2’, ‘3’, ‘5’, ‘6’, ‘8’, ‘9’] and [‘1’, ‘4’, ‘7’]. We train a binary classifier with on the corrupted dataset with 10 epochs. Then we further use the obtained classifier on training and testing set to split them into 2 parts. For any $\boldsymbol{x}$ that is classified as [‘0’, ‘2’, ‘3’, ‘5’, ‘6’, ‘8’, ‘9’], we set $c(\boldsymbol{x}) = c_1$ and c_2 otherwise. In the experiments, Adam with default momentum is used with learning rate, batch size and weight decay set to 1e-3, 1024, and 1e-4, respectively. The model used is ResNet-18.

Table 3: The mean and standard error of the zero-one-c losses (rescaled to 0-100), rejection ratio, and missclassification rates of the accepted data for 5 trails. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Method	$(c_1,\ c_2)$	CE			DEFER			GCE		
		01c	Rej	01	01c	Rej	01	01c	Rej	01
SVHN	(0.50, 0.10)	8.03	4.46	7.60	8.00	9.20	5.07	**7.20**	6.73	5.13
		(0.16)	(0.54)	(0.01)	(0.30)	(0.72)	(0.25)	**(0.17)**	(6.73)	(5.13)
	(0.45, 0.15)	7.80	4.36	7.03	9.10	9.93	**5.07**	6.93	7.00	4.70
		(0.26)	(0.31)	(0.23)	(0.46)	(0.41)	(0.42)	**(0.31)**	(0.35)	(0.26)
	(0.40, 0.20)	7.70	4.50	6.83	7.80	11.13	5.00	**7.03**	7.93	4.80
		(0.10)	(0.50)	(0.25)	(0.26)	(1.27)	(0.44)	**(0.21)**	(0.55)	(0.17)
	(0.35, 0.25)	7.76	4.90	6.67	7.70	11.93	4.80	**6.83**	8.43	4.63
		(0.12)	(0.20)	(0.15)	(0.26)	(0.45)	(0.10)	**(0.20)**	(0.31)	(0.15)

The experimental results are reported in the table above. It can be seen that in the scenario of instance-dependent cost, the prop osed surrogate with GCE loss still outperforms baseline methods, which aligns with the observations in Section 6.

H Limitations and Potential Negative Social Impacts

Limitations: This framework is used for multi-class classification with rejection, while there are also other scenarios for learning with rejection, e.g., AUC optimization with rejection [51]. We believe that extensions to CwR with complex evaluation is a promising future direction.

Potential Negative Social Impacts: Though classification with rejection can be useful in risk-critical missions, it can lead to inefficient services once abused, i.e., used in risk-insensitive missions. This is also the potential negative social impact of all the methods for CwR.