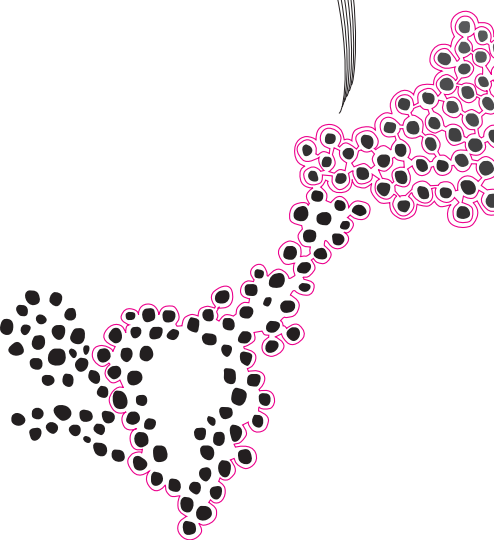




MSc Thesis Applied Mathematics

Leveraged Calibrated Loss for Learning to Defer

J.M. ter Steege



Supervisor: prof. dr. Johannes Schmidt-Hieber
dr. Hanyuan Hang

Assessment committee: prof. dr. Johannes Schmidt-Hieber
dr. Hanyuan Hang
dr. Georg Loho

March, 2023

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

Preface

This master thesis was written to fulfill the graduation requirements of the Master Applied Mathematics at the University of Twente, covering the period from September 2022 to March 2023.

First and foremost, I would like to express my sincere gratitude to my daily supervisor, Hongwei Wen, for his invaluable guidance and support throughout the entire research process. His insightful feedback, patience, and flexibility have been instrumental in shaping this thesis. I truly appreciate the time and effort that he has invested in this graduation project, which has created a motivating work environment for me. It was a pleasure to collaborate with him, and I enjoyed our stimulating conversations about both work and leisure activities, such as football.

I also extend my thanks to my graduation supervisor, dr. Hanyuan Hang, for providing me with an interesting and challenging graduation project, and for his guidance and feedback. I appreciate the opportunity to work on this project under his supervision.

Moreover, I would like to express my appreciation to the other members of my graduation committee, prof. dr. Johannes Schmidt-Hieber and dr. Georg Loho, for taking the time to read and evaluate my thesis.

I would like to acknowledge the unwavering support of my family and friends. My parents and grandpa have always been my pillars of strength and encouragement. I cannot thank them enough for their unconditional love and support throughout my academic journey. Finally, my friends have been my constant source of motivation and joy, and I appreciate the many fun times we have shared together. Special thanks to Maik and Peter for proof-reading my thesis.

In conclusion, I hope this thesis will reflect the effort and hard work put into this project, and I thank all those who have contributed to its completion.

Contents

1	Introduction	1
2	Related Work	3
3	Preliminaries	4
3.1	Classification problem	4
3.2	Learning to Defer	4
4	Methodology	6
4.1	Leveraged One-vs-All (LOvA) Surrogate Loss	6
4.2	Optimal decision rules	7
5	Theoretical Results	9
5.1	Bayes consistency of LOvA loss	9
5.1.1	Definition of Bayes consistency in L2D	9
5.1.2	Strictly proper composite loss	10
5.1.3	Convex, differentiable, decreasing loss	11
5.2	Margin theory	12
5.2.1	Logistic loss	13
5.2.2	Square Loss	14
6	Experiments	15
7	Conclusion & Discussion	18
8	Proofs	19
8.1	Proofs Related to Section 4	19
8.2	Proofs Related to Section 5.1	22
8.2.1	Bayes Consistency and Classification Calibration	22
8.2.2	Proof of Theorem 3	24
8.2.3	Proof of Theorem 4	25
8.3	Proofs Related to Section 5.2	27
8.3.1	Proof of Theorem 5	27
8.3.2	Proof of Theorem 6	27

Leveraged Calibrated Loss for Learning to Defer

J.M. ter Steege

March, 2023

Abstract

The Learning to Defer (L2D) framework is designed to enhance the safety of AI systems by incorporating human intervention in decision-making when it is likely to lead to more accurate results than the model alone. In this paper, we propose a new family of surrogate losses, called the *Leveraged One-vs-All (LOvA)* loss, which for the first time introduces a leverage parameter to consider the trade-off between expert correctness and incorrectness. Our theoretical analysis derives a generalized result for Bayes risk consistency of the LOvA loss in the L2D system, providing guidance for selecting the leverage parameter. Additionally, we establish that the decision *margin* increases, which lowers the misclassification rate, resulting in a more robust and deterministic classification by our system. In our experiments, we validate the guidance offered by our theoretical analysis and demonstrate that our proposed LOvA loss performs significantly better than other state-of-the-art L2D systems on real-world datasets.

Keywords: Learning to Defer (L2D), Artificial Intelligence (AI), Surrogate Loss, Bayes Risk Consistency, Machine Learning, Deep Learning

1 Introduction

Machine learning is becoming increasingly prevalent in various fields, such as healthcare [10, 11], autonomous driving [26] and the stock market [1]. However, complex algorithms in high-stakes scenarios are prone to overfitting or being too general for specific cases. This can result in inaccurate estimations, which may have serious consequences. To address this issue, human expertise can be utilized in uncertain cases, since humans often possess additional information that can aid decision-making.

Learning with a rejection option [5], also known as rejection learning, is a solution that allows the model to abstain from making a decision and defer the burden to a human. For example, a self-driving car’s algorithm may decline to make a decision when a road has a sharp turn and resume operation once the road becomes straight again. The rejection learning framework assumes a constant cost c of deferring, making the problem to predict if the model is $1-c$ confident. Approaches to tackle the rejection learning problem are categorized into two paradigms: confidence-based and classifier-rejector approaches. Confidence-based methods usually focus on the model’s uncertainty, while classifier-rejector methods learn the classifier and rejector simultaneously.

However, the existing rejection learning framework does not explore the interaction between the expert and the classifier. To address the drawback, the novel *Learning to Defer* (L2D) framework [14] takes the human expert’s prediction into account. Recently, there has been research into the L2D framework, particularly in a multiclass setting [16]. The proposed approach is based on a novel reduction to cost-sensitive learning and they propose a consistent surrogate loss function. However, this surrogate loss function has some drawbacks concerning the calibration of the expert probability, as discovered by Proposition 3.1 in [24]. It was also confirmed through experimentation that such a scenario was possible. Consequently, a different surrogate loss function based on One-vs-All classification was proposed [24]. This surrogate loss function is well-calibrated and consistent, but the classifier remains highly dependent on the human expert. We think it is possible to reduce this dependence, while simultaneously keeping the output accuracy high.

In this paper, we propose a new surrogate loss function called the *Leveraged One-vs-All* (*LOvA*) loss, which introduces a leverage parameter to consider the trade-off between expert correctness and incorrectness. We show that this loss function satisfies the Bayes consistency property and theoretically demonstrate that the decision margin increases as the leverage parameter enlarges under mild assumptions. We also conduct experiments to verify the theoretical results and compare the performance with other existing methods.

To summarize, the contributions of this paper are:

- We propose a new surrogate loss function for the L2D problem that introduces a leverage parameter, which decreases the proportion of deferring and increases the performance of the machine.
- We provide a theoretical analysis to prove the Bayes consistency of our proposed loss function and show an increased decision margin as the leverage parameter enlarges under mild assumptions.
- We explore the effect of different values of α on the system performance, classifier performance and proportion of deferred samples. Additionally, we compare the experimental results with other methods on the CIFAR-10 dataset.

2 Related Work

In 1957, rejection learning was introduced in [5], where optimal decision rules were found by minimizing the error rate for a fixed rejection rate. This sparked research into methods that required predetermined confidence rates, where the algorithm decides to reject when it is too uncertain. Since then, numerous confidence-based approaches for binary classification have been proposed [2, 25, 9, 19].

In the binary setting, [6] was the first to introduce the simultaneous learning of a classifier and a rejector. Later, in [17], both the confidence-based and classifier-rejector approaches were extended to a multiclass setting. [17] derived a general condition for calibration to the Bayes optimal solution for the classifier-rejector method, which suggested that calibration is hard to achieve with general loss functions. For the confidence-based technique, they proposed rejection criteria for more general losses and guaranteed calibration to the Bayes optimal solution.

While [17] extended existing work to the multiclass setting, [4] was the first to propose a surrogate loss inspired by the cost-sensitive learning for general classification. However, the above methods for rejection learning neglect the expert’s decision for the samples. To fill this gap, [14] for the first time proposed the L2D framework based on the classifier-rejector approach by exploring the interaction between the expert and the classifier. They found that L2D is a generalization of the previous rejection learning. However, the loss in [14] was found to be inconsistent, and [18] introduced a confidence-based method that compares the confidence levels of the classifier and expert to decide which to defer to. Unfortunately, [16] provided an example showing that this method fails to adapt to the expert’s strengths and weaknesses.

Closely related to our work, [16] proposed the consistent surrogate loss for the multiclass problem, namely the softmax loss. [24] found that the framework in [16] was not calibrated with respect to expert correctness and proposed a different consistent surrogate loss based on One-vs-All classifiers called the OvA loss. Our work aims to decrease expert dependency by adding a leverage term while maintaining the consistency of the proposed surrogate loss.

3 Preliminaries

In this section, we introduce the basic mathematical definitions and notations for the classification problem and Learning to Defer (L2D) problem based on it.

3.1 Classification problem

In machine learning, classification is the task of assigning input data to a certain class [3]. Examples of classification problems include spam detection [22] or handwritten digit recognition [13]. These are typical examples of binary and multiclass classification problems respectively within supervised machine learning.

In a general classification problem, we observe the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which are independently and identically distributed from an unknown probability distribution. Here, each input \mathbf{x}_i is a vector of d features in the input space $\mathcal{X} \subseteq \mathbb{R}^d$, and each label y_i is in the true output space $\mathcal{Y} = [K] := \{1, \dots, K\}$. Our goal is to learn a prediction function called the *classifier* $f: \mathcal{X} \rightarrow \mathcal{Y}$, which maps an input instance $\mathbf{x} \in \mathcal{X}$ to its corresponding label $y \in \mathcal{Y}$. To evaluate the performance of the classifier, we use the 0 – 1 loss function $\ell_{0-1}(y, f(\mathbf{x})) := \mathbb{1}(y \neq f(\mathbf{x}))$.

Let $\boldsymbol{\eta}(\mathbf{x})$ denote the posterior probability function as

$$\boldsymbol{\eta}(\mathbf{x}) := [\eta_k(\mathbf{x})]_{k=1}^K \quad \text{with} \quad \eta_k(\mathbf{x}) := \mathbb{P}(Y = k | X = \mathbf{x}) \quad \text{for } k \in [K] \text{ and } \mathbf{x} \in \mathcal{X}. \quad (1)$$

Then the probability simplex is the collection of all possible posterior probability vectors, denoted as

$$\mathcal{S}_K := \left\{ (\eta_1(\mathbf{x}), \dots, \eta_K(\mathbf{x})) \mid \mathbf{x} \in \mathcal{X}, \sum_{k=1}^K \eta_k(\mathbf{x}) = 1, 0 \leq \eta_k(\mathbf{x}) \leq 1, \forall k \in [K] \right\}.$$

The corresponding risk with respect to the loss ℓ_{0-1} is defined as

$$\mathcal{R}^{\ell_{0-1}}[f] := \mathbb{E}_{\mathbf{x}, y}[\ell_{0-1}(y, f(\mathbf{x}))] = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y|\mathbf{x}}[\mathbb{1}(y \neq f(\mathbf{x}))] = \sum_{k=1}^K \eta_k(\mathbf{x}) \mathbb{P}_{\mathbf{x}}(f(\mathbf{x}) \neq k), \quad (2)$$

where $\mathbb{P}_{\mathbf{x}}(f(\mathbf{x}) \neq k)$ is the probability that the classifier f makes a mistake when predicting the label of input \mathbf{x} as k . The minimal ℓ_{0-1} -risk is called the *Bayes risk*, which is given by

$$\mathcal{R}^{\ell_{0-1},*} := \inf_{f: \mathcal{X} \rightarrow \mathcal{Y}} \mathcal{R}^{\ell_{0-1}}[f].$$

The classifier that achieves the Bayes risk is called the *Bayes classifier*, which is given by

$$f^*(\mathbf{x}) := \arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}).$$

3.2 Learning to Defer

In addition to the standard classification samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the L2D problem assumes access expert demonstrations. The expert's prediction for feature \mathbf{x}_i is denoted as $m_i \in \mathcal{M}$, where \mathcal{M} is the expert's prediction space. Usually, we take $\mathcal{M} = \mathcal{Y}$. Each sample in the dataset $\mathcal{D} = \{\mathbf{x}_i, y_i, m_i\}_{i=1}^n$ is drawn from the same distribution \mathbb{P} of $(\mathcal{X}, \mathcal{Y}, \mathcal{M})$.

Learning to Defer

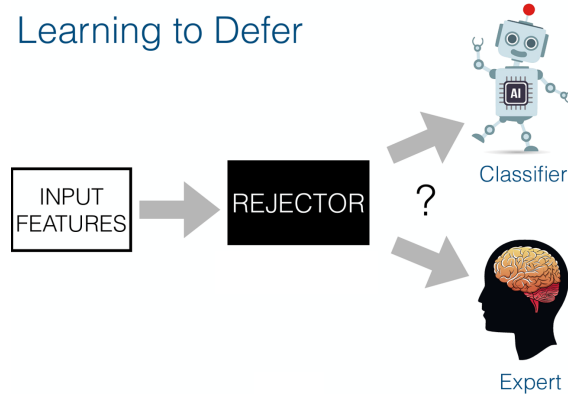


FIGURE 1: Schematic overview of the L2D framework [24].

We denote (X, Y, M) as the variable distributed from \mathbb{P} . Additionally, the probability that the expert correctly classifies the input \mathbf{x} given the true label $k \in \mathcal{Y}$ is denoted as

$$p_{m,k}(\mathbf{x}) = \mathbb{P}(M = k | X = \mathbf{x}, Y = k). \quad (3)$$

while the probability that the expert's prediction matches the true label Y is denoted as

$$p_m(\mathbf{x}) = \mathbb{P}(M = Y | X = \mathbf{x}). \quad (4)$$

The L2D framework is a classification system with a rejection option, where both a *classifier* and a *rejector* are learned simultaneously. The *rejector* $r : \mathcal{X} \rightarrow \{0, 1\}$ determines whether the expert should be consulted based on confidence. If the machine is more certain about deferring than about classifying one of K labels, it decides to defer. The *classifier* $f : \mathcal{X} \rightarrow \mathcal{Y}$ is used when the rejector decides not to defer, i.e. when $r(\mathbf{x}) = 0$. Figure 1 provides a schematic overview.

To evaluate the classification performance of the L2D system, which applies either the classifier $f(\mathbf{x})$ or the expert m as the prediction according to the rejector's decision $r(\mathbf{x})$, we introduce the 0 – 1 loss within the L2D framework. The loss function is defined as

$$L_{0-1}(y, m, f(\mathbf{x}), r(\mathbf{x})) := (1 - r(\mathbf{x}))\mathbb{1}[f(\mathbf{x}) \neq y] + r(\mathbf{x})\mathbb{1}[m \neq y], \quad (5)$$

where $\mathbb{1}$ denotes the indicator function that checks if the prediction and label are equal. For a fixed classifier f and a rejector r , the corresponding risk with respect to the L_{0-1} loss is defined as

$$\mathcal{R}^{L_{0-1}}[f, r] := \mathbb{E}_{\mathbf{x}, y, m}[L_{0-1}(y, m, f(\mathbf{x}), r(\mathbf{x}))].$$

The minimal L_{0-1} -risk is called the *Bayes risk* and is given by

$$\mathcal{R}^{L_{0-1},*} := \inf_{f: \mathcal{X} \rightarrow \mathcal{Y}, r: \mathcal{X} \rightarrow \{0,1\}} \mathcal{R}^{L_{0-1}}[f, r].$$

The classifier and rejector that achieve the Bayes risk are called the *Bayes classifier* and *Bayes rejector*, respectively. Equation 4 in [24] tells us that the Bayes classifier and the Bayes rejector of the L_{0-1} loss (5) are

$$f^*(\mathbf{x}) = \arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}),$$

$$r^*(\mathbf{x}) = \mathbb{1} \left[p_m(\mathbf{x}) \geq \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \right].$$

4 Methodology

In this section, we introduce our proposed Leveraged One-vs-All (LOvA) surrogate loss function for the L2D problem in Section 4.1, and present the corresponding Bayes score function and optimal decision rules in Section 4.2.

4.1 Leveraged One-vs-All (LOvA) Surrogate Loss

The *Leveraged One-vs-All* (LOvA) loss is inspired by [24] and uses $K + 1$ score functions $g_1(\mathbf{x}), \dots, g_K(\mathbf{x})$, and $g_\perp(\mathbf{x})$, where $g_y : \mathcal{X} \rightarrow \mathbb{R}$ for $y \in \mathcal{Y}^\perp := \{1, \dots, K, \perp\}$. The score function vector $\mathbf{g}(\mathbf{x}) := [g_y(\mathbf{x})]_{y \in \mathcal{Y}^\perp}$ indicates the likelihood of \mathbf{x} being labeled as $y \in \mathcal{Y}^\perp$. If $g_\perp(\mathbf{x})$ is the largest score function, \mathbf{x} is deferred to the expert for a decision. Given the fitted score function \mathbf{g} , the classifier and rejector are given by

$$\begin{aligned} f(\mathbf{x}) &= \arg \max_{y \in \mathcal{Y}} g_y(\mathbf{x}), \\ r(\mathbf{x}) &= \mathbb{1} \left[g_\perp(\mathbf{x}) \geq \max_{y \in \mathcal{Y}} g_y(\mathbf{x}) \right]. \end{aligned} \tag{6}$$

Our LOvA loss function introduces a leverage parameter α into a surrogate loss function for L2D, which leverages losses on expert correctness and incorrectness. The loss function has the following point-wise form:

$$\begin{aligned} \psi_L(y, m, \mathbf{g}(\mathbf{x})) &= \phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})] \\ &\quad + \mathbb{1}[m = y] (\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})] + \alpha \phi[g_y(\mathbf{x})]), \end{aligned} \tag{7}$$

where y denotes the correct label and y' denotes the wrong labels. It consists of three components:

- (1) A binary loss function $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$, where $\phi[g_y(\mathbf{x})]$ forces g_y to be larger when y is the correct label, while $\phi[-g_{y'}(\mathbf{x})]$ punishes large $g_{y'}$ for wrong label y' .
- (2) An indicator function $\mathbb{1}$ that determines whether the expert would predict the right label ($m = y$) or wrong label ($m \neq y$).
- (3) The leverage parameter α distinguishes between the score functions of correct and incorrect labels. Larger α causes the score function of the correct label to enlarge and the score functions of incorrect labels to reduce.

We highlight that this is the first time that the leverage parameter α is introduced into surrogate loss functions for the L2D problem. The LOvA loss function of (7) is in a general form, but some special cases are worth mentioning:

- (i) Taking $\alpha = 0$, we achieve the OvA surrogate loss proposed by [24], the form of which is

$$\begin{aligned} \psi_{\text{OvA}}(y, m, \mathbf{g}(\mathbf{x})) &= \phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})] \\ &\quad + \mathbb{1}[m = y] (\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})]). \end{aligned}$$

- (ii) Taking $m \neq y$, the part consisting of the leverage term α disappears and we end up with

$$\psi_L(y, m, \mathbf{g}(\mathbf{x})) = \phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})].$$

- (iii) Taking $m = y$ and $\alpha \neq 0$, we achieve the case of interest, where our leverage term α has utility. The loss has the following form

$$\psi_L(y, m, \mathbf{g}(\mathbf{x})) = (1 + \alpha)\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[g_\perp(\mathbf{x})].$$

As can be seen, the leverage term α is only applicable in the third case, so when the expert's prediction is right. In the context of minimizing the loss, a larger α puts more weight on the score function of the correct label g_y . Consequently, we expect that the score function g_y will increase when it is correctly labelled, while simultaneously the other score functions $g_{y'}$ will reduce. Experiments should be executed to confirm this estimation, but first of all we will show the theoretical consequences of our added leverage term α in Section 5.

4.2 Optimal decision rules

The optimal classifier and rejector are induced by the optimal score function \mathbf{g}^* that minimizes the corresponding risk $\mathbb{E}_{\mathbf{x}, y, m}[\psi_L]$. We will evaluate the integrated binary loss ϕ to be the logistic loss in order to get these decision rules.

Theorem 1. Let ϕ be the logistic loss integrated in the LOvA loss (7). Moreover, let $\eta_k(\mathbf{x})$, $p_{m,k}(\mathbf{x})$ and $p_m(\mathbf{x})$ be defined as in (1), (3) and (4) respectively. Then the optimal score function \mathbf{g}^* has the form

$$\begin{aligned} g_k^*(\mathbf{x}) &= \log\left(\frac{\eta_k(\mathbf{x})}{1 - \eta_k(\mathbf{x})} \cdot (1 + \alpha p_{m,k}(\mathbf{x}))\right), \quad \forall k \in \mathcal{Y}, \\ g_\perp^*(\mathbf{x}) &= \log\left(\frac{p_m(\mathbf{x})}{1 - p_m(\mathbf{x})}\right). \end{aligned} \tag{8}$$

The proof of this theorem is given in Section 8.1.

As described in (6), the largest score function determines which action will be taken. Therefore, the optimal decision rules for the LOvA loss are determined by the optimal score function in (8) in the same way, i.e.

$$\begin{aligned} f_L^*(\mathbf{x}) &:= \arg \max_{k \in \mathcal{Y}} g_k^*(\mathbf{x}), \\ r_L^*(\mathbf{x}) &:= \mathbf{1}\left[g_\perp^*(\mathbf{x}) \geq \max_{k \in \mathcal{Y}} g_k^*(\mathbf{x})\right]. \end{aligned} \tag{9}$$

Using (9) and Theorem 1, we can derive the explicit formulation of the Bayes decision rules, which enables us to determine the specific decision rules for the LOvA loss equipped with the logistic loss. However, before proceeding, we need to make an assumption about the maximum probability $p_{m,k}$.

Assumption 1. Let $\eta_k(\mathbf{x})$ be the posterior probability function defined as in (1) and let $p_{m,k}(\mathbf{x})$ be defined as in (3). Assume that for any \mathbf{x} , there holds

$$\arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \subset \arg \max_{k \in \mathcal{Y}} p_{m,k}(\mathbf{x}).$$

Assumption 1 assumes that the class with the largest posterior probability $\eta_k(\mathbf{x})$ is the class with the largest probability $p_{m,k}(\mathbf{x})$.

Theorem 2. Let ϕ be the logistic loss in the LOvA loss (7). Moreover, let $p_{m,k}(\mathbf{x})$ and $p_m(\mathbf{x})$ be defined as in (3) and (4), respectively. Suppose that Assumption 1 is satisfied. Then the optimal decision rules for the LOvA loss are given by

$$\begin{aligned} f_L^*(\mathbf{x}) &= \arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}), \\ r_L^*(\mathbf{x}) &= \mathbb{1} \left[p_m(\mathbf{x}) \geq \frac{R(\mathbf{x})}{1 + R(\mathbf{x})} \right], \end{aligned} \tag{10}$$

where

$$R(\mathbf{x}) := \max_{k \in \mathcal{Y}} \frac{\eta_k(\mathbf{x})}{1 - \eta_k(\mathbf{x})} (1 + \alpha p_{m,k}(\mathbf{x})).$$

The proof of this theorem is given in Section 8.1.

These decision rules can be used for each possible observed training data \mathcal{D} to classify new data \mathbf{x} or to decide to reject and refer to the expert. Note that the decision rules in (10) are optimal for the logistic loss ϕ integrated in the LOvA loss (7). The optimal decision rules for other integrated losses can be computed similarly.

The optimal rejector $r_L^*(\mathbf{x})$ as in (10) is depending on the leverage term α . For a given instance \mathbf{x} , a larger α causes a larger rejection threshold $\frac{R(\mathbf{x})}{1+R(\mathbf{x})}$. Therefore, the system is less likely to reject new instances when α increases and is more dependent on the classifier $f_L^*(\mathbf{x})$.

The addition of the leverage term reduces the workload of the expert, but this does not say anything about the performance of the system. It is therefore not clear what value of α should be used in order to maximize the system accuracy. Consequently, we must conduct experiments for different values of α to test how large the leverage term should be. We will show these experimental consequences in Section 6.

5 Theoretical Results

In this section, we explore two important theoretical properties of the LOvA loss. More precisely, in Section 5.1, we prove that when equipped with two commonly-used binary losses, the LOvA loss is Bayes consistent with respect to the ℓ_{0-1} -loss for any $\alpha \geq 0$. While this result provides valuable insights into the performance of the LOvA loss, it does not necessarily shed light on the necessity of the leverage parameter α . To address this, we move on to Section 5.2, where we investigate how different values of α impact the margin of the Bayes score functions of the LOvA loss. This analysis provides a more complete picture of the role of α in the performance of the LOvA loss.

5.1 Bayes consistency of LOvA loss

In this section, we will first define Bayes consistency in the L2D problem, and then demonstrate that our proposed LOvA loss serves as a Bayes consistent surrogate loss for the 0–1 L2D loss in (5), using two commonly-used binary losses.

5.1.1 Definition of Bayes consistency in L2D

First, we define some notations for a general surrogate loss $\psi : \mathcal{Y} \times \mathcal{M} \times \mathbb{R}^{K+1} \rightarrow \mathbb{R}_+$, designed for $K+1$ score functions $\mathbf{g}(\mathbf{x}) := [g_y(\mathbf{x})]_{y \in \mathcal{Y}}$. The ψ -risk and Bayes ψ -risk of \mathbf{g} , denoted as \mathcal{R}^ψ and $\mathcal{R}^{\psi,*}$, respectively, are defined by

$$\begin{aligned} \mathcal{R}^\psi(\mathbf{g}) &:= \mathbb{E}_{\mathbf{x}, y, m}[\psi(y, m, \mathbf{g}(\mathbf{x}))], \\ \mathcal{R}^{\psi,*} &:= \min_{\mathbf{g}: \mathcal{X} \rightarrow \mathbb{R}^{K+1}} \mathcal{R}^\psi(\mathbf{g}). \end{aligned}$$

The optimal score functions that achieve the Bayes ψ -risk are defined by

$$\mathbf{g}^{\psi,*} := \arg \min_{\mathbf{g}: \mathcal{X} \rightarrow \mathbb{R}^{K+1}} \mathcal{R}^\psi(\mathbf{g}).$$

In order to estimate the optimal score function $\mathbf{g}^{\psi,*}$, we minimize the empirical ψ -risk over the function space \mathcal{F}_n , i.e.

$$\hat{\mathbf{g}}^\psi := \arg \min_{\mathbf{g} \in \mathcal{F}_n} \hat{\mathcal{R}}^\psi(\mathbf{g}) := \arg \min_{\mathbf{g} \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n \psi(y_i, m_i, \mathbf{g}(\mathbf{x}_i)). \quad (11)$$

Correspondingly, based on the estimated score functions $\hat{\mathbf{g}}^\psi$, we get the classifier

$$\hat{f}^\psi := \arg \max_{k \in \mathcal{Y}} \hat{g}_k^\psi(\mathbf{x}).$$

Under suitable conditions for a surrogate loss ψ , minimizing its empirical risk $\hat{\mathcal{R}}^\psi(\mathbf{g})$ over a sequence of function classes \mathcal{F}_n approximately minimizes the ψ -risk $\mathcal{R}^\psi(\mathbf{g})$. However, one goal in the L2D problem is to find a classifier f whose classification risk $\mathcal{R}^{\ell_{0-1}}[f]$ in (2) (called the *risk* of f) is close to the possible minimum, i.e. the Bayes risk $\mathcal{R}^{\ell_{0-1},*}$. Therefore, we investigate the conditions which guarantee that if the ψ -risk of \mathbf{g} gets close to its Bayes ψ -risk, then the risk of the induced classifier $f(\mathbf{x}) = \arg \max_{k \in \mathcal{Y}} g_k(\mathbf{x})$ also approaches the Bayes risk. When this happens, we say that the L2D classification method based on ψ is Bayes consistent. Mathematically speaking, we provide the formal definition as follows:

Definition 1. (Bayes consistency) Let $\psi : \mathcal{Y} \times \mathcal{M} \times \mathbb{R}^{K+1} \rightarrow \mathbb{R}_+$ be a surrogate loss function for the L2D problem. The surrogate loss function ψ is said to be Bayes consistent with respect to the classification loss ℓ_{0-1} if for any sequence of score functions $\mathbf{g}_n : \mathcal{X} \rightarrow \mathbb{R}^{K+1}$ with $\mathbf{g}_n(\mathbf{x}) := [g_{n,y}(\mathbf{x})]_{y \in \mathcal{Y}^\perp}$, the following holds:

$$\mathcal{R}^\psi(\mathbf{g}_n) \rightarrow \mathcal{R}^{\psi,*} \implies \mathcal{R}^{\ell_{0-1}}(f_n) \rightarrow \mathcal{R}^{\ell_{0-1},*},$$

where the classifier is

$$f_n(\mathbf{x}) := \arg \max_{k \in \mathcal{Y}} g_{n,k}(\mathbf{x}).$$

The Bayes consistency property is crucial in determining the success of a classifier learned by minimizing the surrogate loss. If the surrogate loss is Bayes consistent with respect to the classification loss ℓ_{0-1} , then the convergence of the surrogate ψ -risk to its Bayes ψ -risk implies the convergence of the original classification risk to its Bayes risk. This property enables us to solve the minimization of a surrogate loss instead of the 0 – 1 loss, providing theoretical guarantees for the usage of the surrogate loss function in the L2D problem.

Next, we equip our LOvA loss with two widely-used binary loss functions ϕ and then explore the Bayes consistency of the LOvA loss.

5.1.2 Strictly proper composite loss

The first binary loss function that we consider is the strictly proper composite loss [20], which is defined as follows:

Definition 2. (Strictly proper composite loss) Let $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$ be a binary surrogate loss. Then, ϕ is said to be *proper composite* if there exists a strictly increasing link function $\gamma : [0, 1] \mapsto \mathbb{R}$ such that for any $p(\mathbf{x}) := \mathbb{P}(Y = 1 | \mathbf{x}) \in (0, 1)$, there holds:

$$\gamma(p(\mathbf{x})) \in \arg \min_{f(\mathbf{x}) \in \mathbb{R}} p(\mathbf{x})\phi[f(\mathbf{x})] + (1 - p(\mathbf{x}))\phi[-f(\mathbf{x})].$$

Moreover, ϕ is said to be *strictly proper composite* if the above minimizer is unique for all $p(\mathbf{x}) \in (0, 1)$.

This definition shows that if a binary surrogate loss is strictly proper composite, we can use the associated link function to find the unique minimizer of a binary classification problem. This is a useful trick that we will use to prove the consistency of our LOvA loss in Theorem 3.

Theorem 3. Let ψ_L (7) be the LOvA loss equipped with a strictly proper composite loss ϕ , and let $p_{m,k}(\mathbf{x})$ be defined as in (3). Suppose that Assumption 1 is satisfied. Then, for any $\alpha \geq 0$, ψ_L is Bayes consistent with respect to the classification loss ℓ_{0-1} .

The complete proof is provided in Section 8.2.2. Theorem 3 shows that, under a reasonable assumption, our surrogate loss is consistent with the non-convex 0 – 1 loss. As mentioned earlier, this is especially useful in the learning phase, where finding the minimizer of the 0 – 1 loss appears to be NP-hard.

Some examples of strictly proper composite losses are the logistic loss, exponential loss, and square loss. The logistic loss is often used in LogitBoost [8] and the exponential loss in AdaBoost [7], while the square loss is frequently used for regression but can also be applied to classification [15]. In this context, link functions play a crucial role by connecting the output of a model to a target variable. Here are these examples:

Loss name	$\phi(v)$	$\gamma(p)$	$\gamma^{-1}(v)$
Logistic	$\log(1 + \exp(-v))$	$\log(\frac{p}{1-p})$	$\frac{1}{1+\exp(-v)}$
Exponential	$\exp(-v)$	$\frac{1}{2} \log(\frac{p}{1-p})$	$\frac{1}{1+\exp(-2v)}$
Square	$(1 - v)^2$	$2p - 1$	$\frac{1}{2}(v + 1)$

TABLE 1: Surrogate losses with their respective link and inverse link functions.

Example 1. (Logistic loss) This strictly proper composite loss function is defined as

$$\phi[f(\mathbf{x})] = \log(1 + \exp(-f(\mathbf{x}))).$$

The corresponding link function $\gamma(p) = \log(p/(1-p))$ is strictly increasing and minimizes the inner ψ -risk. Conversely, the inverse link function $\gamma^{-1}[f(\mathbf{x})] = 1/[1 + \exp(-f(\mathbf{x}))]$ can be used as an approximation of the posterior probability p .

Example 2. (Exponential loss) This strictly proper composite loss function is defined as

$$\phi[f(\mathbf{x})] = \exp(-f(\mathbf{x})).$$

The corresponding link function $\gamma(p) = \frac{1}{2} \log(p/(1-p))$ is strictly increasing and minimizes the inner ψ -risk. Conversely, the inverse link function $\gamma^{-1}[f(\mathbf{x})] = 1/[1 + \exp(-2f(\mathbf{x}))]$ can be used as an approximation of the posterior probability p .

Example 3. (Square loss) This strictly proper composite loss function is defined as

$$\phi[f(\mathbf{x})] = (1 - f(\mathbf{x}))^2.$$

The corresponding link function $\gamma(p) = 2p - 1$ is strictly increasing and minimizes the inner ψ -risk. Conversely, the inverse link function $\gamma^{-1}[f(\mathbf{x})] = \frac{1}{2}(f(\mathbf{x}) + 1)$ can be used as an approximation of the posterior probability p .

These three losses satisfy the conditions given in Theorem 3, and they are summarized in Table 1. By using one of these binary surrogate losses ϕ with our LOvA loss ψ_L (7), we can obtain a calibrated surrogate loss for the 0 – 1 classification loss.

5.1.3 Convex, differentiable, decreasing loss

Another group of commonly used loss functions that can be minimized are those that are convex, differentiable and decreasing. These types of loss functions can be minimized using standard optimization techniques [28]. Convex functions have the advantageous property that we only need to aim for local optima, as they are equivalent to global optima. This means that we can minimize our loss function and know that it is the lowest possible loss, making the decision that causes this minimal loss optimal. Additionally, since the loss is both differentiable and decreasing, we have the following theorem:

Theorem 4. Let ψ_L (7) be the LOvA loss equipped with a convex, differentiable and decreasing loss ϕ . Suppose that Assumption 1 is satisfied. Then, for any $\alpha \geq 0$, ψ_L is Bayes consistent with respect to the classification loss ℓ_{0-1} .

The complete proof is provided in Section 8.2.3. Theorem 4 allows for a variety of surrogate loss functions to be used to approximate the non-convex 0 – 1 loss. The following examples demonstrate some of these loss functions and their associated results.

Example 4. (Logistic loss) This convex, differentiable and decreasing loss function is defined as

$$\phi[f(\mathbf{x})] = \log(1 + \exp(-f(\mathbf{x}))).$$

For this logistic loss, we have the derivative $\phi'[f(\mathbf{x})] = -1/(1 + \exp(-f(\mathbf{x})))$. By applying the first-order optimality condition to a binary classification problem, we can find the minimizer of the risk. This gives $p(\mathbf{x})\phi'[g(\mathbf{x})] + (1 - p(\mathbf{x}))\phi'[-g(\mathbf{x})] = 0$ where $p(\mathbf{x}) := \mathbb{P}(Y = 1|\mathbf{x})$, which leads to the optimal score function $g(\mathbf{x}) = \log(p(\mathbf{x})/(1 - p(\mathbf{x})))$.

Example 5. (Exponential loss) This convex, differentiable and decreasing loss function is defined as

$$\phi[f(\mathbf{x})] = \exp(-f(\mathbf{x})).$$

For this exponential loss, we have the derivative $\phi'[f(\mathbf{x})] = -\exp(-f(\mathbf{x}))$. By applying the first-order optimality condition to a binary classification problem, we can find the minimizer of the risk. This gives $p(\mathbf{x})\phi'[g(\mathbf{x})] + (1 - p(\mathbf{x}))\phi'[-g(\mathbf{x})] = 0$ where $p(\mathbf{x}) := \mathbb{P}(Y = 1|\mathbf{x})$, which leads to the optimal score function $g(\mathbf{x}) = \frac{1}{2} \log(p(\mathbf{x})/(1 - p(\mathbf{x})))$.

These two losses satisfy the conditions given in Theorem 4. By using one of these binary surrogate losses ϕ with our LOvA loss ψ_L (7), we can obtain a calibrated surrogate loss for the 0–1 classification loss. Unfortunately, the square loss is increasing on $f(\mathbf{x}) \in [1, +\infty)$, so it does not belong to the type of loss functions we discussed in this section.

5.2 Margin theory

In this section, we explore the effect of the leverage parameter α on the performance of the LOvA loss from the perspective of margin theory. Theorem 3 and 4 both show that the LOvA loss is always Bayes consistent for any $\alpha \geq 0$. In other words, the class with the largest Bayes score function $g_k^*(\mathbf{x})$ is the same as the class with the largest posterior probability $\eta_k(\mathbf{x})$, i.e.

$$\arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}).$$

However, since only finite samples can be observed in reality, we can only get the estimated score functions $\hat{\mathbf{g}}$ by minimizing the empirical risk $\widehat{\mathcal{R}}_L^\psi[\mathbf{g}]$ using (11). Unfortunately, due to the randomness of samples and the estimation error, for some instances $\mathbf{x} \in \mathcal{X}$, the class with the largest estimated score function $\hat{g}_k(\mathbf{x})$ is likely to be different from the one with the largest Bayes score function $g_k^*(\mathbf{x})$, i.e.

$$\arg \max_{y \in \mathcal{Y}} \hat{g}_y(\mathbf{x}) \neq \arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}),$$

which incurs a larger classification error compared to the Bayes classifier, especially for some indecisive cases where the largest two posterior probabilities are very close. Therefore, to accurately classify such points, we encourage the Bayes score functions to enlarge the difference between its largest two score functions. In this way, in spite of some randomness from samples and the estimation error, the estimated score functions are able to achieve

$$\arg \max_{y \in \mathcal{Y}} \hat{g}_y^*(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x})$$

and therefore have high prediction accuracy.

First, for a fixed \mathbf{x} , we rank the score functions $[g_k(\mathbf{x})]_{k \in \mathcal{Y}}$ of the K classes in order and denote them as $g_{(1)}(\mathbf{x}) \geq g_{(2)}(\mathbf{x}) \geq \dots \geq g_{(K)}(\mathbf{x})$. In the next definition, we define the *margin* as the difference between the largest two score functions as follow.

Definition 3. (Margin) Let $\mathbf{g}(\mathbf{x}) := [g_y(\mathbf{x})]_{y \in \mathcal{Y}}$ be the vector of score functions and let $[g_{(y)}(\mathbf{x})]_{y \in \mathcal{Y}}$ be the ordered vector of score functions correspondingly. Then the difference between $g_{(1)}(\mathbf{x})$ and $g_{(2)}(\mathbf{x})$ denoted as

$$\Delta_{\mathbf{g}}(\mathbf{x}) = g_{(1)}(\mathbf{x}) - g_{(2)}(\mathbf{x})$$

is called the *margin* of the score functions \mathbf{g} .

We aim to prove that the margin for our proposed LOvA loss is larger than the margin for other loss functions. For binary surrogate losses satisfying the strictly proper composite property of Definition 2, we can compare the estimated score functions using its link functions. Therefore, we will investigate the margin of the LOvA loss equipped with different strictly proper composite loss functions.

5.2.1 Logistic loss

We will first investigate the logistic loss, which has a continuous inverse link function [19]. In Example 1, we saw that the logistic loss has inverse link function $\gamma^{-1}[f(\mathbf{x})] = 1/[1 + \exp(-f(\mathbf{x}))]$. We can approximate $g_y(\mathbf{x})$ by this link function combined with its binary probability.

It is important to note that the probabilities $\eta_k(\mathbf{x}) = \mathbb{P}(Y = k|X = \mathbf{x})$ and $p_{m,k}(\mathbf{x}) = \mathbb{P}(M = k|X = \mathbf{x}, Y = k)$ are closely related. When η_k is large (i.e. \mathbf{x} will get label k), we expect $p_{m,k}$ to be large as well and vice versa. In this case, the expert will also predict the label that has high probability. Therefore, it is desirable to have the probabilities of η_k 's and $p_{m,k}$'s have the same order.

To achieve this, we introduce an assumption about the order-preserving property, as follows:

Assumption 2. (Order-preserving property) Let $\eta_k(\mathbf{x})$ and $p_{m,k}(\mathbf{x})$ be defined as in (1) and (3) respectively. Assume that

$$(p_{m,i}(\mathbf{x}) - p_{m,j}(\mathbf{x}))(\eta_i(\mathbf{x}) - \eta_j(\mathbf{x})) > 0, \quad \forall \mathbf{x} \in \mathcal{X}$$

for any $i, j \in \mathcal{Y}$.

When the property of Assumption 2 holds for a classifier, we get that $(p_{m,i}(\mathbf{x}) - p_{m,j}(\mathbf{x}))$ and $(\eta_i(\mathbf{x}) - \eta_j(\mathbf{x}))$ should have an equal sign, thus can be interpreted as

$$p_{m,i}(\mathbf{x}) > p_{m,j}(\mathbf{x}) \Leftrightarrow \eta_i(\mathbf{x}) > \eta_j(\mathbf{x})$$

and conversely. This does not necessarily mean that these probabilities should be directly proportional to each other, but ordering its respective labels based on these probabilities should give equal results. Assumption 2 is stronger than Assumption 1, which only assumes that the class with the largest posterior probability $\eta_k(\mathbf{x})$ is also the one with the largest probability $p_{m,k}(\mathbf{x})$.

Theorem 5. Let $\eta_k(\mathbf{x})$ and $p_{m,k}(\mathbf{x})$ be defined as in (1) and (3) respectively and let Assumption 2 hold. If ψ_L (7) is the LOvA loss equipped with the logistic loss ϕ , then for any \mathbf{x} , the *margin* $\Delta_{\mathbf{g}^*}(\mathbf{x})$ of the Bayes score function $\mathbf{g}^*(\mathbf{x})$ becomes larger as $\alpha \geq 0$ increases.

The complete proof is provided in Section 8.3. Theorem 5 shows that under a reasonable assumption, we are able to obtain an increased margin for the logistic loss, thus a better and more deterministic classifier.

Note that the link function of the exponential loss has a similar form, thus the same proof suffices. Therefore, we also get an increased margin for the exponential loss as α increases.

5.2.2 Square Loss

Next, we turn our attention to the square loss, which also has a continuous inverse link function [19]. In Example 3, we saw that the square loss has inverse link function $\gamma^{-1}[f(\mathbf{x})] = (f(\mathbf{x}) + 1)/2$. We can approximate $g_y(\mathbf{x})$ by this link function combined with its binary probability.

Theorem 6. Let $\eta_k(\mathbf{x})$ and $p_{m,k}(\mathbf{x})$ be defined in (1) and (3) respectively and let Assumption 2 hold. If $\max \eta_k(\mathbf{x}) < 1/2$ for any \mathbf{x} and ψ_L (7) is the LOvA loss equipped with the square loss ϕ , the *margin* $\Delta_{\mathbf{g}^*}(\mathbf{x})$ of the Bayes score function $\mathbf{g}^*(\mathbf{x})$ becomes larger as

$$\alpha \in \left[0, \min_{\mathbf{x}} \min_{k \in \mathcal{Y}} (\eta_k(\mathbf{x})^{-1} - 2) / p_{m,k}(\mathbf{x}) \right]$$

increases for any \mathbf{x} .

The complete proof is provided in Section 8.3.2. Theorem 6 has some limitations. The feasible range for α is complex and difficult to understand, making it less practical for experiments. Moreover, the assumption that $\max \eta_k(\mathbf{x}) < 1/2$ is not true for most datasets, meaning that the theorem does not guarantee a larger margin for simple samples where $\max \eta_k(\mathbf{x}) \geq 1/2$. Consequently, in the next section, we will focus on the logistic loss for experimental purposes.

6 Experiments

In this section, we conduct numerical experiments to evaluate the effectiveness of our proposed LOvA surrogate loss, comparing its performance with the softmax loss from [16] and the OvA loss from [24]. We perform simulations on the CIFAR-10 dataset, with the training and test sets provided according to [12]. We further divide the training set into a 90% training subset and a 10% validation subset. We standardized the dataset to have zero mean and unit standard deviation.

To generate expert predictions from the training labels, we assign 70% probability of providing an accurate label for images belonging to the classes $[1, k]$, but a random label for images belonging to the classes $(k, 10]$. We vary the value of k from 2 to 8 to simulate different scenarios for experts with varying predictive capabilities.

We use the same neural network architecture and training configurations for all methods. In accordance with [16] and [24], we utilize a 28-layer Wide Residual Network [27] to parameterize the $\mathbf{g}(\mathbf{x})$ functions. The optimization process is performed using stochastic gradient descent with a momentum of 0.9, weight decay of $5e - 4$, and an initial learning rate of 0.1 with the cosine annealing learning rate schedule. The models are trained with a batch size of 1024, and we do not employ any data augmentation techniques in line with [16] and [24]. To prevent overfitting, we monitor the validation loss throughout the training process and choose the model in the epoch whose validation loss is the lowest. In the evaluation of OvA and LOvA results, the logistic loss function is used as the integrated binary surrogate loss.

Firstly, we analyze the impact of various leverage parameters α on the proposed LOvA loss function by choosing three distinct values of k to simulate varying levels of expert proficiency. We then modify the value of α and compare the overall accuracy of the L2D system (system accuracy), the accuracy of the classifier on the entire test set (classifier accuracy), and the coverage, which indicates the percentage of samples that the system has not deferred. We compute all three measurements by averaging the results of multiple runs with unique random seeds. We employ a total of six seeds in the analysis.

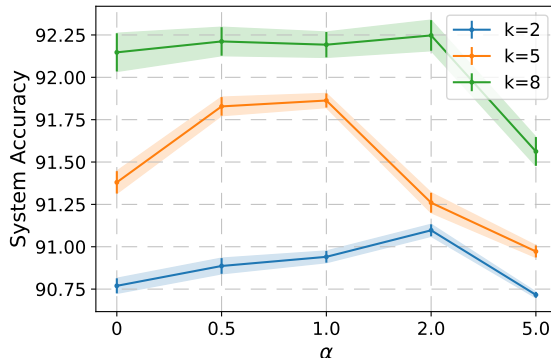


FIGURE 2: System accuracy as the function of the leverage parameter α in the proposed LOvA surrogate loss.

Figures 2 and 3 demonstrate that selecting an appropriate value for the leverage parameter α can significantly improve both the system accuracy and the classifier accuracy across various values of k compared to the scenario where $\alpha = 0$. This suggests that the LOvA

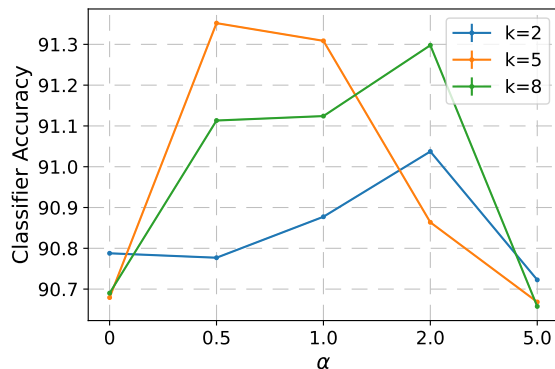


FIGURE 3: Classifier accuracy as the function of the leverage parameter α in the proposed LOvA surrogate loss.

loss ($\alpha \neq 0$) can outperform the OvA loss ($\alpha = 0$) when an optimal α is chosen. Therefore, adjusting the leverage parameter α can maximize the performance of the L2D system. In addition, the empirical superiority of the LOvA loss over the OvA loss on classification accuracy shown in Figure 3 verifies the larger margin of Bayes score functions of the LOvA loss as proven in Theorem 5 and 6. Furthermore, we discovered that the optimal α value for the system accuracy is similar to that for the classifier accuracy.

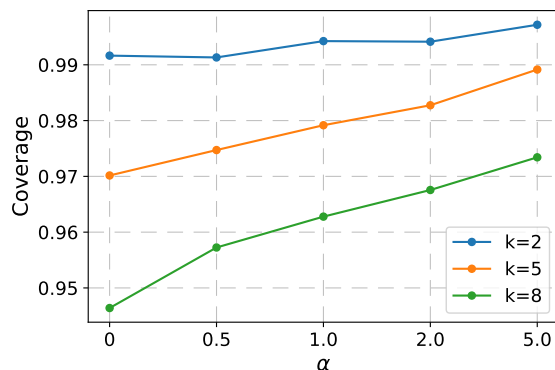


FIGURE 4: Coverage as the function of the leverage parameter α in the proposed LOvA surrogate loss.

Figure 4 shows that, under different k values, a higher α value corresponds to a greater coverage, indicating that the system relies less on expert input. Our LOvA loss has a lower deferring rate compared to the OvA loss (i.e. LOvA loss with $\alpha = 0$), providing empirical evidence supporting the theoretical result presented in Theorem 2 from Section 4.2.

Secondly, we assess the performance of our proposed LOvA loss function, OvA loss function, and softmax loss function on the L2D system’s accuracy and classifier accuracy under various expert capabilities k . We tune the parameter α in the LOvA loss function by choosing the model with the lowest validation loss. The relationship between the number of classes the expert can predict correctly k and the two types of accuracy (system and classifier) is illustrated in Figures 5 and 6. Our LOvA loss function consistently outperforms the other two loss functions, demonstrating the effectiveness of our approach for various levels of expert competence.

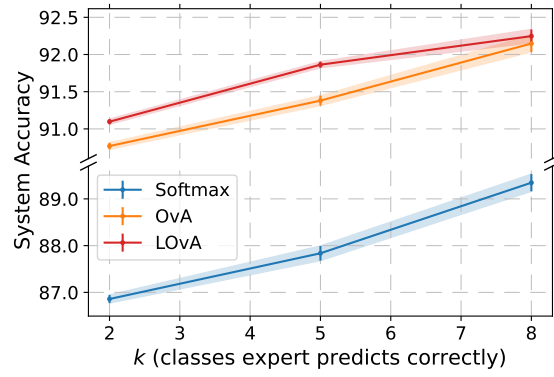


FIGURE 5: System accuracy as the function of an expert with increasing expertise.

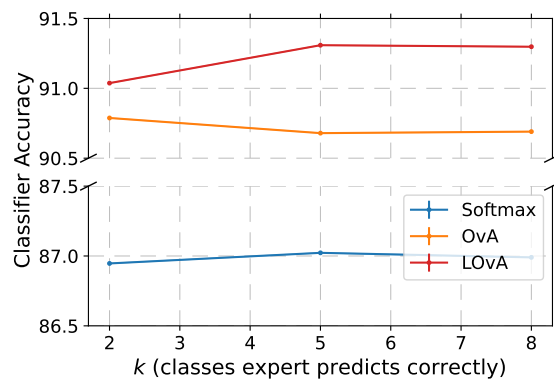


FIGURE 6: Classifier accuracy as the function of an expert with increasing expertise.

7 Conclusion & Discussion

This paper presents a novel approach for improving the performance of multiclass L2D frameworks called the *Leveraged One-vs-All (LOvA)* loss function by decreasing the reliance on expert deferring. We provide theoretical justification for the Bayes risk consistency of our proposed LOvA loss and demonstrate that, under reasonable assumptions, our method can increase the decision margin proportionally to the leverage parameter. Our experimental results confirm the effectiveness of our proposed approach, outperforming other state-of-the-art L2D systems. Overall, this work contributes a valuable loss function with potential applications across a wide range of fields.

One possible area for future work is to extend the multiclass L2D framework to incorporate multiple experts, following the approach of the softmax and OvA losses [23]. In such cases, the expert with the highest confidence score should be utilized if deferral occurs, resulting in an overall increase in deferral confidence. This extension could further improve the classification performance of the L2D system.

8 Proofs

8.1 Proofs Related to Section 4

Proof of Theorem 1. The LOvA loss is defined by

$$\begin{aligned}\psi_L(y, m, \mathbf{g}(\mathbf{x})) &= \phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})] \\ &\quad + \mathbf{1}[m = y](\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})] + \alpha\phi[g_y(\mathbf{x})]).\end{aligned}$$

Since the Bayes score function $\mathbf{g}_L^* = [g_y^*]_{y \in \mathcal{Y}^\perp}$ is the minimizer of the risk $\mathcal{R}^\Psi(\mathbf{g})$, i.e.

$$\begin{aligned}\mathbf{g}_L^* &= \arg \min_{\mathbf{g}} \mathcal{R}^\Psi(\mathbf{g}) = \arg \min_{\mathbf{g}} \mathbb{E}_{\mathbf{x}, y, m} \psi_L(y, m, \mathbf{g}(\mathbf{x})) \\ &= \arg \min_{\mathbf{g}} \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y, m | \mathbf{x}} \psi_L(y, m, \mathbf{g}(\mathbf{x})),\end{aligned}$$

we have

$$\mathbf{g}_L^*(\mathbf{x}) = \arg \min_{\mathbf{g}(\mathbf{x})} \mathbb{E}_{y, m | \mathbf{x}} \psi_L(y, m, \mathbf{g}(\mathbf{x})) = \arg \min_{\mathbf{g}(\mathbf{x})} \mathcal{C}_{\mathbf{x}}^{\Psi_L}.$$

Simplifying the inner risk by expanding the expectations gives

$$\begin{aligned}\mathcal{C}_{\mathbf{x}}^{\Psi_L}(\mathbf{g}) &:= \mathbb{E}_{y, m | \mathbf{x}} \psi_L(y, m, \mathbf{g}(\mathbf{x})) = \mathbb{E}_{y | \mathbf{x}} \mathbb{E}_{m | \mathbf{x}, y} \psi_L(y, m, \mathbf{g}(\mathbf{x})) \\ &= \mathbb{E}_{y | \mathbf{x}} \left[\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})] \right. \\ &\quad \left. + \sum_{m \in \mathcal{Y}} \mathbb{P}(M = m | X = \mathbf{x}, Y = y) \mathbf{1}[m = y](\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})] + \alpha\phi[g_y(\mathbf{x})]) \right] \\ &= \mathbb{E}_{y | \mathbf{x}} \left[\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] + \phi[-g_\perp(\mathbf{x})] \right. \\ &\quad \left. + \mathbb{P}(M = y | X = \mathbf{x}, Y = y)(\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})] + \alpha\phi[g_y(\mathbf{x})]) \right].\end{aligned}$$

Expanding the outer expectation and using $\eta_y(\mathbf{x}) = \mathbb{P}(Y = y | X = \mathbf{x})$, we get

$$\begin{aligned}\mathcal{C}_{\mathbf{x}}^{\Psi_L}(\mathbf{g}) &= \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \left[\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] \right] + \phi[-g_\perp(\mathbf{x})] \\ &\quad + \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \mathbb{P}(M = y | X = \mathbf{x}, Y = y)(\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})] + \alpha\phi[g_y(\mathbf{x})]).\end{aligned}$$

By the law of total probability, we have

$$\sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \mathbb{P}(M = y | X = \mathbf{x}, Y = y) = \mathbb{P}(M = Y | X = \mathbf{x}).$$

Therefore, we get

$$\begin{aligned}\mathcal{C}_{\mathbf{x}}^{\Psi_L}(\mathbf{g}) &= \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \left[\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] \right] + \phi[-g_\perp(\mathbf{x})] \\ &\quad + \mathbb{P}(M = Y | X = \mathbf{x})(\phi[g_\perp(\mathbf{x})] - \phi[-g_\perp(\mathbf{x})])\end{aligned}$$

$$\begin{aligned}
& + \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \mathbb{P}(M = y | X = \mathbf{x}, Y = y) \alpha \phi[g_y(\mathbf{x})] \\
= & \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \left[\phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] \right] + \mathbb{P}(M \neq Y | X = \mathbf{x}) \phi[-g_{\perp}(\mathbf{x})] \\
& + \mathbb{P}(M = Y | X = \mathbf{x}) \phi[g_{\perp}(\mathbf{x})] + \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \mathbb{P}(M = y | X = \mathbf{x}, Y = y) \alpha \phi[g_y(\mathbf{x})] \\
= & \sum_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \left[(1 + \alpha \mathbb{P}(M = y | X = \mathbf{x}, Y = y)) \phi[g_y(\mathbf{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\mathbf{x})] \right] \\
& + \mathbb{P}(M = Y | X = \mathbf{x}) \phi[g_{\perp}(\mathbf{x})] + \mathbb{P}(M \neq Y | X = \mathbf{x}) \phi[-g_{\perp}(\mathbf{x})].
\end{aligned}$$

Using the notations $p_m(\mathbf{x}) := \mathbb{P}(M = Y | X = \mathbf{x})$ and $p_{m,k}(\mathbf{x}) := \mathbb{P}(M = k | X = \mathbf{x}, Y = k)$, $k \in [K]$, we get

$$\begin{aligned}
\mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g}) &= \sum_{y \in \mathcal{Y}} \left[\eta_y(\mathbf{x}) (1 + \alpha p_{m,y}(\mathbf{x})) \phi[g_y(\mathbf{x})] + (1 - \eta_y(\mathbf{x})) \phi[-g_y(\mathbf{x})] \right] \\
& + p_m(\mathbf{x}) \phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x})) \phi[-g_{\perp}(\mathbf{x})].
\end{aligned} \tag{12}$$

For the logistic loss $\phi(v) := \log(1 + \exp(-v))$, it is obvious that $\mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})$ is convex w.r.t. all $g_y(\mathbf{x})$, where $y \in \mathcal{Y}^{\perp}$. Therefore, in order to obtain the minimizer $\mathbf{g}^*(\mathbf{x})$, it suffices to differentiate $\mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})$ w.r.t. each g_y and set it equal to zero. For any $y \in \mathcal{Y}$, the partial derivative of $\mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})$ with respect to $g_y(\mathbf{x})$ is

$$\begin{aligned}
\frac{\partial \mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})}{\partial g_y(\mathbf{x})} &= -\frac{\eta_y(\mathbf{x}) (1 + \alpha p_{m,y}(\mathbf{x}))}{1 + \exp(g_y(\mathbf{x}))} + \frac{(1 - \eta_y(\mathbf{x})) \exp(g_y(\mathbf{x}))}{1 + \exp(g_y(\mathbf{x}))} \\
&= \frac{(1 - \eta_y(\mathbf{x})) \exp(g_y(\mathbf{x})) - \eta_y(\mathbf{x}) (1 + \alpha p_{m,y}(\mathbf{x}))}{1 + \exp(g_y(\mathbf{x}))}.
\end{aligned}$$

Letting the above equation be zero, we get the solution as

$$g_y^*(\mathbf{x}) = \log\left(\frac{\eta_y(\mathbf{x}) (1 + \alpha p_{m,y}(\mathbf{x}))}{1 - \eta_y(\mathbf{x})}\right).$$

Similarly, the partial derivative of $\mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})$ w.r.t. $g_{\perp}(\mathbf{x})$ is

$$\begin{aligned}
\frac{\partial \mathcal{C}_{\mathbf{x}}^{\psi^L}(\mathbf{g})}{\partial g_{\perp}(\mathbf{x})} &= -\frac{p_m(\mathbf{x})}{1 + \exp(g_{\perp}(\mathbf{x}))} + \frac{(1 - p_m(\mathbf{x})) \exp(g_{\perp}(\mathbf{x}))}{1 + \exp(g_{\perp}(\mathbf{x}))} \\
&= \frac{(1 - p_m(\mathbf{x})) \exp(g_{\perp}(\mathbf{x})) - p_m(\mathbf{x})}{1 + \exp(g_{\perp}(\mathbf{x}))}.
\end{aligned}$$

Setting this to be zero, we get the optimal score function for the deferring as

$$g_{\perp}^*(\mathbf{x}) = \log\left(\frac{p_m(\mathbf{x})}{1 - p_m(\mathbf{x})}\right).$$

Therefore, we finish the proof. \square

Proof of Theorem 2. Using the decision rules in (9) and Theorem 1, we get the optimal score functions as

$$g_y^*(\mathbf{x}) = \log\left(\frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x}))\right), \quad \forall y \in \mathcal{Y},$$

$$g_{\perp}^*(\mathbf{x}) = \log\left(\frac{p_m(\mathbf{x})}{1 - p_m(\mathbf{x})}\right).$$

The optimal classifier for the LOvA loss is given by

$$\begin{aligned} f_L^*(\mathbf{x}) &= \arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}) \\ &= \arg \max_{y \in \mathcal{Y}} \log\left(\frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x}))\right) \\ &= \arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})). \end{aligned} \quad (13)$$

First, we prove that

$$\arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})) \subset \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$$

by contradiction. Assume that the assertion does not hold. Then there exists some $k \in \mathcal{Y}$ satisfying

$$k \in \arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})) \quad \text{and} \quad k \notin \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}).$$

Moreover, there exists another index $i \neq k$ satisfying $\eta_i(\mathbf{x}) > \eta_k(\mathbf{x})$. Since

$$\arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \subset \arg \max_{k \in \mathcal{Y}} p_{m,k}(\mathbf{x}),$$

we have $p_{m,i}(\mathbf{x}) > p_{m,k}(\mathbf{x})$. Therefore, we have

$$\frac{\eta_i(\mathbf{x})}{1 - \eta_i(\mathbf{x})}(1 + \alpha p_{m,i}(\mathbf{x})) > \frac{\eta_k(\mathbf{x})}{1 - \eta_k(\mathbf{x})}(1 + \alpha p_{m,k}(\mathbf{x})).$$

This implies

$$k \notin \arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})),$$

which yields the contradiction.

Next, we prove that

$$\arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \subset \arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})).$$

For any $k \in \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$ and any $i \neq k$, we have $\eta_k(\mathbf{x}) \geq \eta_i(\mathbf{x})$. Due to Assumption 1, we have $p_{m,k}(\mathbf{x}) \geq p_{m,i}(\mathbf{x})$. Thus, we have

$$\frac{\eta_k(\mathbf{x})}{1 - \eta_k(\mathbf{x})}(1 + \alpha p_{m,k}(\mathbf{x})) \geq \frac{\eta_i(\mathbf{x})}{1 - \eta_i(\mathbf{x})}(1 + \alpha p_{m,i}(\mathbf{x}))$$

for any $i \neq k$ and therefore,

$$k \in \arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})}(1 + \alpha p_{m,y}(\mathbf{x})).$$

Combining these two parts of the proof, we get

$$\arg \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x})) = \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}).$$

This together with (13) yields

$$f_L^*(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}).$$

Moreover, the optimal rejector for the LOvA loss satisfies

$$\begin{aligned} r_L^*(\mathbf{x}) &= \mathbb{1} \left[g_{\perp}^*(\mathbf{x}) \geq \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}) \right] \\ &= \mathbb{1} \left[\log \left(\frac{p_m(\mathbf{x})}{1 - p_m(\mathbf{x})} \right) \geq \max_{y \in \mathcal{Y}} \log \left(\frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x})) \right) \right] \\ &= \mathbb{1} \left[\frac{p_m(\mathbf{x})}{1 - p_m(\mathbf{x})} \geq \max_{y \in \mathcal{Y}} \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x})) \right] \\ &= \mathbb{1} \left[p_m(\mathbf{x}) \geq \max_{y \in \mathcal{Y}} \frac{\frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x}))}{1 + \frac{\eta_y(\mathbf{x})}{1 - \eta_y(\mathbf{x})} (1 + \alpha p_{m,y}(\mathbf{x}))} \right], \end{aligned}$$

which yields the assertion. \square

8.2 Proofs Related to Section 5.1

8.2.1 Bayes Consistency and Classification Calibration

The definition of Bayes consistency stated in Definition 1 is not concrete enough to be used in checking the consistency of a surrogate loss function. In this section, we aim to find a necessary and sufficient condition of Bayes consistency called *classification calibration*, which is easier to be checked. Before introducing the concept of classification calibration, we first provide some notations concerning it. To this aim, we first write the ψ -risk as

$$\mathcal{R}^\psi[\mathbf{g}] = \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{y,m|\mathbf{x}} [\psi(y, m, \mathbf{g}(\mathbf{x}))] \right].$$

Finding the Bayes score function of ψ -risk $\mathcal{R}^\psi[\mathbf{g}]$ is equivalent to finding the minimizer of the inner conditional expectation $\mathbb{E}_{y,m|\mathbf{x}} [\psi(y, m, \mathbf{g}(\mathbf{x}))]$ for each $\mathbf{x} \in \mathcal{X}$. We denote the inner ψ -risk as

$$\mathcal{C}_{\mathbf{x}}^\psi[\mathbf{g}] := \mathbb{E}_{y,m|\mathbf{x}} [\psi(y, m, \mathbf{g}(\mathbf{x}))]$$

and the Bayes inner ψ -risk as

$$\mathcal{C}_{\mathbf{x}}^{\psi,*} := \inf_{f: \mathcal{X} \rightarrow \mathcal{Y}} \mathcal{C}_{\mathbf{x}}^\psi[\mathbf{g}].$$

Similarly for the 0 – 1 loss ℓ_{0-1} on the classifier

$$f(\mathbf{x}) := \arg \max_{k \in \mathcal{Y}} g_k(\mathbf{x}),$$

we define the inner ℓ_{0-1} -risk by

$$\mathcal{C}_{\mathbf{x}}^{\ell_{0-1}}[f] := \mathbb{E}_{y|\mathbf{x}} \ell_{0-1}(y, f(\mathbf{x}))$$

and Bayes inner ℓ_{0-1} -risk by

$$\mathcal{C}_{\mathbf{x}}^{\ell_{0-1},*} := \inf_{f:\mathcal{X}\rightarrow\mathcal{Y}} \mathcal{C}_{\mathbf{x}}^{\ell_{0-1}}[f].$$

Both inner risks make use of the conditional expectations of the output space \mathcal{Y} given the input \mathbf{x} . Therefore, we will look into their respective posterior probability functions denoted as $\eta_k(\mathbf{x}) = \mathbb{P}(Y = k|X = \mathbf{x})$, i.e. the probability that a given input \mathbf{x} is labelled as class k . Intuition suggests that the class k with the largest corresponding $\eta_k(\mathbf{x})$ should be chosen.

We can now give the formal definition of the *classification calibration* property.

Definition 4. (Classification calibration) Let $\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^{K+1}$ be the score function vector in L2D problem and $f(\mathbf{x}) := \arg \max_{k \in \mathcal{Y}} g_k(\mathbf{x})$ be the induced classifier. Then we say the surrogate loss ψ on the set $\Omega \subset \mathbb{R}^{K+1}$ is *classification calibrated* if there exists a predictor function such that

$$\inf_{\mathbf{g}(\mathbf{x}) \in \Omega: \eta_{f(\mathbf{x})}(\mathbf{x}) < \max_y \eta_y(\mathbf{x})} \mathcal{C}_{\mathbf{x}}^{\psi}[\mathbf{g}] > \inf_{\mathbf{g}(\mathbf{x}) \in \Omega} \mathcal{C}_{\mathbf{x}}^{\psi}[\mathbf{g}], \quad \forall \boldsymbol{\eta} \in \mathcal{S}_K,$$

where $\mathcal{S}_K \subset \mathbb{R}^K$ is the probability simplex.

Definition 4 states that classification calibration involves minimizing the inner ψ -risk, which yields an optimal $\mathbf{g}^*(\mathbf{x}) \in \Omega$. This optimal score function ensures that the label with the highest score value matches the one with the largest posterior probability, i.e.

$$\arg \max_{k \in \mathcal{Y}} g_k^*(\mathbf{x}) = \arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}).$$

Note that this must hold for any probability vector $\boldsymbol{\eta}$ and any instance \mathbf{x} .

The next theorem shows us that the *classification calibration* property is indeed the necessary and sufficient condition to investigate, since the convergence of the inner risk will imply the convergence of the whole risk, and thus Bayes consistency will be satisfied.

Theorem 7. Let $\psi : \mathcal{Y} \times \mathcal{M} \times \mathbb{R}^{K+1} \rightarrow \mathbb{R}_+$ be a loss function and \mathcal{G} be the set of score function vectors satisfying $\mathcal{G} := \{\mathbf{g} : \forall \mathbf{x}, \mathbf{g}(\mathbf{x}) \in \Omega\}$. Moreover, let $\{\mathcal{G}_n\}$ be a sequence of function classes satisfying $\mathcal{G}_n \subseteq \mathcal{G}$ and $\bigcup_n \mathcal{G}_n = \mathcal{G}$. Then ψ is classification calibrated on Ω if and only if for any sequence of $\mathbf{g}_n \in \mathcal{G}_n$, there holds

$$\mathcal{R}^{\psi}[\mathbf{g}_n] \rightarrow \mathcal{R}^{\psi,*} \implies \mathcal{R}^{\ell}[f_n] \rightarrow \mathcal{R}^{\ell,*},$$

where $f_n := \arg \max_{k \in \mathcal{Y}} g_{n,k}$.

Proof. See Appendix A of [21]. □

Since Bayes consistency, as introduced in Definition 1, is the immediate consequence in Theorem 7, we get that classification calibration is indeed the necessary and sufficient condition to study. Therefore, proving if the surrogate loss ψ is Bayes consistent is equivalent to checking the property of classification calibration for the loss ψ .

8.2.2 Proof of Theorem 3

Proof of Theorem 3. By (12), we have the following form of the inner risk

$$\mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}] := \mathbb{E}_{y,m|\mathbf{x}}[\psi_L(y, m, \mathbf{g}(\mathbf{x}))],$$

i.e.

$$\begin{aligned} \mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}] &= \sum_{y \in \mathcal{Y}} \left[\eta_y(\mathbf{x})(1 + \alpha p_{m,y}(\mathbf{x}))\phi[g_y(\mathbf{x})] + (1 - \eta_y(\mathbf{x}))\phi[-g_y(\mathbf{x})] \right] \\ &\quad + p_m(\mathbf{x})\phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x}))\phi[-g_{\perp}(\mathbf{x})] \\ &= \sum_{y \in \mathcal{Y}} (1 + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})) \left(\frac{(\eta_y(\mathbf{x}) + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x}))\phi[g_y(\mathbf{x})]}{1 + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})} + \frac{(1 - \eta_y(\mathbf{x}))\phi[-g_y(\mathbf{x})]}{1 + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})} \right) \\ &\quad + p_m(\mathbf{x})\phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x}))\phi[-g_{\perp}(\mathbf{x})] \\ &= \sum_{y \in \mathcal{Y}} (1 + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})) (Q_y(\mathbf{x})\phi[g_y(\mathbf{x})] + (1 - Q_y(\mathbf{x}))\phi[-g_y(\mathbf{x})]) \\ &\quad + p_m(\mathbf{x})\phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x}))\phi[-g_{\perp}(\mathbf{x})], \end{aligned}$$

where we denote

$$Q_y(\mathbf{x}) := \frac{\eta_y(\mathbf{x}) + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})}{1 + \alpha \eta_y(\mathbf{x})p_{m,y}(\mathbf{x})}, \quad y \in \mathcal{Y}.$$

Since the binary loss ϕ is a strictly proper composite loss, there exists an increasing function γ such that

$$\gamma(p(\mathbf{x})) := \arg \min_{g_y(\mathbf{x}) \in \mathbb{R}} p(\mathbf{x})\phi[g_y(\mathbf{x})] + (1 - p(\mathbf{x}))\phi[-g_y(\mathbf{x})] \quad \forall y \in \mathcal{Y}^{\perp}.$$

Therefore, the unique minimizer $\mathbf{g}^*(\mathbf{x})$ of the inner risk $\mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}]$ is given by

$$\begin{aligned} g_k^*(\mathbf{x}) &= \gamma(Q_k(\mathbf{x})), \quad \forall k \in \mathcal{Y}, \\ g_{\perp}^*(\mathbf{x}) &= \gamma(p_m(\mathbf{x})). \end{aligned}$$

Since the link function γ is strictly increasing, we have

$$\arg \max_{y \in \mathcal{Y}} Q_y(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}).$$

For any $k \in \arg \max_{k \in \mathcal{Y}} g_k^*(\mathbf{x})$, we have $Q_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} Q_y(\mathbf{x})$. Now, we prove

$$\eta_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$$

by contradiction. Assume that $\eta_k(\mathbf{x}) \neq \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$. Then there exists an index $i \neq k$ such that $\eta_k(\mathbf{x}) < \eta_i(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$. Since

$$\arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \subset \arg \max_{k \in \mathcal{Y}} p_{m,k}(\mathbf{x}),$$

by Assumption 1 we get $i \in \arg \max_{k \in \mathcal{Y}} p_{m,k}(\mathbf{x})$ and thus $p_{m,i}(\mathbf{x}) \geq p_{m,k}(\mathbf{x})$. Since the function $f(t) := t/(1 + ct)$ with $t > 0$ and $c \geq 0$ is strictly increasing w.r.t. t , we get

$$f(t_1) > f(t_2) \quad \text{if } t_1 > t_2. \quad (14)$$

Taking $t_1 := \eta_i(\mathbf{x})$, $t_2 := \eta_k(\mathbf{x})$ and $c := p_{m,i}(\mathbf{x})$, we get

$$\frac{\eta_i(\mathbf{x})}{1 + \alpha\eta_i(\mathbf{x})p_{m,i}(\mathbf{x})} > \frac{\eta_k(\mathbf{x})}{1 + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})}.$$

Multiplying both sides of the above equation by $1 + \alpha p_{m,i}(\mathbf{x})$, we obtain

$$Q_i(\mathbf{x}) = \frac{\eta_i(\mathbf{x}) + \alpha\eta_i(\mathbf{x})p_{m,i}(\mathbf{x})}{1 + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})} > \frac{\eta_k(\mathbf{x}) + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})}{1 + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})}. \quad (15)$$

Applying (14) with $t_1 := p_{m,i}(\mathbf{x})$, $t_2 := p_{m,k}(\mathbf{x})$, $c := \alpha\eta_k(\mathbf{x})$ with $\alpha \geq 0$, we get

$$\frac{p_{m,i}(\mathbf{x})}{1 + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})} \geq \frac{p_{m,k}(\mathbf{x})}{1 + \alpha\eta_k(\mathbf{x})p_{m,k}(\mathbf{x})}.$$

Multiplying $\alpha(1 - \eta_k(\mathbf{x}))$, adding 1 and then multiplying $\eta_k(\mathbf{x})$ to both hand sides of the above equation yields

$$\frac{\eta_k(\mathbf{x})(1 + \alpha p_{m,i}(\mathbf{x}))}{1 + \alpha\eta_k(\mathbf{x})p_{m,i}(\mathbf{x})} \geq \frac{\eta_k(\mathbf{x})(1 + \alpha p_{m,k}(\mathbf{x}))}{1 + \alpha\eta_k(\mathbf{x})p_{m,k}(\mathbf{x})} = Q_k(\mathbf{x}). \quad (16)$$

Combining (15) and (16), we get $Q_i(\mathbf{x}) > Q_k(\mathbf{x})$ and thus we derive the contradiction to $Q_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} Q_y(\mathbf{x})$. Therefore, we finish the proof of $\eta_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})$. Thus, we prove that

$$\arg \max_{y \in \mathcal{Y}} g_y^*(\mathbf{x}) \subset \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}).$$

As a result, we have

$$\mathbf{g}^*(\mathbf{x}) \notin \left\{ \mathbf{g}(\mathbf{x}) : \eta_{\arg \max_{k \in \mathcal{Y}} g_k(\mathbf{x})}(\mathbf{x}) < \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \right\},$$

which implies

$$\inf_{\mathbf{g}(\mathbf{x}) \in \mathbb{R}^{K+1} : \eta_{f(\mathbf{x})}(\mathbf{x}) < \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x})} \mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}] > \mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}^*] = \inf_{\mathbf{g}(\mathbf{x}) \in \mathbb{R}^{K+1}} \mathcal{C}_{\mathbf{x}}^{\psi_L}[\mathbf{g}], \quad \forall \boldsymbol{\eta} \in \mathcal{S}_K,$$

where $f(\mathbf{x}) := \arg \max_{y \in \mathcal{Y}} g_y(\mathbf{x})$. Therefore, we prove that the surrogate loss ψ_L is classification calibrated. By the equivalence between classification calibration and Bayes consistency in Theorem 7, we get the assertion. \square

8.2.3 Proof of Theorem 4

Proof of Theorem 4. Using (12), we get

$$\begin{aligned} \mathcal{C}_x^{\psi_L}[\mathbf{g}] &= \sum_{y \in \mathcal{Y}} (1 + \alpha\eta_y(\mathbf{x})p_{m,y}(\mathbf{x})) (Q_y(\mathbf{x})\phi[g_y(\mathbf{x})] + (1 - Q_y(\mathbf{x}))\phi[-g_y(\mathbf{x})]) \\ &\quad + p_m(\mathbf{x})\phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x}))\phi[-g_{\perp}(\mathbf{x})], \end{aligned}$$

where

$$Q_y(\mathbf{x}) := \frac{\eta_y(\mathbf{x}) + \alpha\eta_y(\mathbf{x})p_{m,y}(\mathbf{x})}{1 + \alpha\eta_y(\mathbf{x})p_{m,y}(\mathbf{x})}, \quad y \in \mathcal{Y}. \quad (17)$$

Since the binary surrogate loss ϕ is a convex, differentiable, and decreasing loss for binary classification, the minimizer \mathbf{g}^* satisfies the first-order condition, i.e.

$$\begin{aligned} Q_k(\mathbf{x})\phi'[g_k^*(\mathbf{x})] - (1 - Q_k(\mathbf{x}))\phi'[-g_k^*(\mathbf{x})] &= 0, & k \in \mathcal{Y}, \\ p_m(\mathbf{x})\phi'[g_\perp^*(\mathbf{x})] - (1 - Q_i)\phi'[-g_\perp^*(\mathbf{x})] &= 0. \end{aligned}$$

Note that the assumptions imply that $\phi'(0) < 0$ and ϕ is a decreasing function. It suffices to consider the case that $Q_i \neq 0.5$, else the first condition cannot be satisfied since $\phi'(0) \neq 0$. By the assumption that ϕ is a decreasing function and $\phi'(0) < 0$, we get that $\phi[g_i^*(\mathbf{x})] < \phi[-g_i^*(\mathbf{x})]$ if $g_i^*(\mathbf{x}) > 0$. Therefore if $Q_i(\mathbf{x}) > 0.5$, we get that $g_i^* > 0$ and else if $Q_i(\mathbf{x}) < 0.5$, there holds $g_i^*(\mathbf{x}) < 0$. Next, we will show that $Q_i < Q_j$ implies $g_i < g_j$ with the following cases.

Case 1: When $Q_i < Q_j < 0.5$, we know that $g_i < 0$ and $g_j < 0$. Since we prove it by contradiction, we assume $g_j < g_i < 0$. Then by the convexity of ϕ , we get that ϕ' is non-decreasing and thus $\phi'[g_j] \leq \phi'[g_i] < 0$. Then we can use the first-order optimality condition to get

$$\phi'[-g_i] = \frac{Q_i\phi'[g_i]}{1 - Q_i} > \frac{Q_j\phi'[g_i]}{1 - Q_j} \geq \frac{Q_j\phi'[g_j]}{1 - Q_j} = \phi'[-g_j].$$

Again, using the fact that ϕ' is non-decreasing for convex ϕ , we get $-g_i > -g_j$ and thus $g_i < g_j$. This contradicts our assumptions, so we have $g_i < g_j$.

Case 2: When $Q_i < 0.5 < Q_j$, we immediately know that $g_i < 0$ and $g_j > 0$ and hence $g_i < g_j$ is satisfied.

Case 3: When $0.5 < Q_i < Q_j$, we know that $g_i > 0$ and $g_j > 0$. Since we prove it by contradiction, we assume $0 < g_j < g_i$. Then by the convexity of ϕ , we get that ϕ' is non-decreasing and thus $\phi'[g_j] \leq \phi'[g_i] < 0$. Then we can use the first-order optimality condition to get

$$\phi'[-g_i] = \frac{Q_i\phi'[g_i]}{1 - Q_i} > \frac{Q_j\phi'[g_i]}{1 - Q_j} \geq \frac{Q_j\phi'[g_j]}{1 - Q_j} = \phi'[-g_j].$$

Again, using the fact that ϕ' is non-decreasing for convex ϕ , we $-g_i > -g_j$ and thus $g_i < g_j$. This contradicts our assumptions, so we must have that $g_i < g_j$.

Notice that

$$\arg \max_{k \in \mathcal{Y}} \eta_k(\mathbf{x}) \subset \arg \max_{k \in \mathcal{Y}} p_{m,k}(\mathbf{x})$$

by Assumption 1. Therefore, if there exist two different indices $i \neq k$ satisfying

$$\eta_i(\mathbf{x}) < \eta_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}),$$

then we have

$$p_{m,i}(\mathbf{x}) < p_{m,k}(\mathbf{x}) = \max_{y \in \mathcal{Y}} p_{m,y}(\mathbf{x}).$$

By the definition of Q_k in (17), we then have

$$Q_i(\mathbf{x}) < Q_k(\mathbf{x}) = \max_y Q_y(\mathbf{x}).$$

Since $Q_i < Q_j \implies g_i < g_j$, we obtain that

$$\eta_i(\mathbf{x}) < \eta_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \implies g_i(\mathbf{x}) < g_k(\mathbf{x}) = \max_{y \in \mathcal{Y}} g_y(\mathbf{x}).$$

Therefore, we finish the proof. \square

8.3 Proofs Related to Section 5.2

8.3.1 Proof of Theorem 5

Proof of Theorem 5. Using (12), we get

$$\begin{aligned} \mathcal{C}_{\mathbf{x}}^{\Psi_L}[\mathbf{g}] &= \sum_{y \in \mathcal{Y}} (1 + \alpha \eta_y(\mathbf{x}) p_{m,y}(\mathbf{x})) (Q_y(\mathbf{x}) \phi[g_y(\mathbf{x})] + (1 - Q_y(\mathbf{x})) \phi[-g_y(\mathbf{x})]) \\ &\quad + p_m(\mathbf{x}) \phi[g_{\perp}(\mathbf{x})] + (1 - p_m(\mathbf{x})) \phi[-g_{\perp}(\mathbf{x})], \end{aligned}$$

where

$$Q_y(\mathbf{x}) := \frac{\eta_y(\mathbf{x}) + \alpha \eta_y(\mathbf{x}) p_{m,y}(\mathbf{x})}{1 + \alpha \eta_y(\mathbf{x}) p_{m,y}(\mathbf{x})}, \quad y \in \mathcal{Y}.$$

Denote $Q_{(1)}(\mathbf{x})$ and $Q_{(2)}(\mathbf{x})$ as the largest two values among $[Q_y(\mathbf{x})]_{y \in \mathcal{Y}}$. Without loss of generality, we let $\eta_i(\mathbf{x}) = \eta_{(1)}(\mathbf{x})$ and $\eta_j(\mathbf{x}) = \eta_{(2)}(\mathbf{x})$. By Assumption 2, we have $p_{m,i}(\mathbf{x}) > p_{m,j}(\mathbf{x})$ and thus $Q_{(1)}(\mathbf{x}) = Q_i(\mathbf{x})$ and $Q_{(2)}(\mathbf{x}) = Q_j(\mathbf{x})$. Since the logistic loss has the link function $\gamma(u) = -\ln\left(\frac{1}{u} - 1\right)$ [15], the largest two Bayes score functions are $g_{(1)}^*(\mathbf{x}) = \gamma(Q_{(1)}(\mathbf{x}))$ and $g_{(2)}^*(\mathbf{x}) := \gamma(Q_{(2)}(\mathbf{x}))$. Therefore, it follows that the margin of the Bayes functions for the LOvA loss is

$$\begin{aligned} \Delta_{\mathbf{g}_L^*}(\mathbf{x}) &= \gamma(Q_i) - \gamma(Q_j) = -\ln\left(\frac{1}{Q_i} - 1\right) + \ln\left(\frac{1}{Q_j} - 1\right) \\ &= -\ln\left(\frac{1 + \alpha \eta_i p_{m,i}}{\eta_i(1 + \alpha p_{m,i})} - 1\right) + \ln\left(\frac{1 + \alpha \eta_j p_{m,j}}{\eta_j(1 + \alpha p_{m,j})} - 1\right) \\ &= \ln\left(\frac{1 + \alpha \eta_j p_{m,j} - \eta_j(1 + \alpha p_{m,j})}{\eta_j(1 + \alpha p_{m,j})}\right) - \ln\left(\frac{1 + \alpha \eta_i p_{m,i} - \eta_i(1 + \alpha p_{m,i})}{\eta_i(1 + \alpha p_{m,i})}\right) \\ &= \ln\left(\frac{1 + \alpha \eta_j p_{m,j} - \eta_j(1 + \alpha p_{m,j})}{\eta_j(1 + \alpha p_{m,j})} \cdot \frac{\eta_i(1 + \alpha p_{m,i})}{1 + \alpha \eta_i p_{m,i} - \eta_i(1 + \alpha p_{m,i})}\right) \\ &= \ln\left(\frac{1 - \eta_j}{\eta_j(1 + \alpha p_{m,j})} \cdot \frac{\eta_i(1 + \alpha p_{m,i})}{(1 - \eta_i)}\right) \\ &= \ln\left(\frac{1 - \eta_j}{\eta_j} \cdot \frac{\eta_i}{1 - \eta_i} \cdot \frac{1 + \alpha p_{m,i}}{1 + \alpha p_{m,j}}\right). \end{aligned}$$

Due to $\eta_i(\mathbf{x}) > \eta_j(\mathbf{x})$ and Assumption 2, it is easy to see that $p_{m,i}(\mathbf{x}) > p_{m,j}(\mathbf{x})$. Therefore as the parameter $\alpha \geq 0$ increases, $\Delta_{\mathbf{g}_L^*}(\mathbf{x})$ becomes larger. \square

8.3.2 Proof of Theorem 6

Proof of Theorem 6. Since our LOvA loss is equipped with the square loss ϕ , the largest two Bayes score functions are $g_{(1)}^*(\mathbf{x}) = \gamma(Q_{(1)}(\mathbf{x}))$ and $g_{(2)}^*(\mathbf{x}) = \gamma(Q_{(2)}(\mathbf{x}))$ with $\gamma(\eta_k) := 2\eta_k - 1$. Without loss of generality, we let $\eta_i(\mathbf{x}) = \eta_{(1)}(\mathbf{x})$ and $\eta_j(\mathbf{x}) = \eta_{(2)}(\mathbf{x})$, where $i, j \in \mathcal{Y}$. By Assumption 2, we have $p_{m,i}(\mathbf{x}) > p_{m,j}(\mathbf{x})$ and thus $Q_{(1)}(\mathbf{x}) = Q_i(\mathbf{x})$ and $Q_{(2)}(\mathbf{x}) = Q_j(\mathbf{x})$. Similar to the proof of Theorem 5, the margin of the Bayes score functions turns out to be

$$\Delta_{\mathbf{g}_L^*}(\mathbf{x}) = \gamma(Q_i) - \gamma(Q_j) = 2(Q_i - Q_j) = \frac{\eta_i(1 + \alpha p_{m,i})}{1 + \alpha \eta_i p_{m,i}} - \frac{\eta_j(1 + \alpha p_{m,j})}{1 + \alpha \eta_j p_{m,j}}.$$

Taking the derivative of $\Delta_{\mathbf{g}_L^*}(\mathbf{x})$ w.r.t. α , we get

$$\frac{d\Delta_{\mathbf{g}_L^*}(\mathbf{x})}{d\alpha} = \eta_i \cdot \frac{p_{m,i}(1 + \alpha \eta_i p_{m,i}) - (1 + \alpha p_{m,i}) \eta_i p_{m,i}}{(1 + \alpha \eta_i p_{m,i})^2}$$

$$\begin{aligned}
& -\eta_j \cdot \frac{p_{m,j}(1 + \alpha\eta_j p_{m,j}) - (1 + \alpha p_{m,j})\eta_j p_{m,j}}{(1 + \alpha\eta_j p_{m,j})^2} \\
&= \frac{\eta_i(1 - \eta_i)p_{m,i}}{(1 + \alpha\eta_i p_{m,i})^2} - \frac{\eta_j(1 - \eta_j)p_{m,j}}{(1 + \alpha\eta_j p_{m,j})^2}.
\end{aligned} \tag{18}$$

We now need to show that the above derivative is positive, in order to achieve an increasing margin. We will now use some formulas of useful forms to show this increase.

Denote the function $f(t) := t(1 - t)/(1 + ct)^2$ with $t \in [0, 1]$ and $c \geq 0$. We will show that this function is increasing. The derivative of f w.r.t. t is

$$\frac{df(t)}{dt} = \frac{(1 - 2t)(1 + ct)^2 - (t - t^2)2(1 + ct)c}{(1 + ct)^4} = \frac{1 - (c + 2)t}{(1 + ct)^3}.$$

Taking $t := \eta_k(\mathbf{x})$ and $c := \alpha p_{m,k}(\mathbf{x})$, we get that $f(t)$ is increasing when $(2 + \alpha p_{m,k}(\mathbf{x}))\eta_k(\mathbf{x}) < 1$ for any $k \in \mathcal{Y}$. Therefore we get $f(\eta_i(\mathbf{x})) > f(\eta_j(\mathbf{x}))$ and consequently

$$\frac{\eta_i(1 - \eta_i)p_{m,i}}{(1 + \alpha\eta_i p_{m,i})^2} > \frac{\eta_j(1 - \eta_j)p_{m,i}}{(1 + \alpha\eta_j p_{m,i})^2} \tag{19}$$

when the condition $(2 + \alpha p_{m,k}(\mathbf{x}))\eta_k(\mathbf{x}) < 1$ is satisfied for any $k \in \mathcal{Y}$.

Next, we define the function $h(t) := t/(1 + ct)^2$ with $t \in [0, 1]$ and $c \geq 0$. We will also show that this function is increasing. The derivative of h w.r.t. t is

$$\frac{dh(t)}{dt} = \frac{(1 + ct)^2 - t(1 + ct)(2c)}{(1 + ct)^4} = \frac{1 - ct}{(1 + ct)^3}.$$

Taking $t := p_{m,k}(\mathbf{x})$ and $c := \alpha\eta_k(\mathbf{x})$, we get that $h(t)$ is increasing when $\alpha p_{m,k}(\mathbf{x})\eta_k(\mathbf{x}) < 1$ for any $k \in \mathcal{Y}$. Therefore, we get $h(p_{m,i}(\mathbf{x})) > h(p_{m,j}(\mathbf{x}))$ and consequently

$$\frac{\eta_j(1 - \eta_j)p_{m,i}}{(1 + \alpha\eta_j p_{m,i})^2} > \frac{\eta_j(1 - \eta_j)p_{m,j}}{(1 + \alpha\eta_j p_{m,j})^2} \tag{20}$$

when the condition $\alpha p_{m,k}(\mathbf{x})\eta_k(\mathbf{x}) < 1$ for any $k \in \mathcal{Y}$.

Merging the conditions of (19) and (20)

$$\begin{aligned}
2\eta_k(\mathbf{x}) + \alpha p_{m,k}(\mathbf{x})\eta_k(\mathbf{x}) &< 1 & \forall k \in \mathcal{Y} \\
\alpha p_{m,k}(\mathbf{x})\eta_k(\mathbf{x}) &< 1 & \forall k \in \mathcal{Y}
\end{aligned}$$

gives the total condition, but is equal to only the first condition. The first condition yields

$$\alpha < (\eta_k(\mathbf{x})^{-1} - 2)/p_{m,k}(\mathbf{x}) \quad \forall k \in \mathcal{Y}$$

Combining (19) and (20) yields

$$\frac{\eta_i(1 - \eta_i)p_{m,i}}{(1 + \alpha\eta_i p_{m,i})^2} > \frac{\eta_j(1 - \eta_j)p_{m,j}}{(1 + \alpha\eta_j p_{m,j})^2},$$

so $\frac{d\Delta_{g^*}(\mathbf{x})}{d\alpha} > 0$ of (18) is satisfied, which finishes the proof. \square

References

- [1] Md Mobin Akhtar et al. “Stock market prediction based on statistical data using machine learning algorithms”. In: *Journal of King Saud University-Science* 34.4 (2022), p. 101940.
- [2] Peter L Bartlett and Marten H Wegkamp. “Classification with a Reject Option using a Hinge Loss”. In: *Journal of Machine Learning Research* 9.8 (2008).
- [3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*. Vol. 4. 4. Springer, 2006.
- [4] Nontawat Charoenphakdee et al. “Classification with Rejection Based on Cost-sensitive Classification”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1507–1517.
- [5] Chi-Keung Chow. “An optimum character recognition system using decision functions”. In: *IRE Transactions on Electronic Computers* 4 (1957), pp. 247–254.
- [6] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. “Learning with rejection”. In: *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*. Springer. 2016, pp. 67–82.
- [7] Yoav Freund and Robert E Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.
- [8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)”. In: *The annals of statistics* 28.2 (2000), pp. 337–407.
- [9] Yves Grandvalet et al. “Support Vector Machines with a Reject Option”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc., 2008, pp. 537–544.
- [10] Celestine Iwendi et al. “COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients”. In: *Journal of Experimental & Theoretical Artificial Intelligence* (2022), pp. 1–21.
- [11] Deeksha Kaul, Harika Raju, and BK Tripathy. “Deep learning in healthcare”. In: *Deep Learning in Data Analytics*. Springer, 2022, pp. 97–115.
- [12] Alex Krizhevsky and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. Toronto, Ontario: University of Toronto, 2009.
- [13] Yann LeCun et al. “Handwritten Digit Recognition with a Back-Propagation Network”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 2. Morgan-Kaufmann, 1989, pp. 396–404.
- [14] David Madras, Toni Pitassi, and Richard Zemel. “Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018, pp. 6150–6160.
- [15] Hamed Masnadi-shirazi and Nuno Vasconcelos. “On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc., 2008, pp. 1049–1056.

- [16] Hussein Mozannar and David Sontag. “Consistent Estimators for Learning to Defer to an Expert”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 7076–7087.
- [17] Chenri Ni et al. “On the Calibration of Multiclass Classification with Rejection”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019, pp. 2586–2596.
- [18] Maithra Raghu et al. “The algorithmic automation problem: Prediction, triage, and human effort”. In: *arXiv preprint arXiv:1903.12220* (2019).
- [19] Harish G Ramaswamy et al. “On the consistency of output code based learning algorithms for multiclass learning problems”. In: *Conference on Learning Theory*. PMLR. 2014, pp. 885–902.
- [20] Mark D Reid and Robert C Williamson. “Composite binary losses”. In: *Journal of Machine Learning Research* 11 (2010), pp. 2387–2422.
- [21] Ambuj Tewari and Peter L. Bartlett. “On the Consistency of Multiclass Classification Methods”. In: *Journal of Machine Learning Research* 8.36 (2007), pp. 1007–1025.
- [22] Shrawan Kumar Trivedi. “A study of machine learning classifiers for spam detection”. In: *The 4th International Symposium on Computational and Business Intelligence*. IEEE. 2016, pp. 176–180.
- [23] Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. “Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles”. In: *arXiv preprint arXiv:2210.16955* (2022).
- [24] Rajeev Verma and Eric Nalisnick. “Calibrated Learning to Defer with One-vs-All Classifiers”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 22184–22202.
- [25] Ming Yuan and Marten Wegkamp. “Classification Methods with Reject Option Based on Convex Risk Minimization”. In: *Journal of Machine Learning Research* 11.1 (2010).
- [26] Éloi Zablocki et al. “Explainability of deep vision-based autonomous driving systems: Review and challenges”. In: *International Journal of Computer Vision* (2022), pp. 1–28.
- [27] Sergey Zagoruyko and Nikos Komodakis. “Wide Residual Networks”. In: *British Machine Vision Conference 2016*. British Machine Vision Association. 2016.
- [28] Tong Zhang. “Statistical analysis of some multi-category large margin classification methods”. In: *Journal of Machine Learning Research* 5.Oct (2004), pp. 1225–1251.