Mitigating Underfitting in Learning to Defer with Consistent Losses

Shuqi Liu[1†] **Yuzhou Cao**[2†] **Qiaozhen Zhang**[1] **Lei Feng**[2] **Bo An**[3,2]
[1]School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Skywork AI, Singapore
(correspondence to zhangqz@nankai.edu.cn)
† Equal contribution

Abstract

Learning to defer (L2D) allows the classifier to defer its prediction to an expert for safer predictions, by balancing the system's accuracy and extra costs incurred by consulting the expert. Various loss functions have been proposed for L2D, but they were shown to cause the *underfitting* of trained classifiers when extra consulting costs exist, resulting in degraded performance. In this paper, we propose a novel loss formulation that can mitigate the underfitting issue while maintaining statistical consistency. We first show that our formulation can avoid a common characteristic shared by most existing losses, which has been shown to be a cause of underfitting, and show that it can be combined with the representative losses for L2D to enhance their performance and yield consistent losses. We further study the regret transfer bounds of the proposed losses and experimentally validate its improvements over existing methods.

1 Introduction

For machine learning models deployed in risk-sensitive tasks (e.g., medical diagnosis (Kadampur and Al Riyaee, 2020), autonomous driving (Grigorescu et al., 2020), and healthcare (Beede et al., 2020)), an incorrect prediction can result in serious and even fatal consequences. To meet the requirements of these tasks, the learning to defer (L2D) framework was studied (Madras et al., 2018; Bansal et al., 2021; Okati et al., 2021; De et al., 2021; Mozannar and Sontag, 2020; Verma and Nalisnick, 2022; Mozannar et al., 2022; Charusaie et al., 2022; Verma et al., 2023; Mozannar et al., 2023; Straitouri et al., 2023; Narasimhan et al., 2022; Mao et al., 2023a) that aims to help reduce critical mistakes by enabling models to defer to human experts, which can be seen as the generalization of the classification with rejection task (Chow, 1970; Bartlett and Wegkamp, 2008; Cortes et al., 2016a,b; Ramaswamy et al., 2018; Ni et al., 2019; Cao et al., 2022; Mao et al., 2023b). In L2D, the classifier is equipped with an option of deferral that allows the classifier to refrain from making decisions and seeking answers from an expert who is more likely to be correct for certain predictions.

In L2D, the performance of the classifier augmented with a deferral option is evaluated by a system loss. In most previous studies, the target system loss is set to the system's misclassification cost and an extra cost incurred by the choice of deferring to the expert, which can be either positive or zero. Then L2D is further formulated as a risk minimization problem that aims at minimizing the expected system loss over an underlying data distribution. Given the discontinuous nature of the target system loss, various surrogate losses were designed to make the optimization problem tractable while ensuring statistical consistency. In Mozannar and Sontag (2020), a cross-entropy-like surrogate loss was used to integrate the option of deferral into the training of the classifier. Then, OvA-type losses and an asymmetric softmax-based surrogate loss (Verma and Nalisnick, 2022; Verma et al., 2023; Cao et al., 2023) were further proposed to improve the models' calibration of probability forecasts. A unified framework was proposed in Charusaie et al. (2022) that allows the use of any consistent multiclass loss (Tewari and Bartlett, 2007) for constructing a consistent surrogate for L2D, which was further shown in Cao et al. (2023) to include existing representative surrogates (Mozannar and Sontag, 2020; Verma and Nalisnick, 2022; Cao et al., 2023) as its special realizations by incorporating specific base

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

losses for multiclass classification.

Though the existing losses have guarantees on their infinite-sample consistency, it was pointed out by Narasimhan et al. (2022) that their finite-sample efficacy can be greatly weakened and will suffer from the *underfitting* issue when the extra cost of deferring to the expert is non-zero. That is, with a non-zero extra cost, the formulation of these losses all possesses a *label smoothing term* (Szegedy et al., 2016), which is redundant in the task of L2D and will lead to a flattened training distribution. As a result, it can be harder to recognize the true label for each sample and thus the performance of the learned classifier will be degraded. This degradation of the classifier performance can further have negative impacts on the augmented deferring rule, which finally makes them underfit the training set. A post-hoc estimator was proposed (Narasimhan et al., 2022) to alleviate this underfitting issue, however, it requires the re-training of the deferring rule before the deployment of L2D models.

Do there exist consistent surrogate losses that can resist the issue of underfitting caused by the extra costs of consulting the expert? Such an underfitting-resistant consistent surrogate is naturally end-to-end, which can avoid the extra re-training procedure and its required additional data, and thus close the gap between the training and deployment of L2D systems. In this paper, we give a positive answer to this question by proposing a novel loss formulation that is free from the redundant label smoothing term and we further examine various consistent surrogates for L2D whose performance remains satisfactory under the existence of a non-zero expert cost. The main contributions of this paper are four-fold as summarized below:

- We propose a novel label-smoothing-free loss formulation based on the observation that the redundant label smoothing term in the previous general loss formulation (Narasimhan et al., 2022) can be eliminated by utilizing the intermediate result of models in the training process.
- We show that the existing representative surrogate losses for L2D can be label-smoothing-free by plugging their base multiclass losses into our proposed loss formulation and their consistency guarantees can be also attained, which demonstrates the flexibility of our approach.
- We provide regret transfer bounds *w.r.t.* the target system loss for proposed consistent loss formulation, which shows that the models that have low risk *w.r.t.* surrogate losses are also good solutions for L2D. A more detailed regret bound on the model's classification ability further indicates the efficacy of our proposed method.
- Experimental results on benchmark datasets with different experts accuracy and extra costs validate the superiority of our proposed loss formulation and its robustness to underfitting.

2 Preliminaries

In this section, we introduce the problem formulation and existing solutions for L2D, then we review the issue of underfitting and its cause.

2.1 The Problem Formulation and Existing Losses for L2D

Problem Setup: Let us denote by $\mathcal{X}$ and $\mathcal{Y} = [K]$ the feature space and label space, respectively. In the problem of L2D, we focus on the feature-label-expert random variable triplet $X \times Y \times M \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$, which has a joint distribution with density $p(\boldsymbol{x}, y, m)$ where $\boldsymbol{x},\ y,\ m$ are their realizations. The training sample set $\mathcal{S}_n = \{\boldsymbol{x}_i, y_i, m_i\}_{i=1}^n$ is independently and identically drawn from the joint distribution $p(\boldsymbol{x}, y, m)$. We denote by $\mathrm{Acc}(\boldsymbol{x}) = \Pr(M = Y | X = \boldsymbol{x})$ the accuracy of the expert on $\boldsymbol{x}$ in the rest of this paper.

Let us denote by $\perp$ the option of deferring to the expert and $\mathcal{Y}^\perp = \mathcal{Y} \cup \{\perp\}$. L2D aims to learn an augmented classifier $f : \mathcal{X} \to \mathcal{Y}^\perp$ with performance evaluated by a system loss consists of the classifier loss $\ell_{\mathrm{clf}}(\boldsymbol{x}, y, f(\boldsymbol{x}))$ and the expert loss $\ell_{\mathrm{exp}}(\boldsymbol{x}, y, m)$, which are metrics for the classifier and the expert, respectively. The classifier loss ℓ_{clf} is usually set to the ordinary zero-one loss $\mathbb{I}(f(\boldsymbol{x}) \neq y)$, and the expert loss further takes into consideration the extra cost of consulting the expert, e.g., monetary expense and computational cost. As a result, the expert loss is in the form of $c + \mathbb{I}(m \neq y)$, which is a mixture of the expert's misclassification error and the extra cost of consulting the expert. In this way, the target system loss is shown below (Mozannar and Sontag, 2020; Narasimhan et al., 2022):

$$\begin{aligned}\ell_{01}^{\perp}(f(\boldsymbol{x}), y, m) =& \mathbb{I}(f(\boldsymbol{x}) \neq \perp)\mathbb{I}(f(\boldsymbol{x}) \neq y) \\ &+ \mathbb{I}(f(\boldsymbol{x}) = \perp)(c + \mathbb{I}(m \neq y)),\end{aligned} \quad (1)$$

which is determined by both the quality of prediction and whether the decision of deferral is made. Finally, a risk minimization problem *w.r.t.* this loss is formulated, which is the goal of L2D:

$$\min_f R_{01}^{\perp}(f) = \mathbb{E}_{p(\boldsymbol{x},y,m)}[\ell_{01}^{\perp}(f(\boldsymbol{x}), y, m)]. \quad (2)$$

It was shown in Mozannar and Sontag (2020) that f^* is a Bayes optimal solution *w.r.t.* $\ell_{01}^{\perp}$ if and only if it meets the following condition:

$$f^*(\boldsymbol{x}) = \begin{cases} \perp, & \mathrm{Acc}(\boldsymbol{x}) \geq p(y^*|\boldsymbol{x}) + c, \\ \mathrm{argmax}_{y \in [K]}\, p(y|\boldsymbol{x}), & \text{else.} \end{cases} \quad (3)$$

where $y^* = \operatorname{argmax}_{y\in\mathcal{Y}} p(y|\boldsymbol{x})$ is the optimal prediction of multiclass classification. Intuitively, an accepted expert is required to be strictly more accurate than the optimal classifier by c to make sure that the gain of the expert's high accuracy is not offset by the incurred cost of consulting him.

Existing Losses for L2D: Various surrogate losses have been proposed (Mozannar and Sontag, 2020; Verma and Nalisnick, 2022; Cao et al., 2023), to make (2) tractable. Recently, Cao et al. (2023) showed that these surrogate losses can be unified into a general framework proposed by Charusaie et al. (2022), which can be expressed by the following formulation given the target loss (1):

$$\ell_\Psi(\boldsymbol{g}(\boldsymbol{x}), y, m) = \Psi(\boldsymbol{g}(\boldsymbol{x}), y) + c\mathbb{I}(m \neq y)\sum_{y'\in[K]}\Psi(\boldsymbol{g}(\boldsymbol{x}), y') + (1-c)\mathbb{I}(m = y)\Psi(\boldsymbol{g}(\boldsymbol{x}), K+1) \quad (4)$$

where $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^{K+1}$ is a scoring function and $\Psi : \mathbb{R}^{K+1} \times [K+1] \to \mathbb{R}_+$ is a multiclass loss. A link function is further introduced to induce an augmented classifier $f_{\boldsymbol{g}}$ from the scoring function:

$$f_{\boldsymbol{g}}(\boldsymbol{x}) = \begin{cases} \perp, & \operatorname{argmax}_{y\in[K+1]} g_y(\boldsymbol{x}) = K+1. \\ \operatorname{argmax}_y g_y(\boldsymbol{x}), & \text{else.} \end{cases} \quad (5)$$

The loss formulation (4) shares critical properties with a $(K+1)$-class classification problem. We can rewrite the risk $R^{\perp}_{\Psi}(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y,m)}[\ell_\Psi(\boldsymbol{g}(\boldsymbol{x}), y, m)]$ into the following form:

$$R^{\perp}_{\Psi}(\boldsymbol{g}) = \mathbb{E}_{\hat{p}(\boldsymbol{x},y)}[\hat{\xi}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}), y)], \quad (6)$$

where $\hat{p}(\boldsymbol{x}, y)$ is dummy distribution defined as

$$\hat{p}(y|\boldsymbol{x}) = \frac{1}{\hat{\xi}(\boldsymbol{x})}\begin{cases} (1-c)\mathrm{Acc}(\boldsymbol{x}), & y = K+1, \\ p(y|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x})), & \text{else}, \end{cases} \quad (7)$$

and $\hat{\xi}(\boldsymbol{x}) = 1 + (1-c)\mathrm{Acc}(\boldsymbol{x}) + Kc(1-\mathrm{Acc}(\boldsymbol{x}))$ is the normalization variable. This problem reduction is similar to the case of classification with rejection in Cao et al. (2022) and we will elaborate on it in Appendix A. It can be learned that $\operatorname{argmax}_y \hat{p}(y|\boldsymbol{x}) = K+1$ if and only if $\mathrm{Acc}(\boldsymbol{x}) > p(y^*|\boldsymbol{x}) + c$, in this case $f^*(\boldsymbol{x}) = \perp$; otherwise $f^*(\boldsymbol{x}) = \operatorname{argmax}_y \hat{p}(y|\boldsymbol{x})$. As a result, for any $\boldsymbol{g}^* \in \operatorname{argmin}_{\boldsymbol{g}} R^{\perp}_{\Psi}(\boldsymbol{g})$, $\operatorname{argmax}_y g^*_y(\boldsymbol{x}) = \operatorname{argmax}_y \hat{p}(y|\boldsymbol{x})$ and thus $f_{\boldsymbol{g}^*}(\boldsymbol{x}) = f^*(\boldsymbol{x})$ if Ψ is a consistent multiclass loss, which indicates the consistency of (4).

2.2 Underfitting and Redundant Label Smoothing Issues in L2D

While (4) successfully induces a general consistent loss formulation, a natural question is whether it is also troubled by the issue of underfitting as its special realization in Mozannar and Sontag (2020). Unfortunately, we show that this problem still persists.

It is worth noting that $c\mathbb{I}(m \neq y)\sum_{y\in[K]}\Psi(\boldsymbol{g}(\boldsymbol{x}), y)$ in (4) is exactly a label smoothing term that is active with a wrong expert and $c > 0$, which is proved to be the cause of underfitting for the method in Mozannar and Sontag (2020). To verify that such a label smoothing term is also harmful in the general loss formulation (4), we follow the practice in Narasimhan et al. (2022) by inspecting the *probability margin* of both the original distribution p and the dummy distribution $\hat{p}$. For any $y \in \mathcal{Y}/\{y^*\}$, we can learn from the definition of $\hat{p}$ that:

$$\hat{p}(y^*|\boldsymbol{x}) - \hat{p}(y|\boldsymbol{x}) = \frac{p(y^*|\boldsymbol{x}) - p(y|\boldsymbol{x})}{1 + (1-c)\mathrm{Acc}(\boldsymbol{x}) + Kc(1-\mathrm{Acc}(\boldsymbol{x}))}.$$

Notice that the denominator of the equation's *r.h.s.* is larger than 1 in general, we can learn that the margin between the likelihood of the optimal label and non-optimal ones shrinks when we conduct problem reduction from the distribution p to $\hat{p}$. Furthermore, such a margin shrinks in the rate of $\mathcal{O}(\frac{1}{K})$ as long as the expert is not perfect. With an increasing class number K, the probability margin decays rapidly and will make it more challenging to recognize the optimal label y^* in the sense that the training distribution is flattened (Chou et al., 2020). Therefore, all the losses induced from (4) would face the problem of underfitting the training set.

3 Elimination of the Redundant Label Smoothing Term

In the previous section, we revisited the problem of underfitting in L2D with a non-zero cost $c > 0$, which is a common issue shared by existing losses induced from the loss formulation (4). According to the previous study (Narasimhan et al., 2022), the shrinking of the possibility margin is highly related to the existence of the redundant label smoothing term, which finally causes the problem of underfitting. Given these clues, a natural question for the mitigation of underfitting arises: can we derive a novel loss formulation that can avoid the redundant label smoothing term while maintaining its consistency? We confirm the existence of such a loss formulation in two stages: in this section, we trace back to the motivation of (4) and propose a novel loss formulation for the construction of L2D losses that can eliminate the label smoothing term; then we induce various consistent losses using this formulation in the next section.

According to the previous discussion, the redundant label smoothing term in (4) is closely related to the extra $c(1-\mathrm{Acc}(\boldsymbol{x}))$ appearing **uniformly** in all the

labels' modified class probabilities $\hat{p}(y|\boldsymbol{x})$. By recalling the definition (7) of $\hat{p}$ and adding $c\mathrm{Acc}(\boldsymbol{x})/\hat{\xi}(\boldsymbol{x})$ to both $\hat{p}(y|\boldsymbol{x})$ and $\hat{p}(K+1|\boldsymbol{x})$ in (7), we can learn that:

$$\begin{cases} \hat{p}(y|\boldsymbol{x}) \geq \hat{p}(K+1|\boldsymbol{x}) \Leftrightarrow p(y|\boldsymbol{x}) + c \geq \mathrm{Acc}(\boldsymbol{x}), \forall y \in [K], \\ \hat{p}(y|\boldsymbol{x}) \geq \hat{p}(y'|\boldsymbol{x}) \Leftrightarrow p(y|\boldsymbol{x}) \geq p(y'|\boldsymbol{x}), \quad \forall y, y' \in [K]. \end{cases}$$

This property can immediately guarantee the consistency of the deferral rule, according to the definition of Bayes optimal solution (3). However, by recalling the Bayes optimal solution (3) for L2D, we can find there is redundancy in the first requirement above. Concretely, it is unnecessary for all the labels to fulfill this requirement. Then, we denote by $y^* = \mathrm{argmax}_{y\in\mathcal{Y}}\, p(y|\boldsymbol{x})$ the Bayes optimal label, and thus the Bayes optimal solution (3) indicates that we only have to compare $p(y^*|\boldsymbol{x}) + c$ and $\mathrm{Acc}(\boldsymbol{x})$ to correctly judge if we should defer to an expert. This discovery leads to the following simplified requirement for a better distribution $\tilde{p}$:

$$\begin{cases} \tilde{p}(y^*|\boldsymbol{x}) \geq \tilde{p}(K+1|\boldsymbol{x}) \Leftrightarrow p(y^*|\boldsymbol{x}) + c \geq \mathrm{Acc}(\boldsymbol{x}), \\ \tilde{p}(y|\boldsymbol{x}) \geq \tilde{p}(y'|\boldsymbol{x}) \Leftrightarrow p(y|\boldsymbol{x}) \geq p(y'|\boldsymbol{x}), \forall y, y' \in \mathcal{Y}. \end{cases} \tag{8}$$

We squeeze out the redundancy of the first requirement by removing the requirements for the labels other than y^*. Since the previous indiscriminate requirement is implemented by adding $c(1 - \mathrm{Acc}(\boldsymbol{x}))$ uniformly to $p(y|\boldsymbol{x})$, which is the cause of label smoothing term, our simplified requirement (8) is expected to be free from label smoothing. Based on the discussions above, we move to consider the following ideal scenario: suppose we are aware of the Bayes optimal solution y^* for each $\boldsymbol{x}$, we can construct the following formulation for the class-posterior probability $\tilde{p}(y|\boldsymbol{x})$:

$$\tilde{p}(y|\boldsymbol{x}) = \frac{1}{\tilde{\xi}(\boldsymbol{x})} \begin{cases} p(y|\boldsymbol{x}), & y \in [K]/\{y^*\}, \\ p(y|\boldsymbol{x}) + c(1 - \mathrm{Acc}(\boldsymbol{x})), & y = y^*, \\ (1-c)\mathrm{Acc}(\boldsymbol{x}), & y = K+1, \end{cases}$$

where $\tilde{\xi}(\boldsymbol{x}) = 1 + c + (1-2c)\mathrm{Acc}(\boldsymbol{x})$ is to normalize the possibilities into the probability simplex Δ^{K+1}. By inspecting this new distribution $\tilde{p}(y|\boldsymbol{x})$, we can learn that (8) holds, which means that we can reproduce (3) by solving the $(K+1)$-class classification problem on $\tilde{p}(y|\boldsymbol{x})$. Furthermore, the probability margin between y^* and other labels *w.r.t.* $\tilde{p}$ is coherently larger than that in $\hat{p}$, i.e., $\forall y \in [K]/\{y^*\}$, $\tilde{p}(y^*|\boldsymbol{x}) - \tilde{p}(y|\boldsymbol{x}) > \hat{p}(y^*|\boldsymbol{x}) - \hat{p}(y|\boldsymbol{x})$[1], , which can alleviate the difficulty of recognizing y^*, and thus mitigate the issue of underfitting. This enlargement of probability margin is a direct effect of our new distribution $\hat{p}$ that avoids redundant terms added on non-optimal labels, which makes the value of $\hat{p}(y|\boldsymbol{x})$ independent from the class number K. Similarly to the relation between formulation (6) and loss formulation (4), we can derive the following loss formulation, whose expectation *w.r.t.* $\tilde{p}(\boldsymbol{x}, y)$ is $\mathbb{E}_{\tilde{p}(\boldsymbol{x},y)}[\tilde{\xi}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}), y)]$:

$$\begin{aligned} \tilde{\ell}_\Psi(\boldsymbol{g}(\boldsymbol{x}), y, m) = \Psi(\boldsymbol{g}(\boldsymbol{x}), y) + c\mathbb{I}(m \neq y)\Psi(\boldsymbol{g}(\boldsymbol{x}), y^*) \\ + (1-c)\mathbb{I}(m = y)\Psi(\boldsymbol{g}(\boldsymbol{x}), K+1). \end{aligned} \tag{9}$$

It is noticeable that the above formulation is free from label smoothing, where the label smoothing term is substituted by the loss *w.r.t.* the Bayes optimal label y^*. According to the conditional expectation and definition of $\hat{p}$, we can choose any consistent multiclass classification losses to instantiate Ψ to get a consistent surrogate for L2D with a non-zero expert cost. However, we should recall that this is an idealized scenario: in most cases, there is no access to the Bayes optimal label y^*, and thus the loss formulation above is not applicable in practical applications.

Though the loss formulation is not directly applicable, its derivation can still provide quite meaningful insights: focusing on a certain label can help avoid the presence of redundant label smoothing, while the consistency is further guaranteed if the label is Bayes-optimal. While concentrating on a certain label is easy to implement since we can substitute y^* by any other y, it is hard to ensure its Bayes optimality. To obtain a feasible solution, there is a promising scheme: using the **intermediate learning results** as the approximation of Bayes optimal labels. Concretely, for a scoring function $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^{K+1}$, the index of $\mathrm{argmax}_{y\in[K]}\, g_y(\boldsymbol{x})$ is used as the estimate of y^* *w.r.t.* $\boldsymbol{x}$. If $\mathrm{argmax}_{y\in[K]}\, g_y(\boldsymbol{x})$ approaches y^* gradually in the training process, it is reasonable to expect the consistency of using it as the approximation of y^* in (9). Based on this idea, we can correct (9) as follows:

Definition 1. (label-smoothing-free Loss Formulation) Let us denote by $\Psi : \mathbb{R}^{K+1} \times [K+1] \to \mathbb{R}_+$ a multiclass loss and $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^{K+1}$ the scoring function, we have the following formulation without extra label smoothing term:

$$\begin{aligned} \tilde{\ell}_\Psi(\boldsymbol{g}(\boldsymbol{x}), y, m) = \Psi(\boldsymbol{g}(\boldsymbol{x}), y) + c\mathbb{I}(m \neq y) \min_{y'\in[K]} \Psi(\boldsymbol{g}(\boldsymbol{x}), y') \\ + (1-c)\mathbb{I}(m = y)\Psi(\boldsymbol{g}(\boldsymbol{x}), K+1). \end{aligned} \tag{10}$$

Notice that we use $\min_{y'\in[K]} \Psi(\boldsymbol{g}(\boldsymbol{x}), y')$ here instead of $\Psi(\boldsymbol{g}(\boldsymbol{x}), \mathrm{argmax}_{y'\in[K]} g_y(\boldsymbol{x}))$ since they are equivalent for monotonic losses that $g_y(\boldsymbol{x}) > g_{y'}(\boldsymbol{x}) \to \Psi(\boldsymbol{g}(\boldsymbol{x}), y) < \Psi(\boldsymbol{g}(\boldsymbol{x}), y')$.

Finally, the label smoothing term is completely eliminated in our proposed loss formulation (10). In the next section, we will show that our formulation can be instantiated with existing surrogates to make them free from label smoothing with consistency guarantees.

[1]We prove this claim in Appendix B.

Table 1: Multiclass losses used in representative consistent L2D surrogates.

L2D Loss	Multiclass Loss $\Psi(\boldsymbol{g}(\boldsymbol{x}), y)$
Mozannar and Sontag (2020) (Ψ_{CE})	$-\log\left(\frac{\exp(g_y(\boldsymbol{x}))}{\sum_{y'\in[K+1]}\exp(g_y(\boldsymbol{x}))}\right)$
Verma and Nalisnick (2022) (Ψ_{OvA})	$\begin{cases}\log\left(1+\exp(-g_y(\boldsymbol{x}))\right)-\log\left(1+\exp(g_y(\boldsymbol{x}))\right), & y=K+1,\\ \log\left(1+\exp(-g_y(\boldsymbol{x}))\right)+\sum_{y'\in[K+1]/\{y\}}\log\left(1+\exp(g_y(\boldsymbol{x}))\right), & \text{else.}\end{cases}$
Cao et al. (2023) (Ψ_{ASM})	$\begin{cases}-\log\left(\frac{\exp(g_y(\boldsymbol{x}))}{\sum_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))-\max_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))}\right), & y=K+1,\\ -\log\left(\frac{\exp(g_y(\boldsymbol{x}))}{\sum_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))}\right)-\log\left(\frac{\sum_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))-\max_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))}{\sum_{y'\in[K+1]}\exp(g_{y'}(\boldsymbol{x}))-\max_{y'\in[K]}\exp(g_{y'}(\boldsymbol{x}))}\right), & \text{else.}\end{cases}$

4 Theoretical Results: Consistent Realizations and Regret Analysis

In this section, we first instantiate our formulation with representative L2D surrogates to correct them into label-smoothing-free training objectives and their consistency guarantees can be still attained. Then we further provide the regret transfer bounds of the losses for both tasks of L2D and multiclass classification, which demonstrate the efficacy of our formulation on L2D with a non-zero expert cost.

4.1 Consistent Realizations with Representative Surrogates

In Section 3, we have derived a label-smoothing-free loss formulation (10) by removing redundant terms and trusting the intermediate learning results. To find practical consistent surrogates for L2D with our proposed general formulation, the next step is to check whether potential multiclass base losses Ψ can be plugged into our formulation (10) to be free from label smoothing and attain consistency guarantees. Recalling that the representative consistent L2D surrogates proposed in Mozannar and Sontag (2020); Verma and Nalisnick (2022); Cao et al. (2023) are the realizations of (4) with specific multiclass losses, which are summarized in Table 1. The following theorem shows that these base losses can also induce consistent surrogates by using our proposed formulation:

Theorem 1. Let $R^{\perp}_{\Psi}(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y,m)}[\tilde{\ell}_{\Psi}(\boldsymbol{g}(\boldsymbol{x},y,m))]$. The formulation (10) is a consistent L2D loss if Ψ is set to any multiclass loss in Table 1:

$$\forall \boldsymbol{g}^* \in \operatorname{argmin}_{\boldsymbol{g}} R^{\perp}_{\Psi}(\boldsymbol{g}),\ f_{\boldsymbol{g}^*} \in \operatorname{argmin}_f R^{\perp}_{01}(f).$$

Sketch of proof: The proofs of the consistency results for the three losses are similar. The first step is to make sure that the first K dimensions of the scoring function, i.e., $\boldsymbol{g}_{[K]}$, is a consistent multiclass classifier. This point can be proved by substituting the Bayes-optimal label y^* into (9) with different labels $y' \in [K]$ and enumerating the minimums of their risks, which will finally lead to the conclusion that the global minimum can only be achieved when $y' = y^*$, i.e., $\operatorname{argmax}_{y\in[K]} g_y(\boldsymbol{x}) = y^*$. Then, the consistency of the deferral rule can be further proved based on this result. The detailed proof is provided in Appendix C.

Given the consistency results, we can plug the multiclass losses $\Psi(\boldsymbol{g}(\boldsymbol{x}), y)$ listed in Table 1 into (10) to get consistent surrogates. They can also be seen as the label-smoothing-free correction for the corresponding existing losses. In Section 5, we experimentally validate the improvement of our methods against these existing non-corrected losses.

4.2 Regret Transfer Bounds for L2D and Multiclass Classification

The previous section shows the infinite-sample consistency of the proposed surrogates. In this section, we further study the regret transfer bounds of the proposed surrogates, which provide performance guarantees for models that are close to the optimal ones. Denote by $\texttt{Regret}^{\perp}_{01}(\boldsymbol{g}) = R^{\perp}_{01}(f_{\boldsymbol{g}}) - R^{\perp}_{01}(f^*)$ the regret of target loss and $\texttt{Regret}^{\perp}_{\Psi}(\boldsymbol{g}) = R^{\perp}_{\Psi}(\boldsymbol{g}) - R^{\perp}_{\Psi}(\boldsymbol{g}^*)$ the surrogate loss's risk. Then we have the following regret transfer bound:

Theorem 2. For surrogate losses in Table 1, we have the conclusion that $\texttt{Regret}^{\perp}_{\Psi}(\boldsymbol{g}) = \mathcal{O}\left(\sqrt{\texttt{Regret}^{\perp}_{\Psi}(\boldsymbol{g})}\right)$, i.e., $\texttt{Regret}^{\perp}_{\Psi}(\boldsymbol{g}) \leq \alpha_{\Psi}\sqrt{\texttt{Regret}^{\perp}_{\Psi}(\boldsymbol{g})}$, where $\alpha_{\Psi} > 0$ is a constant depend on Ψ.

In practical scenarios, we usually approximate $R^{\perp}_{\Psi}(\boldsymbol{g})$ with its empirical version and obtain a sub-optimal solution that is close to the optimal solution. This conclusion further guarantees the performance of our methods in this case.

While the bound above is similar to those proposed in previous works, the following regret transfer bound for the model's classifier counterpart's misclassification error further is unique for our formulation, which is focused on samples with low expert accuracy:

Table 2: The mean and standard error of the system error *w.r.t.* $\ell_{01}^{\perp}$ (Err, rescaled to 0-100), and Coverage (Cov) for 5 trails, on the CIFAR-100 dataset. We compare the performance of the label-smoothing-free surrogates and their non-label-smoothing-free counterparts in pairs. The best performance is highlighted in boldface.

Method	Cost	**ASM$^+$**		ASM		**CE$^+$**		CE		**OvA$^+$**		OvA$_1$	
		Err	Cov	Err	Cov	Err	Cov	Err	Cov	Err	Cov	Err	Cov
CIFAR-100 (94%)	0.10	**21.71**	**92.66**	22.11	78.02	**20.56**	**81.03**	22.18	80.18	**22.86**	**88.92**	24.69	76.16
		(0.15)	**(0.36)**	(0.19)	(0.43)	**(0.21)**	**(0.61)**	(0.19)	(0.45)	**(0.10)**	**(0.38)**	(0.20)	(1.48)
	0.20	**22.55**	**94.64**	26.30	81.37	**22.37**	**83.88**	25.37	82.78	**23.60**	**91.65**	31.45	77.33
		(0.10)	**(0.22)**	(0.11)	(0.29)	**(0.07)**	**(0.17)**	(0.19)	(0.37)	**(0.13)**	**(0.51)**	(0.07)	(0.40)
	0.30	**23.24**	**96.51**	30.65	84.60	**23.76**	**85.60**	31.65	85.28	**23.99**	**93.16**	42.66	65.63
		(0.04)	**(0.11)**	(0.23)	(0.17)	**(0.33)**	**(0.19)**	(0.26)	(0.27)	**(0.38)**	**(0.21)**	(0.46)	(0.93)
	0.40	**23.42**	**97.79**	35.11	86.97	**24.41**	**88.06**	35.32	87.12	**25.20**	**92.36**	50.65	74.89
		(0.07)	**(0.20)**	(0.17)	(0.23)	**(0.17)**	**(0.15)**	(0.14)	(0.26)	**(0.19)**	**(0.25)**	(0.24)	(0.52)
CIFAR-100 (75%)	0.10	**23.04**	**96.48**	23.14	84.69	**22.82**	84.49	22.97	**89.76**	**23.62**	**91.97**	25.91	87.17
		(0.17)	**(0.19)**	(0.10)	(0.51)	**(0.17)**	(0.28)	(0.08)	**(5.46)**	**(0.12)**	**(0.45)**	(0.10)	(0.64)
	0.20	**23.31**	**97.57**	25.52	87.85	**24.13**	85.69	25.76	**88.39**	**24.28**	**92.71**	28.56	86.54
		(0.10)	**(0.18)**	(0.32)	(0.39)	**(0.19)**	(0.37)	(0.16)	**(0.33)**	**(0.15)**	**(0.26)**	(0.15)	(1.48)
	0.30	**23.63**	**98.06**	29.68	90.95	**24.94**	87.42	29.65	**91.95**	**24.09**	**94.44**	34.64	84.33
		(0.15)	**(0.10)**	(0.18)	(0.28)	**(0.12)**	(0.33)	(0.15)	**(0.25)**	**(0.16)**	**(0.25)**	(0.27)	(0.98)
	0.40	**24.04**	**98.82**	32.80	94.16	**25.62**	89.31	33.27	**94.12**	**24.82**	**96.02**	49.51	77.66
		(0.08)	**(0.12)**	(0.11)	(0.25)	**(0.14)**	(0.28)	(0.17)	**(0.28)**	**(0.21)**	**(0.44)**	(0.39)	(1.26)

Corollary 1. Let us denote by $\mathcal{X}^\beta \subset \mathcal{X}$ the set that $\forall \boldsymbol{x} \in \mathcal{X}^\beta$, $\mathrm{Acc}(\boldsymbol{x}) \leq \beta$. Then we have the following conclusion:

$$\texttt{Regret}_{01}(\boldsymbol{g}|\mathcal{X}^\beta) \leq \alpha_{\beta,\Psi}\texttt{Regret}_{\Psi}^{\perp}(\boldsymbol{g}|\mathcal{X}^\beta),$$

where $\texttt{Regret}_{01}(\boldsymbol{g}|\mathcal{X}^\beta) = \mathbb{E}_{p(\boldsymbol{x},y|\mathcal{X}^\beta)}[\mathbb{I}(\operatorname{argmax}_{y'\in[K]} g_{y'}) \neq y] - \mathbb{E}_{p(\boldsymbol{x},y|\mathcal{X}^\beta)}[\min_{y\in[K]} p(y|\boldsymbol{x})]$ is the regret of $\boldsymbol{g}$'s classifier counterpart *w.r.t* 0-1 loss and $\texttt{Regret}_{\Psi}^{\perp}(\boldsymbol{g}|\mathcal{X}^\beta)$ is the surrogate loss's regret conditioned on $\mathcal{X}^\beta$. $\alpha_{\beta,\Psi} > 0$ is a determined by β, Ψ, and is increasing when β.

The proof of this corollary and Theorem 2 is provided in Appendix D. According to this corollary, we can learn that the regret transfer bound for the classifier's misclassification rate is linear on the set that expert's accuracy is smaller than β. Compared with the quadratic bound in Theorem 2, this bound is a tighter one. Furthermore, when the expert's performance decreases, the bound becomes tighter as $\alpha_{\beta,\Psi}$ is also decreasing. This property conforms to the motivation of L2D since a sample point with low expert accuracy can be more dependent on the classifier's prediction. The observations above indicate that our method can better fit a classifier, which can finally result in a better L2D system. We will experimentally validate it in the next section.

5 Experiments

In this section, we empirically evaluate the improvement of our method over the existing surrogates by experiments. All the experiments are conducted with 8 NVIDIA GeForce 3090 GPUs.

5.1 Experimental Setup

Models, Datasets, and Optimizers: We evaluate our method and baselines on widely used benchmark datasets with different expert settings:

- We first conduct experiments on the CIFAR-100 dataset (Krizhevsky et al., 2009) with data augmentation. Following the settings in Mozannar and Sontag (2020) and Verma and Nalisnick (2022), in order to simulate different expert accuracy, we set the expert to an oracle that can provide the true label of an instance with probability $p \in \{75\%,\ 94\%\}$ if its label is in the first 50 classes, and does random guessing otherwise. The used model is a 28-layer WideResnet (Zagoruyko and Komodakis, 2016).
- For the experiments on the CIFAR-10 dataset (Krizhevsky et al., 2009), we use an expert who can provide the true label of an instance with probability $p = 94\%$ on the first 5 classes. We also conduct experiments without data augmentation to simulate tasks with different difficulties as suggested in Mozannar and Sontag (2020). The used model is ResNet-18 (He et al., 2016).

A batch norm layer is added to the last layer of the model for all the methods based on OvA strategy in the training process for stable outputs as in Charoenphakdee et al. (2021). For all the methods, SGD with cosine annealing is used with learning rate, weight decay, and batch size set to 1e-1, 5e-4, and 128. The epoch numbers for CIFAR-10 and CIFAR-100 are 200 and 400, respectively. 10% of the training set is split

Table 3: The mean and standard error of the system error *w.r.t.* $\ell_{01}^{\perp}$ (Err, rescaled to 0-100), and Coverage (Cov) for 5 trails, on the CIFAR-10 dataset. We compare the performance of the label-smoothing-free surrogates and their non-label-smoothing-free counterparts in pairs. The best performance is highlighted in boldface.

Method	Cost	**ASM$^+$**		ASM		**CE$^+$**		CE		**OvA$^+$**		OvA$_1$	
		Err	Cov	Err	Cov	Err	Cov	Err	Cov	Err	Cov	Err	Cov
CIFAR-10(w)	0.10	**13.21**	**94.99**	15.17	91.92	**12.71**	90.56	14.60	**93.37**	**10.94**	**94.83**	11.33	94.44
		(0.07)	**(0.20)**	(0.11)	(0.14)	**(0.14)**	(0.27)	(0.38)	**(0.31)**	**(0.11)**	**(0.16)**	(0.15)	(0.37)
	0.20	**14.30**	**96.08**	17.59	92.39	**13.19**	92.27	15.51	**93.53**	**11.79**	94.58	12.00	**95.54**
		(0.08)	**(0.06)**	(0.34)	(0.14)	**(0.08)**	(0.11)	(0.38)	**(0.36)**	**(0.14)**	(4.25)	(0.35)	**(0.37)**
	0.30	**14.62**	**99.19**	18.89	94.91	**14.55**	95.40	16.44	**95.49**	**12.05**	**97.25**	13.22	97.11
		(0.28)	**(0.08)**	(0.10)	(1.03)	**(0.30)**	(0.77)	(0.37)	**(0.36)**	**(0.14)**	**(0.21)**	(0.16)	(1.01)
	0.40	**14.62**	**99.82**	19.29	98.13	**15.20**	95.66	19.24	**96.58**	**12.34**	**98.52**	13.68	97.72
		(0.09)	**(0.06)**	(0.17)	(0.15)	**(0.15)**	(0.45)	(0.39)	**(0.26)**	**(0.26)**	**(0.33)**	(0.18)	(0.13)
CIFAR-10(o)	0.10	**23.32**	**97.35**	24.40	92.30	21.08	80.32	**21.05**	**89.01**	**20.84**	**94.79**	21.25	91.80
		(0.15)	**(0.23)**	(0.10)	(0.26)	(0.09)	(0.79)	**(0.19)**	**(4.85)**	**(0.10)**	**(0.29)**	(0.35)	(0.65)
	0.20	**24.30**	**99.33**	27.92	93.87	**23.40**	**84.00**	23.41	63.06	**21.79**	**96.73**	22.41	94.29
		(0.13)	**(0.05)**	(0.09)	(0.16)	**(0.13)**	**(0.31)**	(0.09)	(36.17)	**(0.17)**	**(0.22)**	(0.12)	(0.42)
	0.30	**24.05**	**99.85**	31.74	95.41	**23.99**	87.86	26.70	**92.65**	**21.87**	**97.68**	23.56	94.28
		(0.31)	**(0.04)**	(0.09)	(0.28)	**(0.12)**	(0.36)	(0.12)	**(0.32)**	**(0.11)**	**(0.16)**	(0.20)	(0.46)
	0.40	**24.43**	**99.98**	32.32	97.51	**26.28**	88.37	27.96	**92.39**	**22.67**	**98.26**	25.69	94.30
		(0.06)	**(0.01)**	(0.49)	(0.12)	**(0.09)**	(0.13)	(0.09)	**(0.30)**	**(0.09)**	**(0.09)**	(0.45)	(0.52)

out as the validation set.

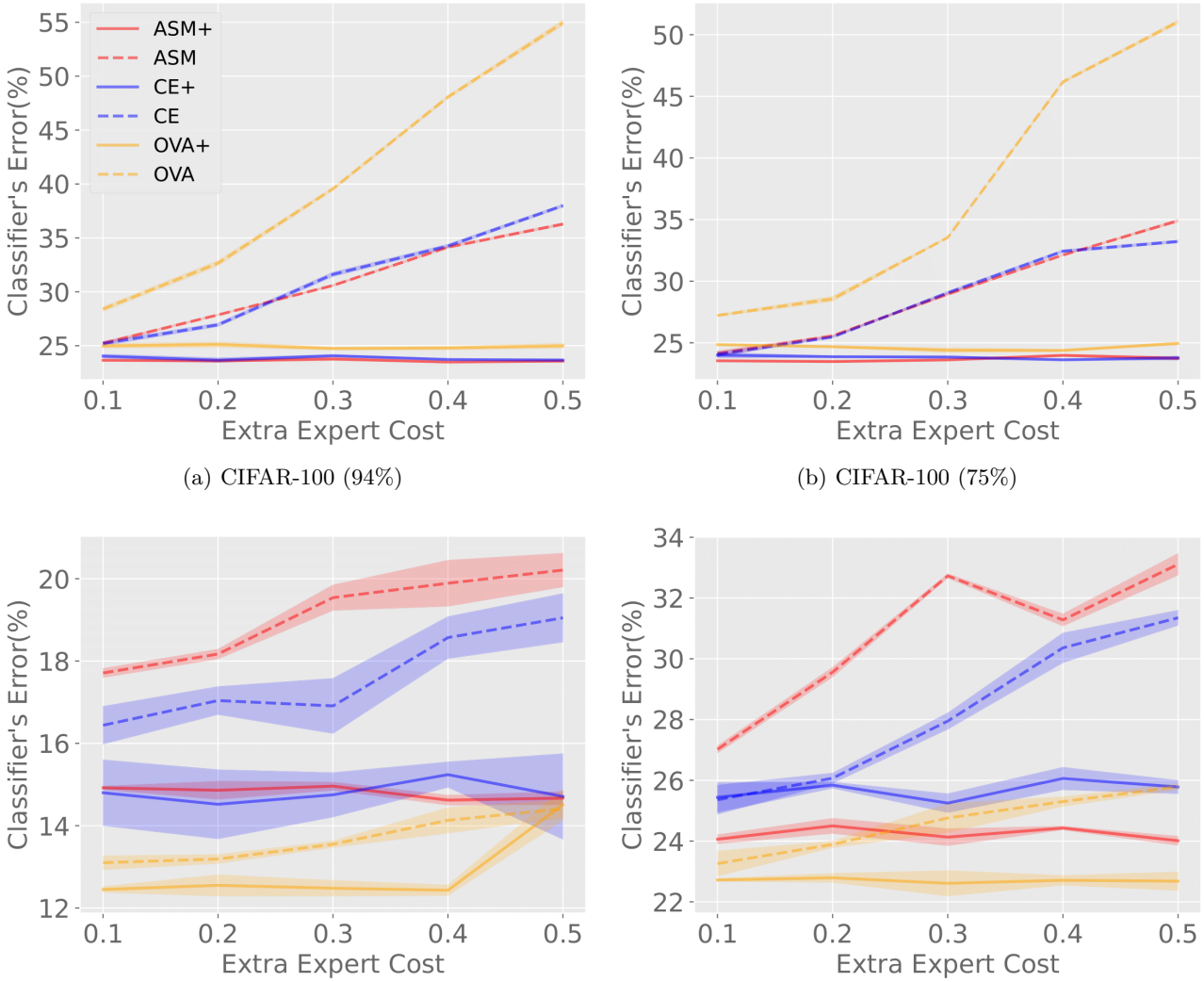
Baselines: In the experiments, we focus on the comparisons between end-to-end methods of consistent surrogates without post-hoc correction, as the correction requires extra retraining. Our methods refer to the consistent realizations of our loss formulation (10) with the three base losses of representative surrogates in Table 1. The baseline methods are the combination of (4) and losses in Table 1, which correspond to the surrogates proposed in Mozannar and Sontag (2020); Verma and Nalisnick (2022); Cao et al. (2023). These methods are abbreviated as **ASM+**/ASM, **CE+**/CE, and **OvA+**/OvA according to the base losses and loss formulations they use. We report the results with an extra expert cost $c \in \{0, 10, 0, 20, 0, 30, 0, 40\}$ since a larger cost will completely deactivate the deferral rule and reduce the problem into ordinary multiclass classification, which is out of the scope of this paper.

5.2 Experimental Results

We first report the obtained models' performance *w.r.t.* the target loss $\ell_{01}^{\perp}$ with different costs c to measure the improvement of our methods in Table 2 and 3, and then report the coverage, i.e., the ratio of non-deferred instances, as in previous works to further characterize the behavior of all the methods. To evaluate all the methods' robustness to underfitting, we report the classifier counterpart's classification error, i.e., misclassification rate, of all the methods in Figure 1.

Comparison of Classification's Error: According to Figure 1, we can see that the novel label-smoothing-free surrogates all outperform their existing counterparts with redundant label smoothing terms in all the setups. In (a) and (b), we can learn that the proposed methods, i.e., **ASM+**, **CE+**, and **OvA+**, are not sensitive to the change of extra cost, and their classifier's performance remains unchanged. However, the surrogates that possess label smoothing terms i.e., ASM, CE, and OvA, suffer seriously from the issue of underfitting, and their performance drops drastically with the increasing of extra expert cost, which is caused by the increasing degree of label smoothing. In (c) and (d), though the change of performance is less obvious compared with the results of CIFAR-100 due to the decreasing of tasks' difficulty, the trend is still unchanged that the baseline accuracy gets lower by increasing c. In conclusion, our proposed methods remain effective under different costs c, thanks to the elimination of label smoothing terms, while baseline methods derived from (4) are all affected by the issue of underfitting.

Comparison of System Error and Coverage: From Table 2 and 3, we can see that our proposed methods enjoy lower system error in most cases. This advantage gets more obvious with larger extra expert costs, which also validates the robustness of our loss formulation to the issue of underfitting. Our loss formulation's coverage is also higher than the baselines' in most cases, except for CE, which was empirically shown in previous studies (Mozannar and Sontag, 2020; Cao et al., 2023) to have lower coverage. Combining the observations above, we can deduce that our formulation can defer to the correct experts, reduce needless deferral decisions, and further classify the accepted samples correctly, which can validate that our methods combat underfitting, as shown in Figure 1.



(a) CIFAR-100 (94%)

(b) CIFAR-100 (75%)

(c) CIFAR-10 (w)

(d) CIFAR-10 (o)

Figure 1: The misclassification rate of the classifier's counterpart's prediction $\operatorname{argmax}_{y\in[K]} g_y$ for all the methods. Solid lines are the methods derived from our proposed formulation (10) and dashed lines are the methods derived from the previous formulation (4).

6 Conclusion and Future Work

Conclusion: In this paper, we investigated the problem of underfitting mitigation with consistent surrogate losses in the scenario of L2D with an extra non-zero expert cost. We first pointed out that the redundant label smoothing term in existing L2D losses that leads to their underfitting is caused by the general loss formulation they belong to. Then we traced back to the motivation of consistent L2D surrogates' design and proposed a novel loss formulation that can avoid redundant label smoothing terms. With this loss formulation, we derived consistent L2D surrogates based on this formulation that can turn the existing L2D surrogates into label-smoothing-free forms and still attain consistency guarantees. We also provided regret transfer bounds, which further justified the efficacy of our formulation. Finally, the experimental result on datasets with different task difficulties demonstrated the improvement of our formulation over the existing surrogate and validated the robustness of our loss formulation to the issue of underfitting.

Future Work: A promising future direction is to explore the conditions that provide consistency guarantees for a multiclass surrogate to be plugged into our loss formulation, which can enlighten the design and enlarge the range of choice of label-smoothing-free surrogates for L2D.

Acknowledgements

This work was supported by the State Key Laboratory of CEMEE (CEMEE2020K0301A) and the National Natural Science Foundation of China (Grant Nos. 12371260 and 12271270). Yuzhou Cao and Bo An are supported by the National Research Foundation, Singapore under its Industry Alignment Fund-Prepositioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). Is the most accurate AI the best teammate? optimizing AI for teamwork. In *AAAI*.

Bartlett, P. L. and Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840.

Beede, E., Baylor, E. E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *CHI*.

Cao, Y., Cai, T., Feng, L., Gu, L., GU, J., An, B., Niu, G., and Sugiyama, M. (2022). Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *NeurIPS*.

Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. (2023). In defense of softmax parametrization for calibrated and consistent learning to defer. In *NeurIPS*.

Charoenphakdee, N., Cui, Z., Zhang, Y., and Sugiyama, M. (2021). Classification with rejection based on cost-sensitive classification. In *ICML*.

Charusaie, M., Mozannar, H., Sontag, D. A., and Samadi, S. (2022). Sample efficient learning of predictors that complement humans. In *ICML*.

Chou, Y., Niu, G., Lin, H., and Sugiyama, M. (2020). Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *ICML*.

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

Cortes, C., DeSalvo, G., and Mohri, M. (2016a). Boosting with abstention. In *NeurIPS*, pages 1660–1668.

Cortes, C., DeSalvo, G., and Mohri, M. (2016b). Learning with rejection. In *ALT*, volume 9925, pages 67–82.

De, A., Okati, N., Zarezade, A., and Rodriguez, M. G. (2021). Classification under human assistance. In *AAAI*.

Grigorescu, S. M., Trasnea, B., Cocias, T. T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *J. Field Robotics*, 37(3):362–386.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Kadampur, M. A. and Al Riyaee, S. (2020). Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images. *Informatics in Medicine Unlocked*, 18:100282.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Madras, D., Pitassi, T., and Zemel, R. S. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*.

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

Mao, A., Mohri, M., and Zhong, Y. (2023b). Ranking with abstention. *CoRR*, abs/2307.02035.

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. A. (2023). Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*.

Mozannar, H., Satyanarayan, A., and Sontag, D. A. (2022). Teaching humans when to defer to a classifier via exemplars. In *AAAI*.

Mozannar, H. and Sontag, D. A. (2020). Consistent estimators for learning to defer to an expert. In *ICML*.

Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. (2022). Post-hoc estimators for learning to defer to an expert. In *NeurIPS*.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. (2019). On the calibration of multiclass classification with rejection. In *NeurIPS*, pages 2582–2592.

Okati, N., De, A., and Gomez-Rodriguez, M. (2021). Differentiable learning under triage. In *NeurIPS*.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554.

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. (2023). Improving expert predictions with conformal prediction. In *ICML*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.

Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *J. Mach. Learn. Res.*, 8:1007–1025.

Verma, R., Barrejon, D., and Nalisnick, E. (2023). Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*.

Verma, R. and Nalisnick, E. T. (2022). Calibrated learning to defer with one-vs-all classifiers. In *ICML*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *BMVC*.

A Derivation of (6)

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\ell_\Psi(\boldsymbol{g}(\boldsymbol{x}),y,m)] = \sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+(1-c)\mathrm{Acc}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)\\
&=\left(\sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)+(1-c)\mathrm{Acc}(\boldsymbol{x})\right)\\
&*\frac{\sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+(1-c)\mathrm{Acc}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)}{\sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)+(1-c)\mathrm{Acc}(\boldsymbol{x})}\\
&=(1+(1-c)\mathrm{Acc}(\boldsymbol{x})+Kc(1-\mathrm{Acc}(\boldsymbol{x})))\\
&*\frac{\sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+(1-c)\mathrm{Acc}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)}{1+(1-c)\mathrm{Acc}(\boldsymbol{x})+Kc(1-\mathrm{Acc}(\boldsymbol{x}))}\\
&=\hat{\xi}(\boldsymbol{x})\frac{\sum_{y\in[K]}\left(p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))\right)\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+(1-c)\mathrm{Acc}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)}{\hat{\xi(\boldsymbol{x})}}\\
&=\mathbb{E}_{\hat{p}(y|\boldsymbol{x})}[\hat{\xi}(\boldsymbol{x})\Psi(\boldsymbol{g}(\boldsymbol{x}),y)]
\end{aligned}
$$

and we can take the expectation on $\mathcal{X}$ with $p(\boldsymbol{x})$ to finish the proof. From the derivation, we can learn that $\hat{\xi}(\boldsymbol{x})$ is used to normalize $\hat{p}$ into a valid distribution.

B Proof of the Claim about Probability Margins

We formulate our claim into the following theorem:

Theorem 3. For any $y\in[K]$:

$$\tilde{p}(y^*|\boldsymbol{x})-\tilde{p}(y|\boldsymbol{x})\geq\hat{p}(y^*|\boldsymbol{x})-\hat{p}(y|\boldsymbol{x})$$

and the equality holds if and only if the margin/the accuracy of expert/cost is equal to 0.

Proof. We give the detailed representation of the two margin terms respectively:

$$\tilde{p}(y^*|\boldsymbol{x})-\tilde{p}(y|\boldsymbol{x})=\frac{p(y^*|\boldsymbol{x})-p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}$$

$$\hat{p}(y^*|\boldsymbol{x})-\hat{p}(y|\boldsymbol{x})=\frac{p(y^*|\boldsymbol{x})-p(y|\boldsymbol{x})}{\hat{\xi}(\boldsymbol{x})}$$

Meanwhile:

$$\hat{\xi}(\boldsymbol{x})-\tilde{\xi}(\boldsymbol{x})=(K-1)c(1-\mathrm{Acc}(\boldsymbol{x}))\geq 0,$$

which means that $\hat{\xi}(\boldsymbol{x})\geq\tilde{\xi}(\boldsymbol{x})$. Combining the conclusions above and we can immediately conclude the proof. □

C Proof of Theorem 1

To prove the Theorem 1, we first construct the following auxiliary lemma. Denote by

$$\ell^i_\Psi(\boldsymbol{g}(\boldsymbol{x}),y,m)=\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+c\mathbb{I}(m\neq y)\Psi(\boldsymbol{g}(\boldsymbol{x}),i)+(1-c)\mathbb{I}(m=y)\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)$$

, and we further give the definition of *inner risk* $R^i_\Psi(\boldsymbol{g}|\boldsymbol{x})=\mathbb{E}_{p(y,m|\boldsymbol{x})}[\ell^i_\Psi(\boldsymbol{g}(\boldsymbol{x}),y,m)]$ and similarly $R^\perp_\Psi(\boldsymbol{g}|\boldsymbol{x})=\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_\Psi(\boldsymbol{g}(\boldsymbol{x}),y,m)]$, we have the following lemma:

Lemma 1. $\min_{\boldsymbol{g}} R^i_\Psi(\boldsymbol{g}|\boldsymbol{x})\geq\min_{\boldsymbol{g}} R^{y^*}_\Psi(\boldsymbol{g}|\boldsymbol{x})$, and the equality holds *i.f.f.* $p(i|\boldsymbol{x})=p(y^*|\boldsymbol{x})$.

Proof. We denote by $s_y(\boldsymbol{x})$ the y-th output of symmetric softmax *w.r.t.* $\boldsymbol{g}$ and $a_y(\boldsymbol{x})$ the y-th output of asymmetric softmax, $\sigma(x)$ is the sigmoid output. We first give the following property that characterizes the minimizers for each inner risk $\boldsymbol{g}^{i*} \in \operatorname{argmin}_{\boldsymbol{g}} R^i_\Psi(\boldsymbol{g}|\boldsymbol{x})$.

- If $\Psi =$CE:

$$\begin{cases} s_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})}{(1+c+(1-2c)\mathrm{Acc}(\boldsymbol{x}))}, y \neq i, y \in [K] \\ s_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))}{(1+c+(1-2c)\mathrm{Acc}(\boldsymbol{x}))}, y = i \\ s_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{(1+c+(1-2c)\mathrm{Acc}(\boldsymbol{x}))}, y = K+1 \end{cases}$$

- If $\Psi =$ASM:

$$\begin{cases} a_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y \neq i, y \in [K] \\ a_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y = i \\ a_y(\boldsymbol{g}^{i*}(\boldsymbol{x})) = \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y = K+1 \end{cases}$$

- If $\Psi =$OvA:

$$\begin{cases} \sigma(\boldsymbol{g}^{i*}_y(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y \neq i, y \in [K] \\ \sigma(\boldsymbol{g}^{i*}_y(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})+c(1-\mathrm{Acc}(\boldsymbol{x}))}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y = i \\ \sigma(\boldsymbol{g}^{i*}_y(\boldsymbol{x})) = \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{1+c-c\mathrm{Acc}(\boldsymbol{x})}, y = K+1 \end{cases}$$

Given these characterizations, we only have to show $R^i_\Psi(\boldsymbol{g}^{i*}|\boldsymbol{x}) - R^{y^*}_\Psi(\boldsymbol{g}^{y^**}|\boldsymbol{x}) \geq 0$ for all the three losses. For all the three losses:

- If $\Psi =$CE:

$$\begin{aligned} R^i_\Psi(\boldsymbol{g}^{i*}|\boldsymbol{x}) - R^{y^*}_\Psi(\boldsymbol{g}^{y^**}|\boldsymbol{x}) &= -(p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))) \log\left(p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))\right) - p(y^*|\boldsymbol{x}) \log\left(p(y^*|\boldsymbol{x})\right) \\ &\quad + p(i|\boldsymbol{x}) \log\left(p(i|\boldsymbol{x})\right) + (p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))) \log\left(p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))\right) \\ &= -p(i|\boldsymbol{x}) \log\left(\frac{p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(i|\boldsymbol{x})}\right) + p(y^*|\boldsymbol{x}) \log\left(\frac{p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(y^*|\boldsymbol{x})}\right) \\ &\quad + c(1-\mathrm{Acc}(\boldsymbol{x})) \log\left(\frac{p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}\right) \geq 0 \end{aligned}$$

  and the equality holds only if $p(i|\boldsymbol{x}) = p(y^*|\boldsymbol{x})$. This holds since $x \log(1 + \frac{1}{x})$ is strictly increasing and the term in the second line is also non-negative.

- If $\Psi =$ASM, the case is the same as in $\Psi =$CE.

- If Ψ =OvA:

$$
\begin{aligned}
&R^{i}_{\Psi}(\boldsymbol{g}^{i*}|\boldsymbol{x}) - R^{y^*}_{\Psi}(\boldsymbol{g}^{y^**}|\boldsymbol{x}) \\
&= -(p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x})))\log\left(\frac{p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}\right) - (1-p(i|\boldsymbol{x}))\log\left(\frac{1-p(i|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&- p(y^*|\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) - (1-p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x})))\log\left(\frac{1-p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&+ (p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x})))\log\left(\frac{p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}\right) - (1-p(y^*|\boldsymbol{x}))\log\left(\frac{1-p(y^*|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&- p(i|\boldsymbol{x})\log\left(\frac{p(i|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) - (1-p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x})))\log\left(\frac{1-p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&= -p(i|\boldsymbol{x})\log\left(\frac{p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(i|\boldsymbol{x})}\right) + p(y^*|\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(y^*|\boldsymbol{x})}\right) \\
&+ c(1-\mathrm{Acc}(\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}\right) \\
&- (1-p(y^*|\boldsymbol{x}))\log\left(\frac{(1-p(y^*|\boldsymbol{x})) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{1-p(y^*|\boldsymbol{x})}\right) + (1-p(i|\boldsymbol{x}))\log\left(\frac{1-p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{1-p(i|\boldsymbol{x})|\boldsymbol{x})}\right) \\
&+ c(1-\mathrm{Acc}(\boldsymbol{x}))\log\left(\frac{1-p(i|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}{1-p(y^*|\boldsymbol{x}) + c(1-\mathrm{Acc}(\boldsymbol{x}))}\right) \geq 0
\end{aligned}
$$

  For the last equation, its first two lines ≥ 0 according to the discussion in the first two cases. Its last two lines is also non-negative, which is similar to the derivation of the first two lines.

Combining the conclusions above, we can conclude the proof. □

Given this lemma, we can simply conclude the proof. First, it is can be learned that any minimizer of $R^{y^*}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$ is also that of $R^{\perp}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$ since $R^{\perp}_{\Psi}(\boldsymbol{g}|\boldsymbol{x}) \geq R^{y^*}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$ and the equality holds when $\boldsymbol{g}$ minimizes $R^{y^*}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$ according to the definition of $R^{\perp}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$. There will not be any $\boldsymbol{g}$ that minimizes $R^{y^*}_{\perp}(\boldsymbol{g}|\boldsymbol{x})$ but does not minimizes $R^{y^*}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$ according to the lemma above. According to the consistency of the three Ψ used in our paper, we can directly get the consistency of our loss according to the definition of $R^{y^*}_{\Psi}(\boldsymbol{g}|\boldsymbol{x})$.

D Proof of Theorem 2 and Corollary 1

Proof. First of all, notice that the first K coordinates of Ψ_{CE}, Ψ_{OvA}, and Ψ_{ASM} are monotonic decreasing, so for any $\boldsymbol{g}$ that $\mathrm{argmax}_{y\in[K]} g_y(\boldsymbol{x}) \neq y^*$, we can swap them and get $\boldsymbol{g}'$ to make the regret smaller. Then we begin with discussing the point-wise regret. In this proof, we denote by $y' = \mathrm{argmax}_{y\in[K]} g_y(\boldsymbol{x})$. $g^*_{\mathrm{CE/ovA/ASM}}$ is the optimal solution *w.r.t.* $\tilde{\ell}_{\mathrm{CE/OvA/ASM}}$, and $g^{i*}_{\mathrm{CE/OvA/ASM}}$ is the optimal solution *w.r.t.* $\tilde{\ell}^{i}_{\mathrm{CE/OvA/ASM}}$. We also denote by $s_y(\boldsymbol{x})$ the y-th output of symmetric softmax *w.r.t.* $\boldsymbol{g}$ and $a_y(\boldsymbol{x})$ the y-th output of asymmetric softmax. Finally, we denote by $y' = \mathrm{argmax}_{y\in[K]} g_y(\boldsymbol{x})$.

D.1 Case of Ψ =CE

First of all, we set Ψ to the CE loss. We consider the two cases of error that:

- $\mathrm{Acc}(\boldsymbol{x}) > p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c, g_{K+1}(\boldsymbol{x}) > g_{y'}(\boldsymbol{x})$.

  Denote by $c(1-\mathrm{Acc}(\boldsymbol{x})) = \epsilon(\boldsymbol{x})$. In these cases, we can learn from our previous discussion and Pinsker's

inequality that for CE:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}(\boldsymbol{x}),y,m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}),y,m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}'(\boldsymbol{x}),y,m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}),y,m)] \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - s_{y^*}(\boldsymbol{x})\right| + \left|\frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - s_{K+1}(\boldsymbol{x})\right|\right)^2 \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} + s_{K+1}(\boldsymbol{x}) - s_{y^*}(\boldsymbol{x})\right|\right)^2 \\
&= \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right| + |s_{K+1}(\boldsymbol{x}) - s_{y^*}(\boldsymbol{x})|\right)^2 \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left|\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right|^2 \\
&= \frac{1}{2\tilde{\xi}(\boldsymbol{x})}\left|p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x}) - (1-c)\mathrm{Acc}(\boldsymbol{x})\right|^2 \\
&\geq \frac{1}{4}\left|p(y^*|\boldsymbol{x}) + c - \mathrm{Acc}(\boldsymbol{x})\right|^2
\end{aligned}
$$

Details: The first inequality holds due to our discussion in the first paragraph of this section. The second one holds due to Pinsker's inequality. The third one holds due to the absolute value inequality. The fourth equation holds according to (1). The last inequality holds since $2 \geq \tilde{\xi}(\boldsymbol{x}) \geq 1$ and the definition of $\epsilon(\boldsymbol{x})$.

- We further consider a special case that $\epsilon(\boldsymbol{x}) \neq 0$ and use the inequalities that $x \geq \log(1+x) \geq \frac{x}{1+x}$:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}(\boldsymbol{x}),y,m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}),y,m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y'}_{\mathrm{CE}}(\boldsymbol{g}^{y'*}_{\mathrm{CE}}(\boldsymbol{x}),y,m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}),y,m)] \\
&= -(p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x}))\log\left(\frac{p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) - p(y^*|\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&+ p(y'|\boldsymbol{x})\log\left(\frac{p(y'|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right) + (p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x}))}{\tilde{\xi}(\boldsymbol{x})}\right) \\
&= -p(y'|\boldsymbol{x})\log\left(\frac{p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y'|\boldsymbol{x})}\right) + p(y^*|\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y^*|\boldsymbol{x})}\right) \\
&+ \epsilon(\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x})}\right) \\
&= (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y^*|\boldsymbol{x})}\right) + (p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x})}\right) - p(y'|\boldsymbol{x})\log\left(\frac{p(y^*|\boldsymbol{x})}{p(y'|\boldsymbol{x})}\right) \\
&= (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y^*|\boldsymbol{x})}\right) \\
&+ \underbrace{(p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x}))\log\left(1+\frac{p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x})}{p(y'|\boldsymbol{x})+\epsilon(\boldsymbol{x})}\right) - p(y'|\boldsymbol{x})\log\left(1+\frac{p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x})}{p(y'|\boldsymbol{x})}\right)}_{\geq 0} \\
&\geq (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\log\left(\frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{p(y^*|\boldsymbol{x})}\right) \\
&\geq (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\frac{\epsilon(\boldsymbol{x})}{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})} \\
&\geq (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\frac{\epsilon(\boldsymbol{x})}{1+\epsilon(\boldsymbol{x})} \\
&= (p(y^*|\boldsymbol{x})-p(y'|\boldsymbol{x}))\left(1-\frac{1}{1+c-c\mathrm{Acc}(\boldsymbol{x})}\right)
\end{aligned}
$$

Then we can see a linear regret bound *w.r.t.* classifier's misclassification rate. Furthermore, a lower expert accuracy can make this bound tighter.

- Then we consider the next case of error:
  $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $p(y'|\boldsymbol{x}) < p(y^*|\boldsymbol{x})$.
  This case can also be split into two cases. First:
  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) > p(y^*|\boldsymbol{x})$:
    In this case, we can use the last inequality in the previous case:

$$\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}), y, m)] \geq (p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x}))\left(1 - \frac{1}{1 + c - c\mathrm{Acc}(\boldsymbol{x})}\right) \\
&\geq (p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x}))\left(\frac{p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x})}{1 + c - c\mathrm{Acc}(\boldsymbol{x})}\right) \\
&\geq \frac{(p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x}))^2}{2}
\end{aligned}$$

  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x})$.

$$\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y'}_{\mathrm{CE}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y^*}_{\mathrm{CE}}(\boldsymbol{g}^{y^*}_{\mathrm{CE}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y'}_{\mathrm{CE}}(\boldsymbol{g}^{y'*}_{\mathrm{CE}}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y^*}_{\mathrm{CE}}(\boldsymbol{g}^*_{\mathrm{CE}}(\boldsymbol{x}), y, m)] \\
&= -\sum_{y\in[K], y\neq y'} p(y|\boldsymbol{x}) \log(p(y|\boldsymbol{x})) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x}) \log(p(y|\boldsymbol{x})) \\
&- (p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \log(p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \log(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \\
&\geq -\sum_{y\in[K], y\neq y'} (p(y|\boldsymbol{x}) + \mathbb{I}(y = y^*)\epsilon(\boldsymbol{x})) \log(p(y|\boldsymbol{x})) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x}) \log(p(y|\boldsymbol{x})) \\
&- p(y'|\boldsymbol{x}) \log(p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \log(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}))
\end{aligned}$$

    The last inequality holds according to the description of this case. Then we define a dummy $K+1$ class distribution with class posterior probability $[\frac{p(1|\boldsymbol{x})}{1+\epsilon(\boldsymbol{x})}, \cdots, \frac{p(y^*|\boldsymbol{x})+\epsilon(\boldsymbol{x})}{1+\epsilon(\boldsymbol{x})}, \cdots, 0]$ and we can apply Pinsker's inequality according to the inequality above and get a quadratic bound:

$$\begin{aligned}
&-\sum_{y\in[K], y\neq y'} (p(y|\boldsymbol{x}) + \mathbb{I}(y = y^*)\epsilon(\boldsymbol{x})) \log(p(y|\boldsymbol{x})) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x}) \log(p(y|\boldsymbol{x})) \\
&- p(y'|\boldsymbol{x}) \log(p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \log(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \\
&= -\sum_{y\in[K], y\neq y'} (p(y|\boldsymbol{x}) + \mathbb{I}(y = y^*)\epsilon(\boldsymbol{x})) \log(\frac{p(y|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x}) \log(\frac{p(y|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}) \\
&- p(y'|\boldsymbol{x}) \log(\frac{p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \log(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}) \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - s_{y^*}(\boldsymbol{x})\right| + \left|\frac{p(y'|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - s_{y'}(\boldsymbol{x})\right|\right)^2 \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - s_{y^*}(\boldsymbol{x}) - \frac{p(y'|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} + s_{y'}(\boldsymbol{x})\right)^2 \\
&\geq \frac{\tilde{\xi}(\boldsymbol{x})}{2}\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})} - \frac{p(y'|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right)^2 \\
&\geq \frac{1}{4}\left(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) - p(y'|\boldsymbol{x})\right)^2 \\
&\geq \frac{1}{4}\left(p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x})\right)^2
\end{aligned}$$

And we can generalize the point-wise regret to the distribution using Jensen's inequality for expectation and concave function. □

D.2 Case of Ψ =ASM

According to our proof of Theorem 1, the linear regret case can be directly deduced. Then we focus on the quadratic bound.

- $\mathrm{Acc}(\boldsymbol{x}) > p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c, g_{K+1}(\boldsymbol{x}) > g_{y'}(\boldsymbol{x})$.
  Denote by $\xi(\boldsymbol{x}) = 1 + c - c\mathrm{Acc}(\boldsymbol{x}) \in [1, 1+c]$:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}^*_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}'(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}^*_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] \\
&\geq \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y^*}(\boldsymbol{x})\right|\right)^2 + \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{K+1}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y^*}(\boldsymbol{x})\right| + \left|\frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{K+1}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} + a_{K+1}(\boldsymbol{x}) - a_{y^*}(\boldsymbol{x}) - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})}\right)^2 \\
&\geq \frac{1}{2}\left(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) - (1-c)\mathrm{Acc}(\boldsymbol{x})\right)^2 \\
&= \frac{1}{2}\left(p(y^*|\boldsymbol{x}) + c - \mathrm{Acc}(\boldsymbol{x})\right)^2
\end{aligned}
$$

  Details: The first inequality holds due to our discussion in the first paragraph of this section. The second one holds due to Pinsker's inequality for the classifier and rejector counterpart, respectively. The third one holds due to Holder's inequality. The fourth one holds due to the absolute value inequality. The fourth equation holds according to (1).

  Then we consider the next case of error:

- $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $p(y'|\boldsymbol{x}) < p(y^*|\boldsymbol{x})$.
  We can also split it into two cases.
  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) > p(y^*|\boldsymbol{x})$:
    The result of this case is the same as in the case for CE since they share the same linear lower bound.
  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x})$:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}^*_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y'}_{\mathrm{ASM}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y^*}_{\mathrm{ASM}}(\boldsymbol{g}^{y^*}_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y'}_{\mathrm{ASM}}(\boldsymbol{g}^{y'*}_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}^{y^*}_{\mathrm{ASM}}(\boldsymbol{g}^*_{\mathrm{ASM}}(\boldsymbol{x}), y, m)] \\
&= -\sum_{y\in[K], y\neq y'} p(y|\boldsymbol{x})\log(p(y|\boldsymbol{x})) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x})\log(p(y|\boldsymbol{x})) \\
&- (p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}))\log(p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}))\log(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \\
&\geq -\sum_{y\in[K], y\neq y'} (p(y|\boldsymbol{x}) + \mathbb{I}(y = y^*)\epsilon(\boldsymbol{x}))\log(p(y|\boldsymbol{x})) + \sum_{y\in[K], y\neq y^*} p(y|\boldsymbol{x})\log(p(y|\boldsymbol{x})) \\
&- p(y'|\boldsymbol{x})\log(p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) + (p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}))\log(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})) \\
&\geq \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y^*}(\boldsymbol{x})\right| + \left|\frac{p(y^*|\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y'}(\boldsymbol{x})\right|\right)^2 \\
&\geq \frac{\xi(\boldsymbol{x})}{4}\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} + a_{K+1}(\boldsymbol{x}) - a_{y^*}(\boldsymbol{x}) - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})}\right)^2 \\
&= \frac{1}{4}\left(p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x})\right)^2
\end{aligned}
$$

  Then we can conclude the proof.

D.3 Case of Ψ =OvA

We consider the first type of error:

- $\mathrm{Acc}(\boldsymbol{x}) > p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c, g_{K+1}(\boldsymbol{x}) > g_{y'}(\boldsymbol{x})$.
  Denote by $\xi(\boldsymbol{x}) = 1 + c - c\mathrm{Acc}(\boldsymbol{x}) \in [1, 1+c]$:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g}^*_{\mathrm{OvA}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g'}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{ASM}}(\boldsymbol{g}^*_{\mathrm{OvA}}(\boldsymbol{x}), y, m)] \\
&\geq \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y^*}(\boldsymbol{x})\right|\right)^2 + \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{K+1}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{y^*}(\boldsymbol{x})\right| + \left|\frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})} - a_{K+1}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} + a_{K+1}(\boldsymbol{x}) - a_{y^*}(\boldsymbol{x}) - \frac{(1-c)\mathrm{Acc}(\boldsymbol{x})}{\xi(\boldsymbol{x})}\right)^2 \\
&\geq \frac{1}{2}\left(p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) - (1-c)\mathrm{Acc}(\boldsymbol{x})\right)^2 \\
&= \frac{1}{2}\left(p(y^*|\boldsymbol{x}) + c - \mathrm{Acc}(\boldsymbol{x})\right)^2
\end{aligned}
$$

  Details: The derivation is similar to that of ASM since they both share a structure that can separate the K-class classification task and training of deferral rule.

- The linear regret bound can be derived similarly: according to the discussion in the proof of Theorem 1, we can find the first two lines of the regret of OvA is the same with the regret for CE, and thus we can get the same lower bound. Notice that the last two lines of the regret of OvA can also be lower-bounded by the same lower bound if we substitute $p(y^*|\boldsymbol{x})$ and $p(i|\boldsymbol{x})$ with $1 - p(i|\boldsymbol{x})$ and $1 - p(y^*|\boldsymbol{x})$ respectively in the derivation of the first two lines' lower bound.

- $\mathrm{Acc}(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x}) + c$, $g_{K+1}(\boldsymbol{x}) \leq g_{y'}(\boldsymbol{x})$, $p(y'|\boldsymbol{x}) < p(y^*|\boldsymbol{x})$.
  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) > p(y^*|\boldsymbol{x})$:
    The result of this case is the same as in the case for CE since they share the same linear lower bound.
  - When $p(y'|\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \leq p(y^*|\boldsymbol{x})$:

$$
\begin{aligned}
&\mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g}^*_{\mathrm{OvA}}(\boldsymbol{x}), y, m)] \\
&\geq \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g'}(\boldsymbol{x}), y, m)] - \mathbb{E}_{p(y,m|\boldsymbol{x})}[\tilde{\ell}_{\mathrm{OvA}}(\boldsymbol{g}^*_{\mathrm{OvA}}(\boldsymbol{x}), y, m)] \\
&\geq \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - \sigma_{y^*}(\boldsymbol{x})\right|\right)^2 + \frac{\xi(\boldsymbol{x})}{2}\left(\left|\frac{p(y'|\boldsymbol{x})}{\xi(\boldsymbol{x})} - \sigma_{y'}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\left|\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - \sigma_{y^*}(\boldsymbol{x})\right| + \left|\frac{p(y'|\boldsymbol{x})}{\xi(\boldsymbol{x})} - \sigma_{y'}(\boldsymbol{x})\right|\right)^2 \\
&\geq \xi(\boldsymbol{x})\left(\frac{p(y^*|\boldsymbol{x}) + \epsilon(\boldsymbol{x})}{\xi(\boldsymbol{x})} - \sigma_{y^*}(\boldsymbol{x}) - \frac{p(y'|\boldsymbol{x})}{\xi(\boldsymbol{x})} + \sigma_{y'}(\boldsymbol{x})\right)^2 \\
&= \frac{1}{2}\left(p(y^*|\boldsymbol{x}) - p(y'|\boldsymbol{x})\right)^2
\end{aligned}
$$

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   Please refer to the definitions and description of each theorem/lemma.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**]
   We provide the regret transfer bound to further quantify our methods in Theorem 2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**No**]
   The loss functions in our paper can be easily implemented and we will include a demo for both datasets in the future version.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**]
   (c) Clear explanations of any assumptions. [**Yes**]

   Please refer to the definitions and description of each theorem/lemma, and the proofs are included in the appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]

   Please refer to Section 5 for the detailed experimental setup.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**]
   The used models and data are cited in Zagoruyko and Komodakis (2016); Krizhevsky et al. (2009).
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]