Theoretically Grounded Loss Functions and Algorithms for Score-Based Multi-Class Abstention

Anqi Mao
Courant Institute

Mehryar Mohri
Google Research and Courant Institute

Yutao Zhong
Courant Institute

Abstract

Learning with abstention is a key scenario where the learner can abstain from making a prediction at some cost. In this paper, we analyze the score-based formulation of learning with abstention in the multi-class classification setting. We introduce new families of surrogate losses for the abstention loss function, which include the state-of-the-art surrogate losses in the single-stage setting and a novel family of loss functions in the two-stage setting. We prove strong non-asymptotic and hypothesis set-specific consistency guarantees for these surrogate losses, which upper-bound the estimation error of the abstention loss function in terms of the estimation error of the surrogate loss. Our bounds can help compare different score-based surrogates and guide the design of novel abstention algorithms by minimizing the proposed surrogate losses. We experimentally evaluate our new algorithms on CIFAR-10, CIFAR-100, and SVHN datasets and the practical significance of our new surrogate losses and two-stage abstention algorithms. Our results also show that the relative performance of the state-of-the-art score-based surrogate losses can vary across datasets.

1 Introduction

In many applications, incorrect predictions can be costly and it is then preferable to abstain from making predictions for some input instances, since the cost of abstention is typically less significant. As an example, in medical diagnosis, the cost of an incorrect diagnosis is incommensurable since the patient's health may be jeopardized. In contrast, the cost of abstention is typically that of resorting to some additional laboratory tests. For a spoken-dialog system, an incorrect prediction may result in canceling a credit card, for a bank, or shipping the wrong medication to the patient, for a pharmacy, while the cost of abstaining is limited to that of switching to a human operator.

A related problem arises in applications where a learning model distilled from a very complex one is used, due to its more modest inference cost. However, since it is less accurate, one may need to resort to abstention for some inputs and instead predict using the more complex model, despite its higher inference cost. This problem of *deferring* to an alternative model, in fact to a human in some cases, can also be viewed as a special case of the general abstention scenario (Madras et al., 2018; Raghu et al., 2019a; Mozannar and Sontag, 2020; Okati et al., 2021; Wilder et al., 2021; Verma and Nalisnick, 2022; Narasimhan et al., 2022; Verma et al., 2023). In other applications such as information extraction or natural language text generation or question-answering, the output is sometimes not factual (Filippova, 2020; Maynez et al., 2020). It can then be important to learn to abstain from responding to avoid such *hallucinations* and instead defer to a more costly predictor.

The scenario of classification with abstention is very broad and admits increasingly many important applications, including as a subroutine for other algorithms such as active learning (Zhang and Chaudhuri, 2016a) or dual purpose learning (Amin et al., 2021). But, how should we formulate the problem of multi-class classification with abstention and when should we abstain?

There is a vast literature related to the problem of abstention or rejection. Here, we briefly discuss work directly related to this study and give a more detailed discussion in Appendix A. A standard method for abstention adopted in the past, which covers a very large number of publications (e.g., Herbei and Wegkamp (2005); Bartlett and Wegkamp (2008); Yuan and Wegkamp (2010); Lei (2014); Denis and Hebiri (2020)) and dates back to the early work of Chow (1957, 1970), is the

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

so-called *confidence-based abstention*. This consists of first learning a predictor and then abstaining when the score returned by the predictor falls below some fixed threshold. Herbei and Wegkamp (2005) examined binary classification with abstention by giving the optimal rule for these ternary functions. Bartlett and Wegkamp (2008) formulated a loss function for this setting taking into consideration the abstention cost c and suggested to learn a predictor using a *double hinge loss* that they showed benefits from consistency results. Yuan and Wegkamp (2010) investigated the necessary and sufficient condition for consistency of convex risk minimization with respect to the abstention loss and obtained the corresponding excess error bounds in the same setting. Other variants of this framework have also been studied in (Lei, 2014; Denis and Hebiri, 2020).

However, Cortes, DeSalvo, and Mohri (2016a, 2023) argued that, in general, confidence-based abstention is suboptimal, unless the predictor learned is the Bayes classifier. They showed that, in general, even in simple cases, no threshold-based abstention can achieve the desired result. They introduced a novel framework for abstention that consists of learning *simultaneously* both a predictor h and a rejector r that, in general, can be distinct from a threshold-based function. They further defined a *predictor-rejector formulation* loss function for the pair (h, r), taking into consideration the abstention cost c. The authors gave Rademacher complexity-based generalization bounds for this learning problem. They also suggested several surrogate loss functions for the abstention loss in the binary classification setting, and further showed that these surrogate losses benefitted from consistency guarantees. They designed algorithms based on these surrogate losses, which they showed empirically outperform confidence-based abstention baselines. This work had multiple follow-up studies, including a theoretical and algorithmic study of boosting with abstention (Cortes et al., 2016b) and a study of the extension of the results to multi-class setting (Ni et al., 2019). These authors argued that the design of calibrated or Bayes-consistent surrogate losses in the multi-class classification setting based on the predictor-rejector abstention loss of Cortes et al. (2016a) was difficult and left that as an open problem. Recently, Mao et al. (2024b) introduced several new theoretical and algorithmic findings within this framework, effectively addressing the open question. Furthermore, Mohri et al. (2024) explored the framework from the perspective of learning with a fixed predictor, applying their novel algorithms to decontextualization tasks.

Mozannar and Sontag (2020) proposed instead for the multi-class abstention setting a *score-based formulation*, where, in addition to the standard scoring functions associated to each label, a new scoring function is associated to a new rejection label. Rejection takes places when the score given to the rejection label is higher than other scores and the rejector is therefore implicitly defined via this specific rule. The authors suggested a surrogate loss for their approach based on the cross-entropy (logistic loss with softmax applied to neural networks outputs), which they proved to be Bayes-consistent. More recently, Cao et al. (2022) gave a more general family of Bayes-consistent surrogate losses for the score-based formulation that can be built upon any consistent loss for the standard multi-class classification problem. Most recent research by Mozannar et al. (2023) demonstrates that cross-entropy score-based surrogate losses are not realizable $\mathcal{H}$-consistent, as defined by Long and Servedio (2013); Zhang and Agarwal (2020), in relation to abstention loss. Instead, the authors propose a novel surrogate loss that is proved to be realizable $\mathcal{H}$-consistent when $\mathcal{H}$ is *closed under scaling*, although its Bayes-consistency remains unclear. The challenge of devising a surrogate loss that exhibits both Bayes-consistency and realizable $\mathcal{H}$-consistency remains an open problem.

This paper presents a series of new theoretical and algorithmic results for multi-class classification for the score-based abstention formulation. In Section 2, we formalize the setting and first define explicitly the underlying abstention loss. We then show how the general family of surrogate losses introduced by Cao et al. (2022) can be naturally derived from that expression in Section 3.1.

More importantly, we prove *$\mathcal{H}$-consistency bounds* for these surrogate losses (Section 3.2), which are non-asymptotic and hypothesis set-specific guarantees upper-bounding the estimation error of the abstention loss function in terms of the estimation error of the surrogate loss (Awasthi et al., 2022b). These provide stronger guarantees than Bayes-consistency guarantees, which only provide an asymptotic guarantee and hold only for the full family of measurable functions. We first derive our guarantees for a broad family of score-based abstention surrogates, which we name *cross-entropy score-based surrogate losses*. These include the surrogate losses in (Mozannar and Sontag, 2020; Cao et al., 2022), for which our guarantees admit their Bayes-consistency as a special case. Our theory can also help compare different surrogate losses. To make it more explicit, we give an explicit analysis of the minimizability gaps appearing in our bounds. We further prove a general result showing that an $\mathcal{H}$-consistency bound in standard classification yields immediately an $\mathcal{H}$-consistency bound for score-based abstention losses. Minimization of these new surrogate losses directly leads to new algorithm for multi-class abstention.

In Section 4, we analyze a two-stage algorithmic scheme often more relevant in practice, for which we give surrogate losses that we prove to benefit from $\mathcal{H}$-consistency bounds. These are also non-asymptotic and hypothesis set-specific guarantees upper-bounding the estimation error of the abstention loss function in terms of the estimation error of the first-stage surrogate loss and second-stage one. Minimizing these new surrogate losses directly leads to new algorithm for multi-class abstention.

In Section 5, we demonstrate that our proposed two-stage score-based surrogate losses are not only Bayes-consistent, but also realizable $\mathcal{H}$-consistent. This effectively addresses the open question posed by Mozannar et al. (2023) and highlights the benefits of the two-stage formulation.

In Section 6, we show that our $\mathcal{H}$-consistency bounds can be directly used to derive finite sample estimation bounds for a surrogate loss minimizer of the abstention loss. These are more favorable and more relevant guarantee than a similar finite sample guarantee that could be derived from an excess error bound.

In Section 7, we report the results of several experiments comparing these algorithms and discuss them in light of our theoretical guarantees. Our empirical results show, in particular, that the two-stage score-based abstention surrogate loss consistently outperforms the state-of-the-art cross-entropy scored-based abstention surrogate losses on CIFAR-10, CIFAR-100 and SVHN, while highlighting that the relative performance of the state-of-the-art cross-entropy scored-based abstention losses varies by the datasets. We present a summary of our main contribution as follows and start with a formal description of the problem formulations.

- Derivation of the cross-entropy score-based surrogate loss from first principles, which include the state-of-the-art surrogate losses as special cases.
- $\mathcal{H}$-consistency bounds for cross-entropy score-based surrogate losses, which can help theoretically compare different cross-entropy score-based surrogate losses and guide the design of a multi-class abstention algorithm in comparison to the existing asymptotic consistency guarantees.
- A novel family of surrogate loss functions in the two-stage setting and their strong $\mathcal{H}$-consistency bounds guarantees.
- Realizable $\mathcal{H}$-consistency guarantees of proposed two-stage score-based surrogate loss, which effectively addresses the open question posed by Mozannar et al. (2023) and highlights the benefits of the two-stage formulation.
- Extensive experiments demonstrating the practical significance of our new surrogate losses and the varying relative performance of the state-of-the-art cross-entropy score-based surrogate losses across datasets.

2 Preliminary

We consider the standard multi-class classification setting with an input space $\mathcal{X}$ and a set of $n \geq 2$ classes or labels $\mathcal{Y} = \{1, \ldots, n\}$. We will denote by $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ and by $p(x, y)$, the conditional probability of $Y = y$ given $X = x$, that is $p(x, y) = \mathcal{D}(Y = y \mid X = x)$. We will also use $p(x) = (p(x, 1), \ldots, p(x, n))$ to denote the vectors of these probabilities for a given x.

We study the learning scenario of multi-class classification with abstention in the *score-based formulation* proposed by Mozannar and Sontag (2020) and recently studied by Cao et al. (2022).

Score-Based Abstention Formulation In this formulation of the abstention problem, the label set $\mathcal{Y}$ is augmented with an additional category $(n+1)$ corresponding to abstention. We denote by $\mathcal{Y} \cup \{n+1\} = \{1, \ldots, n, n+1\}$ the augmented set and consider a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X} \times (\mathcal{Y} \cup \{n+1\})$ to $\mathbb{R}$. The label associated by $h \in \mathcal{H}$ to an input $x \in \mathcal{X}$ is denoted by $\mathsf{h}(x)$ and defined by $\mathsf{h}(x) = n+1$ if $h(x, n+1) \geq \max_{y \in \mathcal{Y}} h(x, y)$; otherwise, $\mathsf{h}(x)$ is defined as an element in $\mathcal{Y}$ with the highest score, $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y)$, with an arbitrary but fixed deterministic strategy for breaking ties. When $\mathsf{h}(x) = n+1$, the learner abstains from making a prediction for x and incurs a cost $c(x)$. Otherwise, it predicts the label $y = \mathsf{h}(x)$. The *score-based abstention loss* L_{abs} for this formulation is defined as follows for any $h \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\mathsf{L}_{\text{abs}}(h, x, y) = \mathbb{1}_{\mathsf{h}(x) \neq y} \mathbb{1}_{\mathsf{h}(x) \neq n+1} + c(x) \mathbb{1}_{\mathsf{h}(x) = n+1}. \quad (1)$$

Thus, when it does not abstain, $\mathsf{h}(x) \neq n+1$, the learner incurs the familiar zero-one classification loss and when it abstains, $\mathsf{h}(x) = n+1$, the cost $c(x)$. Given a finite sample drawn i.i.d. from $\mathcal{D}$, the learning problem consists of selecting a hypothesis h in $\mathcal{H}$ with small expected score-based abstention loss, $\mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathsf{L}_{\text{abs}}(h, x, y)]$. Note that the cost c implicitly controls the rejection rate when minimizing the abstention loss.

Optimizing the score-based abstention loss is intractable for most hypothesis sets. Thus, instead, learning algorithms for this scenario must resort to a surrogate loss L for L_{abs}. In the next sections, we will define score-based surrogate losses and analyze

their properties. Given a loss function L, we denote by $\mathcal{E}_{\mathsf{L}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathsf{L}(h,x,y)]$ the generalization error or expected loss of h and by $\mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) = \inf_{h\in\mathcal{H}} \mathcal{E}_{\mathsf{L}}(h)$ the minimal generalization error. In the following, to simplify the presentation, we assume that the cost function $c \in (0,1)$ is constant. However, many of our results extend straightforwardly to the general case.

$\mathcal{H}$-Consistency Bounds We will seek to derive $\mathcal{H}$-*consistency bounds* for L. These are strong guarantees that take the form of inequalities establishing a relationship between the abstention loss L_{abs} of any hypothesis $h \in \mathcal{H}$ and the surrogate loss L associated with it (Awasthi et al., 2021a,b, 2022a,b, 2023, 2024; Mao et al., 2023c,d,e; Zheng et al., 2023; Mao et al., 2023b,f). These are bounds of the form $\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \leq f(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))$, for some non-decreasing function f, that upper-bound the estimation error of the loss L_{abs} in terms of that of L for a given hypothesis set $\mathcal{H}$. Thus, they show that if we can reduce the surrogate estimation error $(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))$ to $\epsilon > 0$, then the estimation error of L_{abs} is guaranteed to be at most $f(\epsilon)$. These guarantees are non-asymptotic and take into consideration the specific hypothesis set $\mathcal{H}$ used.

Minimizability Gaps A key quantity appearing in these bounds is the *minimizability gap*, denoted by $\mathcal{M}_{\mathsf{L}}(\mathcal{H})$ and defined by $\mathcal{M}_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}} \mathbb{E}_y[\mathsf{L}(h,X,y) \mid X = x]\big]$ for a given hypothesis set $\mathcal{H}$. Thus, the minimizability gap for a hypothesis set $\mathcal{H}$ and loss function L measures the difference of the best-in-class expected loss and the expected pointwise infimum of the loss. Since the infimum is super-additive, it follows that the minimizability gap is always non-negative. When the loss function L depends only on $h(x,\cdot)$ for all h, x, and $y \in \mathcal{Y}$, that is, $\mathsf{L}(h,x,y) = \Psi(h(x,1),\ldots,h(x,n+1),y)$ for some function Ψ, it can be shown that the minimizability gap vanishes for the family of all measurable functions: $\mathcal{M}(\mathcal{H}_{\text{all}}) = 0$ (Steinwart, 2007, lemma 2.5). However, in general, the minimizability gap is non-zero for restricted hypothesis sets $\mathcal{H}$ and is therefore essential to analyze. It is worth noting that the minimizability gap can be upper bounded by the approximation error $\mathcal{A}_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\text{all}}} \mathbb{E}_y[\mathsf{L}(h,X,y) \mid X = x]\big]$. However, the minimizability gap is a more refined quantity than the approximation error and can lead to more favorable guarantees (see Appendix D).

3 Single-stage score-based formulation

In this section, we first derive the general form of a family of surrogate loss functions L for L_{abs} by analyzing the abstention loss L_{abs}. Next, we give $\mathcal{H}$-consistency bounds for these surrogate losses, which provide non-asymptotic hypothesis set-specific guarantees upper-bounding the estimation error of the loss function L_{abs} in terms of estimation error of L.

3.1 General surrogate losses

Consider a hypothesis h in the score-based setting. Note that for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $\mathsf{h}(x) = n+1$ implies $\mathsf{h}(x) \neq y$, therefore, we have: $\mathbb{1}_{\mathsf{h}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)=n+1} = \mathbb{1}_{\mathsf{h}(x)=n+1}$. Thus, $\mathsf{L}_{\text{abs}}(h,x,y)$ can be rewritten as follows:

$$\begin{aligned}
\mathsf{L}_{\text{abs}}(h,x,y) &= \mathbb{1}_{\mathsf{h}(x)\neq y}\big(1 - \mathbb{1}_{\mathsf{h}(x)=n+1}\big) + c\mathbb{1}_{\mathsf{h}(x)=n+1} \\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} - \mathbb{1}_{\mathsf{h}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)=n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1} \\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} - \mathbb{1}_{\mathsf{h}(x)=n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1} \\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} + (c-1)\mathbb{1}_{\mathsf{h}(x)=n+1} \\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} + (1-c)\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c - 1.
\end{aligned}$$

In view of this expression, since the last term $(c-1)$ is a constant, if ℓ is a surrogate loss for the zero-one multi-class classification loss over the set of labels $\mathcal{Y}$, then L defined as follows is a natural surrogate loss for L_{abs}: for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$,

$$\mathsf{L}(h,x,y) = \ell(h,x,y) + (1-c)\,\ell(h,x,n+1). \tag{2}$$

This is precisely the form of the surrogate losses proposed by Mozannar and Sontag (2020), for which the analysis just presented gives a natural derivation. This is also the form of the surrogate losses adopted by Cao et al. (2022).

3.2 $\mathcal{H}$-Consistency bounds guarantees

Cao et al. (2022) presented a nice study of the surrogate loss L for a specific family of zero-one loss surrogates ℓ. The authors showed that the surrogate loss L is Bayes-consistent with respect to the score-based abstention loss L_{abs} when ℓ is Bayes-consistent with respect to the multi-class zero-one classification loss ℓ_{0-1}. Bayes-consistency guarantees that, asymptotically, a nearly optimal minimizer of L over the family of all measurable functions is also a nearly optimal minimizer of L_{abs}. However, this does not provide any guarantee for a restricted subset $\mathcal{H}$ of the family of all measurable functions. It also provides no guarantee for approximate minimizers since convergence could be arbitrarily slow and the result is only asymptotic.

In the following, we will prove $\mathcal{H}$-consistency bounds guarantees, which are stronger results that are non-asymptotic and that hold for a restricted hypothesis set $\mathcal{H}$. The specific instance of our results where $\mathcal{H}$ is the family of all measurable functions directly implies the Bayes-consistency results of Cao et al. (2022).

$\mathcal{H}$-Consistency Bounds for Cross-Entropy Abstention Losses We first prove $\mathcal{H}$-consistency bounds for a broad family of score-based abstention surrogate losses L_μ, that we will refer to as *cross-entropy score-based surrogate losses*. These are loss functions defined by

$$\mathsf{L}_\mu(h,x,y) = \ell_\mu(h,x,y) + (1-c)\,\ell_\mu(h,x,n+1), \quad (3)$$

where, for any $h \in \mathcal{H}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\mu \geq 0$,

$$\ell_\mu(h,x,y) = \begin{cases} \frac{1}{1-\mu}\left(\left[\sum_{y' \in \mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y)}\right]^{1-\mu} - 1\right) & \mu \neq 1 \\ \log\left(\sum_{y' \in \mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y)}\right) & \mu = 1. \end{cases}$$

The loss function ℓ_μ coincides with the (multinomial) logistic loss (Verhulst, 1838, 1845; Berkson, 1944, 1951) when $\mu = 1$, matches the generalized cross-entropy loss (Zhang and Sabuncu, 2018) when $\mu \in (1,2)$, and the mean absolute loss (Ghosh et al., 2017) when $\mu = 2$. Thus, the cross-entropy score-based surrogate losses L_μ include the abstention surrogate losses proposed in (Mozannar and Sontag, 2020) which correspond to the special case of $\mu = 1$ and the abstention surrogate losses adopted in (Cao et al., 2022), which correspond to the special case of $\mu \in [1,2]$.

We say that a hypothesis set $\mathcal{H}$ is *symmetric* when the scoring functions it induces do not depend on any particular ordering of the labels, that is when there exists a family $\mathcal{F}$ of functions f mapping from $\mathcal{X}$ to $\mathbb{R}$ such that, for any $x \in \mathcal{X}$, $\{[h(x,1),\ldots,h(x,n),h(x,n+1)]\colon h \in \mathcal{H}\} = \{[f_1(x),\ldots,f_n(x),f_{n+1}(x)]\colon f_1,\ldots,f_{n+1} \in \mathcal{F}\}$. We say that a hypothesis set $\mathcal{H}$ is *complete* if the set of scores it generates spans $\mathbb{R}$, that is, $\{h(x,y)\colon h \in \mathcal{H}\} = \mathbb{R}$, for any $(x,y) \in \mathcal{X} \times \mathcal{Y} \cup \{n+1\}$. Common hypothesis sets used in practice, such as the family of linear models, that of neural networks and of course that of all measurable functions are all symmetric and complete. The guarantees given in the following result are thus general and widely applicable.

Theorem 1 (**$\mathcal{H}$-consistency bounds for cross-entropy score-based surrogates**)**.** *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality holds:*

$$\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \leq \Gamma_\mu\left(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\right),$$

where $\Gamma_\mu(t) = \begin{cases} \sqrt{(2-c)2^\mu(2-\mu)t} & \mu \in [0,1) \\ \sqrt{2(2-c)(n+1)^{\mu-1}t} & \mu \in [1,2) \\ (\mu-1)(n+1)^{\mu-1}t & \mu \in [2,+\infty). \end{cases}$

The proof is given in Appendix C.1. It consists of analyzing the calibration gap of the score-based abstention loss L_{abs} and that of L_μ, and of finding a concave function Γ_μ relating these two quantities. Note that our proofs and results are distinct, original, and more complex than those in the standard setting (Mao et al., 2023c), where the standard loss ℓ_μ is analyzed. Establishing $\mathcal{H}$-consistency bounds for L_μ is more intricate compared to ℓ_μ. This is because the target loss in the score-based multi-class abstention is inherently different from that of the standard multi-class scenario (the multi-class zero-one loss). Thus, we need to tackle a more complex calibration gap, integrating both the conditional probability vector and the cost function. This complexity presents an added layer of challenge when attempting to establish a lower bound for the calibration gap of the surrogate loss in relation to the target loss in the score-based abstention setting.

To understand the result, consider first the case where the minimizability gaps are zero. As mentioned earlier, this would be the case, for example, when $\mathcal{H}$ is the family of all measurable functions or when $\mathcal{H}$ contains the Bayes classifier. In that case, the theorem shows that if the estimation loss $(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}))$ is reduced to ϵ, then, for $\mu \in [0,2)$, in particular for the logistic score-based surrogate ($\mu = 1$) and the generalized cross-entropy score-based surrogate ($\mu \in (1,2)$), modulo a multiplicative constant, the score-based abstention estimation loss $(\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}))$ is bounded by $\sqrt{\epsilon}$. The bound is even more favorable for the mean absolute error score-based surrogate ($\mu = 2$) or for cross-entropy score-based surrogate L_μ with $\mu \in (2,+\infty)$ since in that case, modulo a multiplicative constant, the score-based abstention estimation loss $(\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}))$ is bounded by ϵ.

These are strong results since they are not asymptotic and are hypothesis set-specific. In particular, Theorem 1 provides stronger guarantees than the Bayes-consistency results of Mozannar and Sontag (2020) or Cao et al. (2022) for cross-entropy abstention surrogate losses (3) with the logistic loss ($\mu = 1$), generalized cross-entropy loss ($\mu \in (1,2)$) and mean absolute error loss ($\mu = 2$) adopted for ℓ. These Bayes-consistency results can be obtained by considering the special case of $\mathcal{H}$ being the family of all measurable functions and taking the limit.

Moreover, Theorem 1 also provides similar guarantees for other types of cross-entropy score-based surrogate losses, such as $\mu \in [0,1)$ and $\mu \in [2,+\infty)$, which are new surrogate losses for score-based multi-class abstention that, to the best of our knowledge, have not been previously studied in the literature. In particular, our $\mathcal{H}$-consistency bounds can help theoretically compare different cross-entropy score-based surrogate losses and

guide the design of a multi-class abstention algorithm. In contrast, asymptotic consistency guarantees given for a subset of cross-entropy score-based surrogate losses in (Mozannar and Sontag, 2020; Cao et al., 2022) do not provide any such comparative information.

Recall that the minimizability gap is always upper bounded by the approximation error. By Lemma 5 in Appendix C, the minimizability gap for the abstention loss $\mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H})$ coincides with the approximation error $\mathcal{A}_{\mathsf{L}_{\text{abs}}}(\mathcal{H})$ when the labels generated by the hypothesis set encompass all possible outcomes, which naturally holds true for typical hypothesis sets. However, for a surrogate loss, the minimizability gap is in general a more refined quantity than the approximation error and can lead to more favorable guarantees. More precisely, $\mathcal{H}$-consistency bounds expressed in terms of minimizability gaps are better and more significant than the excess error bounds expressed in terms of approximation errors (See Appendix D for a more detailed discussion).

3.3 Analysis of minimizability gaps

In general, the minimizability gaps do not vanish and their magnitude, $\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$, is important to take into account when comparing cross-entropy score-based surrogate losses, in addition to the functional form of Γ_μ. Thus, we will specifically analyze them below. Note that the dependency of the multiplicative constant on the number of classes in some of these bounds ($\mu \in (1, +\infty)$) makes them less favorable, while for $\mu \in [0, 1]$, the bounds do not depend on the number of classes.

In the deterministic cases where for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, either $p(x, y) = 0$ or 1, the pointwise expected loss admits an explicit form. Thus, the following result characterizes the minimizability gaps directly in those cases.

Theorem 2 (**Characterization of minimizability gaps**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for the cross-entropy score-based surrogate losses L_μ and any deterministic distribution, the minimizability gaps can be characterized as follows:*

$$\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \begin{cases} \frac{1}{1-\mu}\left[\left[1 + (1-c)^{\frac{1}{2-\mu}}\right]^{2-\mu} - (2-c)\right] & \mu \notin \{1, 2\} \\ -\log\left(\frac{1}{2-c}\right) - (1-c)\log\left(\frac{1-c}{2-c}\right) & \mu = 1 \\ 1 - c & \mu = 2. \end{cases}$$

See Appendix C.2 for the proof. By l'Hôpital's rule, $\mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$ is continuous as a function of μ at $\mu = 1$. In light of the equality $\lim_{x \to 0^+}\left(1 + u^{\frac{1}{x}}\right)^x = \max\{1, u\} = 1$, for $u \in [0, 1]$, $\mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$ is continuous as a function of μ at $\mu = 2$. Moreover, for any $c \in (0, 1)$, $\mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$ is decreasing with respect to μ. On the other hand, since the function $\mu \mapsto \frac{1}{1-\mu}\left(t^{1-\mu} - 1\right)\mathbb{1}_{\mu \neq 1} + \log(t)\mathbb{1}_{\mu=1}$ is decreasing for any $t > 0$, we obtain that ℓ_μ is decreasing with respect to μ, which implies that L_μ is decreasing and then $\mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H})$ is decreasing with respect to μ as well. For a specific problem, a favorable $\mu \in [0, \infty)$ is one that minimizes $\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$, which, in practice, can be selected via cross-validation.

3.4 General transformation

More generally, we prove the following result, which shows that an $\mathcal{H}$-consistency bound for ℓ with respect to the zero-one loss, yields immediately an $\mathcal{H}$-consistency bound for L with respect to L_{abs}.

Theorem 3. *Assume that ℓ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1} with a concave function Γ, that is, for all $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma\left(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\right).$$

Then, L defined by (2) admits an $\mathcal{H}$-consistency bound with respect to L_{abs} with the functional form $(2-c)\Gamma(\frac{t}{2-c})$, that is, for all $h \in \mathcal{H}$, we have

$$\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \leq (2-c)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{2-c}\right).$$

The proof is given in Appendix C.3. Awasthi et al. (2022b) recently presented a series of results providing $\mathcal{H}$-consistency bounds for common surrogate losses in the standard multi-class classification, including max losses such as those of Crammer and Singer (2001), sum losses such as those of Weston and Watkins (1998) and constrained losses such as the loss functions adopted by Lee et al. (2004). Thus, plugging in any of those $\mathcal{H}$-consistency bounds in Theorem 3 yields immediately a new $\mathcal{H}$-consistency bound for the corresponding score-based abstention surrogate losses.

4 Two-stage score-based formulation

In the single-stage scenario discussed in Section 3, the learner simultaneously learns when to abstain and how to make predictions otherwise. However, in practice often there is already a predictor available and retraining can be very costly. A two-stage solution is thus much more relevant for those critical applications, where the learner only learns when to abstain in the second stage

based on the predictor trained in the first stage. With the two stage solution, we can improve the performance of a large pre-trained model by teaching it the option of abstaining without having to retrain the model. In this section, we analyze a two-stage algorithmic scheme, for which we propose surrogate losses that we prove to benefit from $\mathcal{H}$-consistency bounds.

Given a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}\times(\mathcal{Y}\cup\{n+1\})$ to $\mathbb{R}$, it can be decomposed into $\mathcal{H}=\mathcal{H}_{\mathcal{Y}}\times\mathcal{H}_{n+1}$, where $\mathcal{H}_{\mathcal{Y}}$ denotes the hypothesis set spanned by the first n scores corresponding to the labels, and $\mathcal{H}_{n+1}$ represents the hypothesis set spanned by the last score corresponding to the additional category. We consider the following two-stage algorithmic scheme: in the first stage, we learn a hypothesis $h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}$ by optimizing a surrogate loss ℓ for standard multi-class classification; in the second stage, we fix the $h_{\mathcal{Y}}$ learned in the first stage and then learn a hypothesis $h_{n+1}\in\mathcal{H}_{n+1}$ by optimizing a surrogate loss function $\ell_{h_{\mathcal{Y}}}$ defined for any $h_{n+1}\in\mathcal{H}_{n+1}$ and $(x,y)\in\mathcal{X}\times\mathcal{Y}$ by

$$\ell_{h_{\mathcal{Y}}}(h_{n+1},x,y)=\mathbb{1}_{h_{\mathcal{Y}}(x)\neq y}\Phi\Big(h_{n+1}(x)-\max_{y\in\mathcal{Y}}h_{\mathcal{Y}}(x,y)\Big)+c\Phi\Big(\max_{y\in\mathcal{Y}}h_{\mathcal{Y}}(x,y)-h_{n+1}(x)\Big),\quad(4)$$

where Φ is a decreasing function. The learned hypothesis $h\in\mathcal{H}$ corresponding to those two stages can be expressed as $h=(h_{\mathcal{Y}},h_{n+1})$. We note that the first stage consists of the familiar task of finding a predictor using a standard surrogate loss such as the logistic loss $\ell(h,x,y)=\log\Big(\sum_{y'\in\mathcal{Y}}e^{h(x,y')-h(x,y)}\Big)$ (or cross-entropy combined with the softmax). Recall that the learner abstains from making a prediction for x and incurs a cost c when $h_{n+1}(x)\geq\max_{y\in\mathcal{Y}}h_{\mathcal{Y}}(x,y)$. In the second stage, the first term of (4) encourages abstention for an input instance whose prediction made by the pre-trained predictor $h_{\mathcal{Y}}$ is incorrect, while the second term penalizes abstention according to the cost c. The function Φ can be chosen as any margin-based loss function in binary classification, including the exponential loss or the logistic loss.

Let $\ell_{0-1}^{\text{binary}}$ be the binary zero-one classification loss. Then, the two-stage surrogate losses benefit from the $\mathcal{H}$-consistency bounds shown in Theorem 4. For a fixed parameter τ, we define the τ-translated hypothesis set of $\mathcal{H}_{n+1}$ by $\mathcal{H}_{n+1}^{\tau}=\{h_{n+1}-\tau:h_{n+1}\in\mathcal{H}_{n+1}\}$.

Theorem 4 (**$\mathcal{H}$-consistency bounds for two-stage surrogates**). *Given a hypothesis set $\mathcal{H}=\mathcal{H}_{\mathcal{Y}}\times\mathcal{H}_{n+1}$. Assume that ℓ admits an $\mathcal{H}_{\mathcal{Y}}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1} and that Φ admits an $\mathcal{H}_{n+1}^{\tau}$-consistency bound with respect to the binary zero-one classification loss $\ell_{0-1}^{\text{binary}}$ for any $\tau\in\mathbb{R}$. Thus, there are non-decreasing concave functions Γ_1 and Γ_2 such that, for all $h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}$, $h_{n+1}^{\tau}\in\mathcal{H}_{n+1}^{\tau}$ and $\tau\in\mathbb{R}$, we have*

$$\begin{aligned}&\mathcal{E}_{\ell_{0-1}}(h_{\mathcal{Y}})-\mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}_{\mathcal{Y}})+\mathcal{M}_{\ell_{0-1}}(\mathcal{H}_{\mathcal{Y}})\\&\quad\leq\Gamma_1\big(\mathcal{E}_{\ell}(h_{\mathcal{Y}})-\mathcal{E}^*_{\ell}(\mathcal{H}_{\mathcal{Y}})+\mathcal{M}_{\ell}(\mathcal{H}_{\mathcal{Y}})\big)\\&\mathcal{E}_{\ell_{0-1}^{\text{binary}}}(h_{n+1}^{\tau})-\mathcal{E}^*_{\ell_{0-1}^{\text{binary}}}(\mathcal{H}_{n+1}^{\tau})+\mathcal{M}_{\ell_{0-1}^{\text{binary}}}(\mathcal{H}_{n+1}^{\tau})\\&\quad\leq\Gamma_2\big(\mathcal{E}_{\Phi}(h_{n+1}^{\tau})-\mathcal{E}^*_{\Phi}(\mathcal{H}_{n+1}^{\tau})+\mathcal{M}_{\Phi}(\mathcal{H}_{n+1}^{\tau})\big).\end{aligned}$$

Then, the following holds for all $h=(h_{\mathcal{Y}},h_{n+1})\in\mathcal{H}$:

$$\begin{aligned}&\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h)-\mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H})+\mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H})\\&\leq\Gamma_1\big(\mathcal{E}_{\ell}(h_{\mathcal{Y}})-\mathcal{E}^*_{\ell}(\mathcal{H}_{\mathcal{Y}})+\mathcal{M}_{\ell}(\mathcal{H}_{\mathcal{Y}})\big)\\&+(1+c)\Gamma_2\left(\frac{\mathcal{E}_{\ell_{h_{\mathcal{Y}}}}(h_{n+1})-\mathcal{E}^*_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1})+\mathcal{M}_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1})}{c}\right),\end{aligned}$$

where the constant factors $(1+c)$ and $\frac{1}{c}$ can be removed when Γ_2 is linear.

The proof is given in Appendix C.4. The assumptions in Theorem 4 are mild and hold for common hypothesis sets such as linear models and neural networks with common surrogate losses in the binary and multi-class classification, as shown by (Awasthi et al., 2022a,b). Recall that the minimizability gaps vanish when $\mathcal{H}_{\mathcal{Y}}$ and $\mathcal{H}_{n+1}$ are the family of all measurable functions or when $\mathcal{H}_{\mathcal{Y}}$ and $\mathcal{H}_{n+1}$ contain the Bayes predictors. In their absence, the theorem shows that if the estimation loss $(\mathcal{E}_{\ell}(h_{\mathcal{Y}})-\mathcal{E}^*_{\ell}(\mathcal{H}_{\mathcal{Y}}))$ is reduced to ϵ_1 and the estimation loss $(\mathcal{E}_{\ell_{h_{\mathcal{Y}}}}(h_{n+1})-\mathcal{E}^*_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1}))$ to ϵ_2, then, modulo constant factors, the score-based abstention estimation loss $(\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h)-\mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}))$ is bounded by $\Gamma_1(\epsilon_1)+\Gamma_2(\epsilon_2)$. Thus, this gives a strong guarantee for the surrogate losses described in this two-stage setting.

5 Realizable $\mathcal{H}$-consistency and benefits of two-stage surrogate losses

Mozannar et al. (2023) recently showed that cross-entropy score-based surrogate losses are not realizable $\mathcal{H}$-consistent, as defined by Long and Servedio (2013); Zhang and Agarwal (2020), in relation to abstention loss. Instead, the authors proposed a novel surrogate loss that is proved to be realizable $\mathcal{H}$-consistent when $\mathcal{H}$ is *closed under scaling*, although its Bayes-consistency remains unclear. Devising a surrogate loss that exhibits both Bayes-consistency and realizable $\mathcal{H}$-consistency remains an open problem. A hypothesis set $\mathcal{H}$ is said to be *closed under scaling* if, for any hypothesis h belonging to $\mathcal{H}$, the scaled hypothesis αh also belongs to $\mathcal{H}$ for all $\alpha\in\mathbb{R}$.

We prove in Theorem 8 of Appendix C.5, that for any realizable distribution, when both the first-stage surrogate estimation loss $\mathcal{E}_{\ell}(h_{\mathcal{Y}})-\mathcal{E}^*_{\ell}(\mathcal{H}_{\mathcal{Y}})$ and the

second-stage surrogate estimation loss $\mathcal{E}_{\ell_{h_y}}(h_{n+1}) - \mathcal{E}^*_{\ell_{h_y}}(\mathcal{H}_{n+1})$ converge to zero, the abstention estimation loss $\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H})$ also approaches zero. This implies that the two-stage score-based surrogate loss is realizable $\mathcal{H}$-consistent with respect to L_{abs}, which provides a significant advantage over the single-stage cross-entropy score-based surrogate loss. It is important to note that Theorem 4 shows that the two-stage formulation is also Bayes-consistent. This addresses the open problem in (Mozannar et al., 2023) and highlights the benefits of the two-stage formulation. In the following section, our empirical results further demonstrate that the two-stage score-based surrogate loss outperforms the state-of-the-art cross-entropy score-based surrogate loss.

6 Finite sample guarantees

Our $\mathcal{H}$-consistency bounds enable the direct derivation of finite-sample estimation bounds for a surrogate loss minimizer. These are expressed in terms of the Rademacher complexity of the hypothesis set $\mathcal{H}$, the loss function, and the minimizability gaps. Here, we provide a simple illustration based on Theorem 1.

Let $\widehat{h}_S$ be the empirical minimizer of the surrogate loss L_μ: $\widehat{h}_S = \operatorname{argmin}_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \mathsf{L}_\mu(h, x_i, y_i)$, for an i.i.d sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of size m. Let $\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H})$ be the Rademacher complexity of the set $\mathcal{H}_{\mathsf{L}_\mu} = \{(x, y) \mapsto \mathsf{L}_\mu(h, x, y) \colon h \in \mathcal{H}\}$ and B_{L_μ} an upper bound on the surrogate loss L_μ. By using the standard Rademacher complexity bounds (Mohri et al., 2018), for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$\left|\mathcal{E}_{\mathsf{L}_\mu}(h) - \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(h)\right| \leq 2\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H}) + B_{\mathsf{L}_\mu}\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Fix $\epsilon > 0$. By the definition of the infimum, there exists $h^* \in \mathcal{H}$ such that $\mathcal{E}_{\mathsf{L}_\mu}(h^*) \leq \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \epsilon$. By definition of $\widehat{h}_S$, we have

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_\mu}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H})\\
&= \mathcal{E}_{\mathsf{L}_\mu}(\widehat{h}_S) - \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H})\\
&\leq \mathcal{E}_{\mathsf{L}_\mu}(\widehat{h}_S) - \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(h^*) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H})\\
&\leq \mathcal{E}_{\mathsf{L}_\mu}(\widehat{h}_S) - \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{\mathsf{L}_\mu,S}(h^*) - \mathcal{E}^*_{\mathsf{L}_\mu}(h^*) + \epsilon\\
&\leq 2\left[2\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H}) + B_{\mathsf{L}_\mu}\sqrt{\frac{\log(2/\delta)}{2m}}\right] + \epsilon.
\end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, it implies:

$$\mathcal{E}_{\mathsf{L}_\mu}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) \leq 4\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H}) + 2B_{\mathsf{L}_\mu}\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Plugging in this inequality in the bound of Theorem 1, we obtain that for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d sample S of size m, the following finite sample guarantee holds for $\widehat{h}_S$:

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\text{abs}}}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H})\\
&\leq \Gamma_\mu\left(4\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H}) + 2B_{\mathsf{L}_\mu}\sqrt{\tfrac{\log\frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\right) - \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}).
\end{aligned}$$

To our knowledge, these are the first abstention estimation loss guarantees for empirical minimizers of a cross-entropy score-based surrogate loss. Our comments about the properties of Γ_μ below Theorem 1, in particular its functional form or its dependency on the number of classes n, similarly apply here. Similar finite sample guarantees can also be derived based on Theorems 3 and 4.

As commented before Section 3.3, for a surrogate loss, the minimizability gap is in general a more refined quantity than the approximation error, while for the abstention loss, these two quantities coincide for typical hypothesis sets (See Appendix D). Thus, our bound can be rewritten as follows for typical hypothesis sets:

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\text{abs}}}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}_{\text{all}})\\
&\leq \Gamma_\mu\left(4\mathfrak{R}^{\mathsf{L}_\mu}_m(\mathcal{H}) + 2B_{\mathsf{L}_\mu}\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\right).
\end{aligned}$$

Our guarantee is thus more favorable and more relevant than a similar finite sample guarantee where $\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})$ is replaced with $\mathcal{A}_{\mathsf{L}_\mu}(\mathcal{H})$, which could be derived from an excess error bound.

7 Experiments

In this section, we report the results of experiments comparing the single-stage and two-stage score-based abstention surrogate losses, for three widely used datasets CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011).

Experimental Settings As with (Mozannar and Sontag, 2020; Cao et al., 2022), we use ResNet (He et al., 2016) and WideResNet (WRN) (Zagoruyko and Komodakis, 2016) with ReLU activations. Here, ResNet-n denotes a residual network with n convolutional layers and WRN-n-k denotes a residual network with n convolutional layers and a widening factor k. We trained ResNet-34 for CIFAR-10 and SVHN, and WRN-28-10 for CIFAR-100. We applied standard data augmentations, 4-pixel padding with 32×32 random crops and random horizontal flips for CIFAR-10 and CIFAR-100. We used Stochastic Gradient Descent (SGD) with Nesterov momentum (Nesterov, 1983) and set batch size 1,024 and weight decay 1×10^{-4} in the training. We trained for 200 epochs using the cosine decay learning

Table 1: Abstention Loss for Models Obtained with Different Surrogate Losses; Mean ± Standard Deviation for Both Two-Stage Score-Based Abstention Surrogate Loss and The State-Of-The-Art Cross-Entropy Score-Based Surrogate Losses in (Mozannar and Sontag, 2020) ($\mu = 1.0$) and (Cao et al., 2022) ($\mu = 1.7$).

METHOD	DATASET	ABSTENTION LOSS
Cross-entropy score-based ($\mu = 1.0$)	CIFAR-10	4.48% ± 0.10%
cross-entropy score-based ($\mu = 1.7$)		3.62% ± 0.07%
Two-stage score-based		**3.22% ± 0.04%**
Cross-entropy score-based ($\mu = 1.0$)	CIFAR-100	10.40% ± 0.10%
Cross-entropy score-based ($\mu = 1.7$)		14.99% ± 0.01%
Two-stage score-based		**9.54% ± 0.07%**
Cross-entropy score-based ($\mu = 1.0$)	SVHN	1.61% ± 0.06%
Cross-entropy score-based ($\mu = 1.7$)		2.16% ± 0.04%
Two-stage score-based		**0.93% ± 0.02%**

rate schedule (Loshchilov and Hutter, 2016) with the initial learning rate of 0.1.

For each dataset, the cost value c was selected to be close to the best-in-class zero-one classification loss, which are $\{0.05, 0.15, 0.03\}$ for CIFAR-10, CIFAR-100 and SVHN respectively, since a too small value leads to abstention on almost all points and a too large one leads to almost no abstention. Other neighboring values for c lead to similar results.

The abstention surrogate loss proposed in (Mozannar and Sontag, 2020) corresponds to the special case of cross-entropy score-based surrogate losses L_μ with $\mu = 1$, and meanwhile the abstention surrogate loss adopted in (Cao et al., 2022) corresponds to the special case of cross-entropy score-based surrogate losses L_μ with $\mu = 1.7$. Note that the simple confidence-based approach by thresholding estimators of conditional probability typically does not perform as well as these state-of-the-art surrogate losses (Cao et al., 2022). For our two-stage score-based abstention surrogate loss, we adopted the logistic loss in the first stage and the exponential loss $\Phi(t) = \exp(-t)$ in the second stage.

Evaluation We evaluated all the models based on the abstention loss L_{abs}, and reported the mean and standard deviation over three trials.

Results Table 1 shows that the two-stage score-based surrogate losses consistently outperform the cross-entropy score-based surrogate losses used in the state-of-the-art algorithms (Mozannar and Sontag, 2020; Cao et al., 2022) for all the datasets. Table 1 also shows the relative performance of the cross-entropy surrogate (3) with ℓ_μ adopted as the generalized cross-entropy loss ($\mu = 1.7$) and that with ℓ_μ adopted as the logistic loss ($\mu = 1.0$) varies by the datasets.

As show in Section 4 and Section 5, the two-stage surrogate losses benefit from the guarantees of both realizable $\mathcal{H}$-consistency and Bayes-consistency while the cross-entropy surrogate loss does not exhibit realizable $\mathcal{H}$-consistency, as shown by Mozannar et al. (2023). This explains the superior performance of two-stage surrogate losses over the cross-entropy surrogate loss. It is worth noting that the hypothesis set we used for each dataset is sufficiently rich, and the experimental setup closely resembles a realizable scenario.

As our theoretical analysis (Theorem 1 and Theorem 2) suggests, the relative performance variation between the cross-entropy surrogate loss with $\mu = 1.0$ used in (Mozannar and Sontag, 2020) and the cross-entropy surrogate loss with $\mu = 1.7$ used in (Cao et al., 2022) can be explained by the functional forms of their $\mathcal{H}$-consistency bounds and the magnitude of their minimizability gaps. Specifically, the dependency of the multiplicative constant on the number of classes in $\mathcal{H}$-consistency bounds (Theorem 1) for the cross-entropy surrogate loss with $\mu = 1.7$ makes it less favorable when dealing with a large number of classes, such as in the case of CIFAR-100. This suggests that the recent observation made in (Cao et al., 2022) that the cross-entropy surrogate with $\mu = 1.7$ outperforms the one with $\mu = 1.0$ does not apply to the scenario where the evaluation involves datasets like CIFAR-100. For a more comprehensive discussion of our experimental results, please refer to Appendix B.

8 Conclusion

Our comprehensive study of score-based multi-class abstention introduced novel surrogate loss families with strong hypothesis set-specific and non-asymptotic theoretical guarantees. Empirical results demonstrate the practical advantage of these surrogate losses and their derived algorithms. This work establishes a powerful framework for designing new, more reliable abstention-aware algorithms applicable across diverse domains.

References

K. Amin, G. DeSalvo, and A. Rostamizadeh. Learning with labeling induced abstentions. In *Advances in Neural Information Processing*, pages 12576–12586, 2021.

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, 2021a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, 2022a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in neural information processing systems*, 2022b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094, 2023.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, pages 1–17, 2024.

G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11405–11414, 2021.

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

J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357—-365, 1944.

J. Berkson. Why I prefer logits to probits. *Biometrics*, 7(4):327—-339, 1951.

A. Bounsiar, E. Grall, and P. Beauseroy. Kernel based rejection method for supervised classification. In *WASET*, 2007.

Y. Cao, T. Cai, L. Feng, L. Gu, J. Gu, B. An, G. Niu, and M. Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *Advances in neural information processing systems*, 2022.

N. Charoenphakdee, Z. Cui, Y. Zhang, and M. Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517, 2021.

G. Chen, X. Li, C. Sun, and H. Wang. Learning to make adherence-aware advice. In *International Conference on Learning Representations*, 2024.

C. Chow. An optimum character recognition system using decision function. *IEEE T. C.*, 1957.

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

E. Chzhen, C. Denis, M. Hebiri, and T. Lorieul. Set-valued classification–overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82, 2016a.

C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2016b.

C. Cortes, G. DeSalvo, and M. Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, pages 1–39, 2023.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec): 265–292, 2001.

C. Denis and M. Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72, 2020.

C. Denis, M. Hebiri, B. N. Njike, and X. Siebert. Active learning algorithm through the lens of rejection arguments. *arXiv preprint arXiv:2208.14682*, 2022.

B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern recognition*, 26(1):155–165, 1993.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

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

C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, pages 973–978, 2001.

K. Filippova. Controlled hallucinations:learning to generate faithfully from noisy data. In *Findings of EMNLP 2020*, 2020.

G. Fumera and F. Roli. Support vector machines with embedded reject option. In *ICPR*, 2002.

G. Fumera, F. Roli, and G. Giacinto. Multiple reject thresholds for improving classification reliability. In *ICAPR*, 2000.

A. Gangrade, A. Kag, and V. Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2179–2187, 2021.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, 2017.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159, 2019.

A. Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

Y. Grandvalet, J. Keshet, A. Rakotomamonjy, and S. Canu. Suppport vector machines with a reject option. In *NIPS*, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

R. Herbei and M. Wegkamp. Classification with reject option. *Can. J. Stat.*, 2005.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Toronto University, 2009.

T. Landgrebe, D. Tax, P. Paclik, and R. Duin. Interaction between classification and reject performance for distance-based reject-option classifiers. *PRL*, 2005.

H. Le Capitaine and C. Frelicot. An optimum class-rejective decision rule and its evaluation. In *International Conference on Pattern Recognition*, pages 3312–3315, 2010.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

J. Lei. Classification with confidence. *Biometrika*, 101 (4):755–769, 2014.

L. Li, M. L. Littman, and T. J. Walsh. Knows what it knows: a framework for self-aware learning. In *International conference on Machine learning*, pages 568–575, 2008.

X. Li, S. Liu, C. Sun, and H. Wang. When no-rejection learning is optimal for regression with rejection. In *International Conference on Artificial Intelligence and Statistics*, 2024.

P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809, 2013.

I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, 2023a.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds: Characterization and extensions. In *Advances in Neural Information Processing Systems*, 2023b.

A. Mao, M. Mohri, and Y. Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, 2023c.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, 2023d.

A. Mao, M. Mohri, and Y. Zhong. Ranking with abstention. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023e.

A. Mao, M. Mohri, and Y. Zhong. Structured prediction with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2023f.

A. Mao, M. Mohri, and Y. Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024a.

A. Mao, M. Mohri, and Y. Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, 2024b.

J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173.

I. Melvin, J. Weston, C. S. Leslie, and W. S. Noble. Combining classifiers for improved classification of proteins from sequence or structure. *BMCB*, 2008.

C. Mohri, D. Andor, E. Choi, M. Collins, A. Mao, and Y. Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087, 2020.

H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pages 10520–10545, 2023.

H. Narasimhan, W. Jitkrittum, A. K. Menon, A. S. Rawat, and S. Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, 2022.

H. Narasimhan, A. K. Menon, W. Jitkrittum, and S. Kumar. Learning to reject meets ood detection: Are all abstentions created equal? *arXiv preprint arXiv:2301.12386*, 2023.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. akad. nauk Sssr*, 269:543–547, 1983.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems*, 2011.

C. Ni, N. Charoenphakdee, J. Honda, and M. Sugiyama. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, pages 2582–2592, 2019.

N. Okati, A. De, and M. Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

C. S. Pereira and A. Pires. On optimal reject rules and ROC curves. *PRL*, 2005.

T. Pietraszek. Optimizing abstaining classifiers using ROC. In *ICML*, 2005.

N. Puchkin and N. Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In *Conference on Learning Theory*, pages 3806–3832, 2021.

M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019a.

M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290, 2019b.

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

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

N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In *Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR, 2021.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2): 225–287, 2007.

D. M. Tax and R. P. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008.

F. Tortorella. An optimal reject rule for binary classifiers. In *ICAPR*, 2001.

P. F. Verhulst. Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique*, 10:113—-121, 1838.

P. F. Verhulst. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1—-42, 1845.

R. Verma and E. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202, 2022.

R. Verma, D. Barrejón, and E. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11415–11434, 2023.

J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

Y. Wiener and R. El-Yaniv. Agnostic selective classification. In *Advances in neural information processing systems*, 2011.

Y. Wiener and R. El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *J. Mach. Learn. Res.*, 16:713–745, 2015.

B. Wilder, E. Horvitz, and E. Kamar. Learning to complement humans. In *International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.

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

M. Yuan and M. Wegkamp. SVMs with a reject option. In *Bernoulli*, 2011.

S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

C. Zhang and K. Chaudhuri. The extended Littlestone's dimension for learning with mistakes and abstentions. In *Conference on Learning Theory*, 2016a.

C. Zhang and K. Chaudhuri. The extended littlestone's dimension for learning with mistakes and abstentions. In *Conference on Learning Theory*, pages 1584–1616, 2016b.

M. Zhang and S. Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, 2020.

Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, 2018.

C. Zheng, G. Wu, F. Bao, Y. Cao, C. Li, and J. Zhu. Revisiting discriminative vs. generative classifiers: Theory and implications. *arXiv preprint arXiv:2302.02334*, 2023.

Y. Zhu and R. Nowak. Efficient active learning with abstention. *arXiv preprint arXiv:2204.00043*, 2022.

L. Ziyin, Z. Wang, P. P. Liang, R. Salakhutdinov, L.-P. Morency, and M. Ueda. Deep gamblers: Learning to abstain with portfolio theory. *arXiv preprint arXiv:1907.00208*, 2019.

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]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

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]

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). [No]
   (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). [Not Applicable]

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]
   (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]

Contents of Appendix

A Related work

The problem of abstention (or rejection) has been studied in several publications in the past. Chow (1957, 1970) studied the trade-off between error rate and rejection rate and also presented an analysis of the Bayes optimal decision for this setting. Later, Fumera et al. (2000) suggested a multiple thresholds rule when the a posteriori probabilities were affected by errors. Tortorella (2001) introduced an optimal rejection rule for binary classifiers based on the Receiver Operating Characteristic curve (ROC curve) and Pereira and Pires (2005) compared their approach with that of Chow (1970). A number of other publications suggested various rejection techniques to decrease the misclassification rate, though without a theoretical analysis (Fumera and Roli, 2002; Pietraszek, 2005; Bounsiar et al., 2007; Landgrebe et al., 2005; Melvin et al., 2008). Classification with a rejection option that incurs a cost was examined by Herbei and Wegkamp (2005), who gave the optimal rule for these ternary functions.

El-Yaniv et al. (2010) and Wiener and El-Yaniv (2011) proposed and studied a framework for *selective classification* based on a classifier and a selector and an objective function defined as the average loss on accepted samples, normalized by the average selection. Several related connections with abstention have been studied, including active learning (El-Yaniv and Wiener, 2012; Wiener et al., 2015; Wiener and El-Yaniv, 2015; Puchkin and Zhivotovskiy, 2021; Denis et al., 2022; Zhu and Nowak, 2022), rejection in the multi-class setting (Dubuisson and Masson, 1993; Tax and Duin, 2008; Le Capitaine and Frelicot, 2010), reinforcement learning (Li et al., 2008), online learning (Zhang and Chaudhuri, 2016b), modern confidence-based rejection techniques (Geifman and El-Yaniv, 2017), neural network architectures for abstention (Geifman and El-Yaniv, 2019), loss functions derived from the doubling rate of gambling (Ziyin et al., 2019), disparity-free methods (Schreuder and Chzhen, 2021), the abstention problem within a "confidence set" framework (Gangrade et al., 2021; Chzhen et al., 2021) and out-of-distribution detection (Narasimhan et al., 2023).

A standard method for abstention adopted in the past, which covers a very large number of publications and dates back to the early work of Chow (1957, 1970), is the so-called *confidence-based abstention*. This consists of first learning a predictor and then abstaining when the score returned by the predictor falls below some fixed threshold. Bartlett and Wegkamp (2008) formulated a loss function for this setting taking into consideration the abstention cost c and suggested to learn a predictor using a *double hinge loss* that they showed benefits from consistency results. Several other publications followed up on this approach (Grandvalet et al., 2008; Yuan and Wegkamp, 2010, 2011). Yuan and Wegkamp (2010) investigated the necessary and sufficient condition for consistency of convex risk minimization with respect to the abstention loss and obtained the corresponding excess error bounds in the same setting. Other variants of this framework have also been studied in (Lei, 2014; Denis and Hebiri, 2020).

However, Cortes, DeSalvo, and Mohri (2016a, 2023) argued that, in general, confidence-based abstention is suboptimal, unless the predictor learned is the Bayes classifier. They showed that, in general, even in simple cases, no threshold-based abstention can achieve the desired result. They introduced a novel framework for abstention that consists of learning *simultaneously* both a predictor h and a rejector r that, in general, can be distinct from a threshold-based function. They further defined a *predictor-rejector formulation* loss function for the pair (h, r), taking into consideration the abstention cost c. The authors gave Rademacher complexity-based generalization bounds for this learning problem. They also suggested several surrogate loss functions for the abstention loss in the binary classification setting, and further showed that these surrogate losses benefitted from consistency guarantees. They designed algorithms based on these surrogate losses, which they showed empirically outperform confidence-based abstention baselines. This work had multiple follow-up studies, including a theoretical and algorithmic study of boosting with abstention (Cortes et al., 2016b) and a study of the extension of the results to multi-class setting (Ni et al., 2019). These authors argued that the design of calibrated or Bayes-consistent surrogate losses in the multi-class classification setting based on the predictor-rejector abstention loss of Cortes et al. (2016a) was difficult and left that as an open problem. Recently, Mao et al. (2024b) introduced several new theoretical and algorithmic findings within this framework, effectively addressing the open question. Furthermore, Mohri et al. (2024) explored the framework from the perspective of learning with a fixed predictor, applying their novel algorithms to decontextualization tasks. Li et al. (2024) investigated the Bayes-consistency of no-rejection learning in the setting of regression with abstention. Ramaswamy et al. (2018) also studied the confidence-based abstention in the multi-class classification, where they show certain multi-class hinge loss formulations and a new constructed polyhedral binary encoded predictions (BEP) surrogate loss are Bayes-consistent. Charoenphakdee et al. (2021) proposed a cost-sensitive approach for the multi-class abstention, where they decomposed the

multi-class problem into multiple binary cost-sensitive classification problems (Elkan, 2001). They proposed a family of cost-sensitive one-versus-all surrogate losses, which are Bayes-consistent in that setting.

Mozannar and Sontag (2020) proposed instead for the multi-class abstention setting a *score-based formulation*, where, in addition to the standard scoring functions associated to each label, a new scoring function is associated to a new rejection label. Rejection takes places when the score given to the rejection label is higher than other scores and the rejector is therefore implicitly defined via this specific rule. The authors suggested a surrogate loss for their approach based on the cross-entropy (logistic loss with softmax applied to neural networks outputs), which they proved to be Bayes-consistent. More recently, Cao et al. (2022) gave a more general family of Bayes-consistent surrogate losses for the score-based formulation that can be built upon any consistent loss for the standard multi-class classification problem.

A problem directly related to our study is that of learning to defer, which can be directly cast as an instance of learning with abstention. There are several recent publications studying this formulation of the problem (Madras et al., 2018; Raghu et al., 2019a,b; Mozannar and Sontag, 2020; Okati et al., 2021; Wilder et al., 2021; Bansal et al., 2021; Verma and Nalisnick, 2022; Narasimhan et al., 2022; Verma et al., 2023; Mao et al., 2023a, 2024a). Raghu et al. (2019b); Wilder et al. (2021); Bansal et al. (2021) studied confidence-based methods to make deferral decisions, which may be sub-optimal for low capital models (Cortes et al., 2016a, 2023). To overcome this limitation, Mozannar and Sontag (2020) proposed cost-sensitive logistic loss and Verma and Nalisnick (2022) proposed cost-sensitive one-versus-all proper composite loss (Reid and Williamson, 2010), both in the score-based formulation. Verma et al. (2023) further generalized the surrogate loss in (Verma and Nalisnick, 2022) to the setting of deferring with multiple experts. Furthermore, Mao et al. (2024a) introduced a new and more general family of surrogate losses specifically tailored for this setting and proved that these surrogate losses benefit from strong $\mathcal{H}$-consistency bounds. More recently, Narasimhan et al. (2022) pointed out that the existing surrogate losses for learning to defer (Mozannar and Sontag, 2020; Verma and Nalisnick, 2022) may underfit in an important practical setting and proposed a post-hoc correction for these loss functions. Moreover, Mao et al. (2023a) studied a two-stage scenario for learning to defer with multiple experts, where a predictor is first trained using a standard loss function such as cross-entropy, and a deferral function is subsequently learned. They introduced a novel family of surrogate loss functions and algorithms for this crucial scenario, supported by $\mathcal{H}$-consistency bounds. Most recently, Chen et al. (2024) incorporated deferral into a sequential decision-making model, leading to improved theoretical convergence and empirical performance.

B Discussion on experiments

This section presents a detailed analysis of the experimental results.

For CIFAR-10, the two-stage score-based abstention surrogate loss outperforms the cross-entropy scored-based abstention surrogate loss ($\mu = 1.0$) used in (Mozannar and Sontag, 2020) by 1.26%, and outperforms the cross-entropy scored-based abstention surrogate loss ($\mu = 1.7$) used in (Cao et al., 2022) by 0.4%. Our results for the score-based surrogate losses are also consistent with those of Cao et al. (2022), who showed that the scored-based abstention loss (2) with ℓ_μ adopted as the generalized cross-entropy loss ($\mu = 1.7$) performs better than the scored-based abstention loss with ℓ_μ adopted as the logistic loss ($\mu = 1$). This agrees with our theoretical analysis based on $\mathcal{H}$-consistency bounds and minimizability gaps in Theorem 1 and Theorem 2, since both losses have the same square-root functional form while the magnitude of the minimizability gap decreases with μ in light of the fact that $\mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H})$ is close for both losses.

Table 1 also shows that on SVHN, using deeper neural networks than (Cao et al., 2022), the cross-entropy scored-based abstention loss ($\mu = 1.7$) actually performs worse than the cross-entropy scored-based abstention loss ($\mu = 1$) in (Mozannar and Sontag, 2020), in contrast with the opposite results observed in (Cao et al., 2022) when using shallower neural networks. This is consistent with our theoretical analysis based on their $\mathcal{H}$-consistency bounds (Theorem 1): the minimizability gaps are basically the same while the dependency of the multiplicative constant on the number of classes appears for $\mu = 1.7$, which makes the scored-based abstention loss (2) with ℓ_μ adopted as the generalized cross-entropy loss ($\mu = 1.7$) less favorable. Here too, the two-stage score-based abstention surrogate loss is superior to both, with an abstention loss 1.23% lower than that of (Cao et al., 2022) and 0.68% lower than that of (Mozannar and Sontag, 2020).

To further test the algorithms, we also carried out experiments on CIFAR-100, with deeper neural networks.

Table 1 shows that score-based abstention loss with generalized cross-entropy adopted in (Cao et al., 2022) does not perform well in this case. In contrast, the score-based abstention loss with the logistic loss adopted in (Mozannar and Sontag, 2020) performs better and surpasses it by 4.59%. Our two-stage score-based abstention loss is still the most favorable, here too, with 0.86% lower abstention loss than that of (Mozannar and Sontag, 2020). As with the case of SVHN, the inferior performance of the cross-entropy scored-based abstention surrogate loss ($\mu = 1.7$) can be seen from the dependency of the multiplicative constant on the number of classes in $\mathcal{H}$-consistency bounds (Theorem 1), which is worse when the number of classes is much larger as in the case of CIFAR-100.

C Proofs for score-based abstention losses

To begin with the proof, we first introduce some notation. Recall that we denote by $p(x, y) = \mathcal{D}(Y = y|X = x)$ the conditional probability of $Y = y$ given $X = x$. For simplicity of the notation, we let $p(x, n+1) = 1 - c$ and denote by $y_{\max} \in \mathcal{Y} \bigcup \{n+1\}$ the label associated to an input $x \in \mathcal{X}$, defined as $y_{\max} = n+1$ if $1 - c \geq \max_{y \in \mathcal{Y}} p(x, y)$; otherwise, $y_{\max}$ is defined as an element in $\mathcal{Y}$ with the highest conditional probability, $y_{\max} = \operatorname{argmax}_{y \in \mathcal{Y}} p(x, y)$, with the same deterministic strategy for breaking ties as that of $\mathsf{h}(x)$. Thus, the generalization error for a score-based abstention surrogate loss can be rewritten as $\mathcal{E}_{\mathsf{L}}(h) = \mathbb{E}_X[\mathcal{C}_{\mathsf{L}}(h, x)]$, where $\mathcal{C}_{\mathsf{L}}(h, x)$ is the conditional L-risk, defined by

$$\mathcal{C}_{\mathsf{L}}(h, x) = \sum_{y \in \mathcal{Y} \bigcup \{n+1\}} p(x, y) \ell(h, x, y).$$

We denote by $\mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x) = \inf_{h \in \mathcal{H}} \mathcal{C}_{\mathsf{L}}(h, x)$ the minimal conditional L-risk. Then, the minimizability gap can be rewritten as follows:

$$\mathcal{M}_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) - \mathbb{E}_X[\mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x)].$$

We further refer to $\mathcal{C}_{\mathsf{L}}(h, x) - \mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x)$ as the calibration gap and denote it by $\Delta\mathcal{C}_{\mathsf{L},\mathcal{H}}(h, x)$. We first prove a lemma on the calibration gap of the score-based abstention loss. For any $x \in \mathcal{X}$, we will denote by $\mathsf{H}(x)$ the set of labels generated by hypotheses in $\mathcal{H}$: $\mathsf{H}(x) = \{\mathsf{h}(x) \colon h \in \mathcal{H}\}$.

Lemma 5. *For any $x \in \mathcal{X}$, the minimal conditional L_{abs}-risk and the calibration gap for L_{abs} can be expressed as follows:*

$$\mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}, x) = 1 - \max_{y \in \mathsf{H}(x)} p(x, y)$$
$$\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h, x) = \max_{y \in \mathsf{H}(x)} p(x, y) - p(x, \mathsf{h}(x)).$$

Proof. The conditional L_{abs}-risk of h can be expressed as follows:

$$\mathcal{C}_{\mathsf{L}_{\text{abs}}}(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \mathbb{1}_{\mathsf{h}(x) \neq y} \mathbb{1}_{\mathsf{h}(x) \neq n+1} + c \mathbb{1}_{\mathsf{h}(x) = n+1} = 1 - p(x, \mathsf{h}(x)).$$

Then, the minimal conditional L_{abs}-risk is given by

$$\mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}, x) = 1 - \max_{y \in \mathsf{H}(x)} p(x, y),$$

and the calibration gap can be expressed as follows:

$$\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h, x) = \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h, x) - \mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}, x) = \max_{y \in \mathsf{H}(x)} p(x, y) - p(x, \mathsf{h}(x)).$$

This completes the proof. □

Note that when $\mathcal{H}$ is symmetric, $\mathsf{H}(x) = \mathcal{Y} \bigcup \{n+1\}$. By Lemma 5, in those cases, we obtain the following result,

Corollary 6. *Assume that $\mathcal{H}$ is symmetric. Then, for any $x \in \mathcal{X}$, the minimal conditional L_{abs}-risk and the calibration gap for L_{abs} can be expressed as follows:*

$$\mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}, x) = 1 - p(x, y_{\max})$$
$$\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h, x) = p(x, y_{\max}) - p(x, \mathsf{h}(x)).$$

C.1 Proof of $\mathcal{H}$-Consistency bounds for Cross-Entropy Score-Based Surrogates (Theorem 1)

Theorem 1 ($\mathcal{H}$**-consistency bounds for cross-entropy score-based surrogates**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality holds:*

$$\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \leq \Gamma_\mu\Big(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\Big),$$

where $\Gamma_\mu(t) = \begin{cases} \sqrt{(2-c)2^\mu(2-\mu)t} & \mu \in [0,1) \\ \sqrt{2(2-c)(n+1)^{\mu-1}t} & \mu \in [1,2) \\ (\mu-1)(n+1)^{\mu-1}t & \mu \in [2,+\infty). \end{cases}$

Proof. The main proof idea is similar for each case of μ: we will lower bound the calibration gap of L_μ by that of L_{abs} by carefully selecting a hypothesis h_λ in the hypothesis set $\mathcal{H}$. In particular, we analyze different cases as follows.

The Case Where $\mu \in [0,1)$ For any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, choose hypothesis $h_\lambda \in \mathcal{H}$ such that

$$h_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\} \\ \log(\exp[h(x,y_{\max})] + \lambda) & \text{if } y = \mathsf{h}(x) \\ \log(\exp[h(x,\mathsf{h}(x))] - \lambda) & \text{if } y = y_{\max}, \end{cases}$$

where $\lambda = \frac{\exp[h(x,\mathsf{h}(x))]p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}} - \exp[h(x,y_{\max})]p(x,y_{\max})^{\frac{1}{2-\mu}}}{p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}$. The existence of such a h_λ in the hypothesis set $\mathcal{H}$ is guaranteed by the assumption that $\mathcal{H}$ is symmetry and complete. Thus, the calibration gap can be expressed and

lower-bounded as follows:

$$
\begin{aligned}
&(1-\mu)\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\\
&= (1-\mu)\Big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}^*_{\mathsf{L}_\mu}(\mathcal{H},x)\Big)\\
&\geq (1-\mu)\Big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}_{\mathsf{L}_\mu}(h_\lambda,x)\Big)\\
&= p(x,y_{\max})\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})}\right]^{1-\mu} - 1\right) + p(x,\mathsf{h}(x))\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))}\right]^{1-\mu} - 1\right)\\
&- p(x,y_{\max})\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))+\lambda}\right]^{1-\mu} - 1\right) - p(x,\mathsf{h}(x))\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})-\lambda}\right]^{1-\mu} - 1\right)\\
&= p(x,y_{\max})\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-e^{h(x,y_{\max})}}\right]^{1-\mu} - p(x,y_{\max})\left[\frac{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}\Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]}{\Big[e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))}\Big]p(x,y_{\max})^{\frac{1}{2-\mu}}}\right]^{1-\mu}\\
&+ p(x,\mathsf{h}(x))\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))}\right]^{1-\mu} - p(x,\mathsf{h}(x))\left[\frac{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}\Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]}{\Big[e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))}\Big]p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}\right]^{1-\mu}\\
&\geq p(x,y_{\max})\Big[e^{h(x,\mathsf{h}(x))-h(x,y_{\max})}+1\Big]^{1-\mu} - p(x,y_{\max})\left[\frac{p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}{p(x,y_{\max})^{\frac{1}{2-\mu}}}\right]^{1-\mu}\\
&+ p(x,\mathsf{h}(x))\Big[e^{h(x,y_{\max})-h(x,\mathsf{h}(x))}+1\Big]^{1-\mu} - p(x,\mathsf{h}(x))\left[\frac{p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}{p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}\right]^{1-\mu}\\
&\qquad\qquad \Big(\textstyle\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')} \geq e^{h(x,\mathsf{h}(x))} + e^{h(x,y_{\max})}\Big)\\
&\geq p(x,y_{\max})2^{1-\mu} - p(x,y_{\max})^{\frac{1}{2-\mu}}\Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]^{1-\mu}\\
&+ p(x,\mathsf{h}(x))2^{1-\mu} - p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]^{1-\mu}\\
&\qquad\qquad \Big(\text{minimum is attained when } e^{h(x,\mathsf{h}(x))} = e^{h(x,y_{\max})}\Big)\\
&= 2^{1-\mu}\big(p(x,y_{\max}) + p(x,\mathsf{h}(x))\big) - \Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]^{2-\mu}\\
&= 2^{2-\mu}\left[\left(\frac{p(x,y_{\max}) + p(x,\mathsf{h}(x))}{2}\right) - \left[\frac{p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}{2}\right]^{2-\mu}\right]\\
&\geq \frac{1-\mu}{(2-c)2^{\mu}(2-\mu)}\big(p(x,y_{\max}) - p(x,\mathsf{h}(x))\big)^2\\
&\qquad\qquad \Big(p(x,y_{\max}) + p(x,\mathsf{h}(x)) \leq 2-c \text{ and by analyzing the Taylor expansion}\Big)\\
&= \frac{1-\mu}{(2-c)2^{\mu}(2-\mu)}\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)^2 \qquad \text{(Corollary 6)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) &= \mathop{\mathbb{E}}_{X}\big[\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)\big]\\
&\leq \mathop{\mathbb{E}}_{X}\Big[\Gamma_\mu\Big(\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\Big)\Big]\\
&\leq \Gamma_\mu\left(\mathop{\mathbb{E}}_{X}\Big[\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\Big]\right) \qquad (\Gamma_\mu \text{ is concave})\\
&= \Gamma_\mu\Big(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\Big),
\end{aligned}
$$

where $\Gamma_\mu(t) = \sqrt{(2-c)2^{\mu}(2-\mu)t}$.

The Case Where $\mu = 1$ For any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, choose hypothesis $h_\lambda \in \mathcal{H}$ such that

$$h_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\} \\ \log(\exp[h(x,y_{\max})] + \lambda) & \text{if } y = \mathsf{h}(x) \\ \log(\exp[h(x,\mathsf{h}(x))] - \lambda) & \text{if } y = y_{\max} \end{cases}$$

where $\lambda = \frac{\exp[h(x,\mathsf{h}(x))]p(x,\mathsf{h}(x)) - \exp[h(x,y_{\max})]p(x,y_{\max})}{p(x,y_{\max}) + p(x,\mathsf{h}(x))}$. The existence of such a h_λ in hypothesis set $\mathcal{H}$ is guaranteed by the fact that $\mathcal{H}$ is symmetry and complete. Thus, the calibration gap can be expressed and lower-bounded as follows:

$$\begin{aligned}
&\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x) \\
&= \mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}^*_{\mathsf{L}_\mu}(\mathcal{H},x) \\
&\geq \mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}_{\mathsf{L}_\mu}(h_\lambda,x) \\
&= -p(x,y_{\max})\log\left[e^{h(x,y_{\max})}\right] - p(x,\mathsf{h}(x))\log\left[e^{h(x,\mathsf{h}(x))}\right] \\
&\quad + p(x,y_{\max})\log\left[e^{h(x,\mathsf{h}(x))} - \lambda\right] + p(x,\mathsf{h}(x))\log\left[e^{h(x,y_{\max})} + \lambda\right] \\
&= p(x,y_{\max})\log\left[\frac{\left[e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))}\right]p(x,y_{\max})}{e^{h(x,y_{\max})}[p(x,y_{\max}) + p(x,\mathsf{h}(x))]}\right] + p(x,\mathsf{h}(x))\log\left[\frac{\left[e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))}\right]p(x,\mathsf{h}(x))}{e^{h(x,\mathsf{h}(x))}[p(x,y_{\max}) + p(x,\mathsf{h}(x))]}\right] \\
&\geq p(x,y_{\max})\log\left[\frac{2p(x,y_{\max})}{p(x,y_{\max}) + p(x,\mathsf{h}(x))}\right] + p(x,\mathsf{h}(x))\log\left[\frac{2p(x,\mathsf{h}(x))}{p(x,y_{\max}) + p(x,\mathsf{h}(x))}\right] \\
&\qquad\qquad \text{(minimum is attained when } e^{h(x,\mathsf{h}(x))} = e^{h(x,y_{\max})}\text{)} \\
&\geq [p(x,y_{\max}) + p(x,\mathsf{h}(x))] \times \frac{1}{2}\left[\left|\frac{p(x,y_{\max})}{p(x,y_{\max}) + p(x,\mathsf{h}(x))} - \frac{1}{2}\right| + \left|\frac{p(x,\mathsf{h}(x))}{p(x,y_{\max}) + p(x,\mathsf{h}(x))} - \frac{1}{2}\right|\right]^2 \\
&\qquad\qquad \text{(Pinsker's inequality (Mohri et al., 2018, Proposition E.7))} \\
&= [p(x,y_{\max}) + p(x,\mathsf{h}(x))] \times \frac{1}{2}\left[\frac{p(x,y_{\max}) - p(x,\mathsf{h}(x))}{p(x,y_{\max}) + p(x,\mathsf{h}(x))}\right]^2 \qquad (p(x,y_{\max}) \geq p(x,\mathsf{h}(x))) \\
&\geq \frac{1}{2(2-c)}(p(x,y_{\max}) - p(x,\mathsf{h}(x)))^2 \\
&\qquad\qquad (p(x,y_{\max}) + p(x,\mathsf{h}(x)) \leq 2 - c) \\
&= \frac{1}{2(2-c)}\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)^2 \qquad \text{(Corollary 6)}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) &= \mathop{\mathbb{E}}_X[\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)] \\
&\leq \mathop{\mathbb{E}}_X\left[\Gamma_\mu\left(\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\right)\right] \\
&\leq \Gamma_\mu\left(\mathop{\mathbb{E}}_X\left[\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\right]\right) \qquad (\Gamma_\mu \text{ is concave}) \\
&= \Gamma_\mu\left(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\right),
\end{aligned}$$

where $\Gamma_\mu(t) = \sqrt{2(2-c)t}$.

The Case Where $\mu \in [2, +\infty)$ For any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, choose hypothesis $h_\lambda \in \mathcal{H}$ such that

$$h_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\} \\ \log(\exp[h(x,y_{\max})] + \lambda) & \text{if } y = \mathsf{h}(x) \\ \log(\exp[h(x,\mathsf{h}(x))] - \lambda) & \text{if } y = y_{\max} \end{cases}$$

where $\lambda = -\exp[h(x, y_{\max})]$. The existence of such a h_λ in hypothesis set $\mathcal{H}$ is guaranteed by the fact that $\mathcal{H}$ is symmetry and complete. Thus, the calibration gap can be expressed and lower-bounded as follows:

$$
\begin{aligned}
&(\mu-1)\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\\
&= (\mu-1)\big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}^*_{\mathsf{L}_\mu}(\mathcal{H},x)\big)\\
&\geq (\mu-1)\big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}_{\mathsf{L}_\mu}(h_\lambda,x)\big)\\
&= p(x,y_{\max})\left(1-\left[\frac{e^{h(x,y_{\max})}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\right) + p(x,\mathsf{h}(x))\left(1-\left[\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\right)\\
&- p(x,y_{\max})\left(1-\left[\frac{e^{h(x,\mathsf{h}(x))}-\mu}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\right) - p(x,\mathsf{h}(x))\left(1-\left[\frac{e^{h(x,y_{\max})}+\mu}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\right)\\
&= p(x,y_{\max})\left[\frac{e^{h(x,\mathsf{h}(x))}+e^{h(x,y_{\max})}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1} - p(x,y_{\max})\left[\frac{e^{h(x,y_{\max})}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1} - p(x,\mathsf{h}(x))\left[\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\\
&\geq p(x,y_{\max})\left[\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1} - p(x,\mathsf{h}(x))\left[\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}}\right]^{\mu-1}\\
&\qquad\qquad\qquad\qquad ((x+y)^{\mu-1} \geq x^{\mu-1} + y^{\mu-1},\ \forall\, x,y\geq 0,\ \mu\geq 2)\\
&\geq \frac{1}{(n+1)^{\mu-1}}\big(p(x,y_{\max}) - p(x,\mathsf{h}(x))\big) \qquad \left(\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}} \geq \frac{1}{n+1}\right)\\
&= \frac{1}{(n+1)^{\mu-1}}\Delta\mathcal{C}_{\mathsf{L}_{\rm abs},\mathcal{H}}(h,x) \qquad \text{(Corollary 6)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\rm abs}}(h) - \mathcal{E}^*_{\mathsf{L}_{\rm abs}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\rm abs}}(\mathcal{H}) &= \mathop{\mathbb{E}}_X\big[\Delta\mathcal{C}_{\mathsf{L}_{\rm abs},\mathcal{H}}(h,x)\big]\\
&\leq \mathop{\mathbb{E}}_X\Big[\Gamma_\mu\big(\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\big)\Big]\\
&\leq \Gamma_\mu\Big(\mathop{\mathbb{E}}_X\big[\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\big]\Big) && (\Gamma_\mu \text{ is concave})\\
&= \Gamma_\mu\big(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\big),
\end{aligned}
$$

where $\Gamma_\mu(t) = (\mu-1)(n+1)^{\mu-1}t$.

The Case Where $\mu \in (1,2)$ For any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, choose hypothesis $h_\lambda \in \mathcal{H}$ such that

$$
h_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\}\\ \log(\exp[h(x,y_{\max})] + \lambda) & \text{if } y = \mathsf{h}(x)\\ \log(\exp[h(x,\mathsf{h}(x))] - \lambda) & \text{if } y = y_{\max} \end{cases}
$$

where $\lambda = \frac{\exp[h(x,\mathsf{h}(x))]p(x,y_{\max})^{\frac{1}{\mu-2}} - \exp[h(x,y_{\max})]p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}{p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}$. The existence of such a h_λ in hypothesis set $\mathcal{H}$ is guaranteed by the fact that $\mathcal{H}$ is symmetry and complete. Thus, the calibration gap can be lower-bounded as

follows:

$$
\begin{aligned}
&(\mu-1)\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\\
&= (\mu-1)\Big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}^*_{\mathsf{L}_\mu}(\mathcal{H},x)\Big)\\
&\geq (\mu-1)\Big(\mathcal{C}_{\mathsf{L}_\mu}(h,x) - \mathcal{C}_{\mathsf{L}_\mu}(h_\lambda,x)\Big)\\
&= p(x,y_{\max})\left(1-\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})}\right]^{1-\mu}\right) + p(x,\mathsf{h}(x))\left(1-\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))}\right]^{1-\mu}\right)\\
&\quad - p(x,y_{\max})\left(1-\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))+\lambda}\right]^{1-\mu}\right) - p(x,\mathsf{h}(x))\left(1-\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})-\lambda}\right]^{1-\mu}\right)\\
&= -p(x,y_{\max})\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-e^{h(x,y_{\max})}}\right]^{1-\mu} + p(x,y_{\max})\left[\frac{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}\Big[p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}\Big]}{\Big[e^{h(x,y_{\max})}+e^{h(x,\mathsf{h}(x))}\Big]p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}\right]^{1-\mu}\\
&\quad - p(x,\mathsf{h}(x))\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,\mathsf{h}(x))}\right]^{1-\mu} + p(x,\mathsf{h}(x))\left[\frac{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}\Big[p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}\Big]}{\Big[e^{h(x,y_{\max})}+e^{h(x,\mathsf{h}(x))}\Big]p(x,y_{\max})^{\frac{1}{\mu-2}}}\right]^{1-\mu}\\
&\geq \frac{1}{(n+1)^{\mu-1}}\left(p(x,y_{\max})\left[\frac{\Big[e^{h(x,y_{\max})}+e^{h(x,\mathsf{h}(x))}\Big]p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}{e^{h(x,\mathsf{h}(x))}\Big[p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}\Big]}\right]^{\mu-1} - p(x,y_{\max})\Big[e^{h(x,y_{\max})-h(x,\mathsf{h}(x))}\Big]^{\mu-1}\right)\\
&\quad + \frac{1}{(n+1)^{\mu-1}}\left(p(x,\mathsf{h}(x))\left[\frac{\Big[e^{h(x,y_{\max})}+e^{h(x,\mathsf{h}(x))}\Big]p(x,y_{\max})^{\frac{1}{\mu-2}}}{e^{h(x,\mathsf{h}(x))}\Big[p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}\Big]}\right]^{\mu-1} - p(x,\mathsf{h}(x))\right)\\
&\hspace{20em}\left(\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')}} \geq \frac{1}{(n+1)^{\mu-1}}\right)\\
&\geq \frac{1}{(n+1)^{\mu-1}}\left(p(x,y_{\max})\left[\frac{2p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}{p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}\right]^{\mu-1} - p(x,y_{\max})\right)\\
&\quad + \frac{1}{(n+1)^{\mu-1}}\left(p(x,\mathsf{h}(x))\left[\frac{2p(x,y_{\max})^{\frac{1}{\mu-2}}}{p(x,y_{\max})^{\frac{1}{\mu-2}} + p(x,\mathsf{h}(x))^{\frac{1}{\mu-2}}}\right]^{\mu-1} - p(x,\mathsf{h}(x))\right)\\
&\hspace{20em}\left(\text{minimum is attained when } e^{h(x,\mathsf{h}(x))} = e^{h(x,y_{\max})}\right)\\
&= \frac{1}{(n+1)^{\mu-1}}\left(2^{\mu-1}\Big[p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}\Big]^{2-\mu} - p(x,y_{\max}) - p(x,\mathsf{h}(x))\right)\\
&= \frac{2}{(n+1)^{\mu-1}}\left(\left[\frac{p(x,y_{\max})^{\frac{1}{2-\mu}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\mu}}}{2}\right]^{2-\mu} - \frac{p(x,y_{\max}) + p(x,\mathsf{h}(x))}{2}\right)\\
&\geq \frac{\mu-1}{2(2-c)(n+1)^{\mu-1}}\big(p(x,y_{\max}) - p(x,\mathsf{h}(x))\big)^2\\
&\hspace{12em}\big(p(x,y_{\max}) + p(x,\mathsf{h}(x)) \leq 2-c \text{ and by analyzing the Taylor expansion}\big)\\
&= \frac{\mu-1}{2(2-c)(n+1)^{\mu-1}}\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{abs}},\mathcal{H}}(h,x)^2 \qquad \text{(Corollary 6)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\mathrm{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\mathrm{abs}}}(\mathcal{H}) &= \mathop{\mathbb{E}}_X\big[\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{abs}},\mathcal{H}}(h,x)\big]\\
&\leq \mathop{\mathbb{E}}_X\Big[\Gamma_\mu\big(\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\big)\Big]\\
&\leq \Gamma_\mu\Big(\mathop{\mathbb{E}}_X\big[\Delta\mathcal{C}_{\mathsf{L}_\mu,\mathcal{H}}(h,x)\big]\Big) \qquad (\Gamma_\mu \text{ is concave})\\
&= \Gamma_\mu\Big(\mathcal{E}_{\mathsf{L}_\mu}(h) - \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H})\Big),
\end{aligned}
$$

where $\Gamma_\mu(t) = \sqrt{2(2-c)(n+1)^{\mu-1}t}$. □

C.2 Characterization of Minimizability Gaps (Theorem 2)

Theorem 2 (**Characterization of minimizability gaps**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for the cross-entropy score-based surrogate losses L_μ and any deterministic distribution, the minimizability gaps can be characterized as follows:*

$$\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \begin{cases} \frac{1}{1-\mu}\left[\left[1+(1-c)^{\frac{1}{2-\mu}}\right]^{2-\mu} - (2-c)\right] & \mu \notin \{1,2\} \\ -\log\left(\frac{1}{2-c}\right) - (1-c)\log\left(\frac{1-c}{2-c}\right) & \mu = 1 \\ 1-c & \mu = 2. \end{cases}$$

Proof. Let $s_h(x,y) = \frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y} \cup \{n+1\}} h(x,y')} \in [0,1]$, $\forall y \in \mathcal{Y}$. By the definition, for any deterministic distribution, $\mathcal{M}_{\mathsf{L}_\mu}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}_\mu}(\mathcal{H}) - \mathbb{E}_X\left[\inf_{h\in\mathcal{H}} \mathcal{C}_{\mathsf{L}_\mu}(\mathcal{H},x)\right]$, where

$$\begin{aligned}
&\mathcal{C}_{\mathsf{L}_\mu}(h,x) \\
&= \sum_{y \in \mathcal{Y}\cup\{n+1\}} p(x,y)\ell_\mu(h,x,y) \\
&= \ell_\mu(h,x,y_{\max}) + (1-c)\ell_\mu(h,x,n+1) \\
&= \begin{cases} \frac{1}{1-\mu}\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})}\right]^{1-\mu} - 1\right) + (1-c)\frac{1}{1-\mu}\left(\left[\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,n+1)}\right]^{1-\mu} - 1\right) & \mu \neq 1 \\ \log\left(\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,y_{\max})}\right) + (1-c)\log\left(\sum_{y'\in\mathcal{Y}\cup\{n+1\}} e^{h(x,y')-h(x,n+1)}\right) & \mu = 1. \end{cases} \\
&= \begin{cases} \frac{1}{1-\mu}\left(s_h(x,y_{\max})^{\mu-1} - 1\right) + (1-c)\frac{1}{1-\mu}\left([s_h(x,n+1)]^{\mu-1} - 1\right) & \mu \neq 1 \\ -\log(s_h(x,y_{\max})) - (1-c)\log(s_h(x,n+1)) & \mu = 1. \end{cases}
\end{aligned}$$

Since $0 \leq s_h(x,y_{\max}) + s_h(x,n+1) \leq 1$, by taking the partial derivative, we obtain that the minimum can be attained by

$$\begin{cases} s^*_h(x,y_{\max}) = \frac{1}{1+(1-c)^{\frac{1}{2-\mu}}} \text{ and } s^*_h(x,n+1) = \frac{(1-c)^{\frac{1}{2-\mu}}}{1+(1-c)^{\frac{1}{2-\mu}}} & \mu \neq 2 \\ s^*_h(x,y_{\max}) = 1 \text{ and } s^*_h(x,n+1) = 0 & \mu = 2. \end{cases} \tag{5}$$

Since $\mathcal{H}$ is symmetric and complete, there exists $h \in \mathcal{H}$ such that (5) is achieved. Therefore,

$$\begin{aligned}
\inf_{h\in\mathcal{H}} \mathcal{C}_{\mathsf{L}_\mu}(\mathcal{H},x) &= \begin{cases} \frac{1}{1-\mu}\left(s^*_h(x,y_{\max})^{\mu-1} - 1\right) + (1-c)\frac{1}{1-\mu}\left([s^*_h(x,n+1)]^{\mu-1} - 1\right) & \mu \neq 1 \\ -\log(s^*_h(x,y_{\max})) - (1-c)\log(s^*_h(x,n+1)) & \mu = 1 \end{cases} \\
&= \begin{cases} \frac{1}{1-\mu}\left[\left[1+(1-c)^{\frac{1}{2-\mu}}\right]^{2-\mu} - (2-c)\right] & \mu \notin \{1,2\} \\ -\log\left(\frac{1}{2-c}\right) - (1-c)\log\left(\frac{1-c}{2-c}\right) & \mu = 1 \\ 1-c & \mu = 2. \end{cases}
\end{aligned}$$

Since $\inf_{h\in\mathcal{H}} \mathcal{C}_{\mathsf{L}_\mu}(\mathcal{H},x)$ is independent of x, we obtain that $\mathbb{E}_X\left[\inf_{h\in\mathcal{H}} \mathcal{C}_{\mathsf{L}_\mu}(\mathcal{H},x)\right] = \inf_{h\in\mathcal{H}} \mathcal{C}_{\mathsf{L}_\mu}(\mathcal{H},x)$, which completes the proof. □

C.3 Proof of General Transformation of $\mathcal{H}$-Consistency Bounds (Theorem 3)

Theorem 3. *Assume that ℓ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1} with a concave function Γ, that is, for all $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})).$$

Then, L *defined by* (2) *admits an* $\mathcal{H}$*-consistency bound with respect to* L_{abs} *with the functional form* $(2-c)\Gamma(\frac{t}{2-c})$*, that is, for all* $h \in \mathcal{H}$*, we have*

$$\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \leq (2-c)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{2-c}\right).$$

Proof. By Lemma 5, the calibration gap of L_{abs} can be expressed and upper-bounded as follows:

$$\begin{aligned}
&\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)\\
&= \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H},x)\\
&= \max_{y\in\mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\\
&= (2-c)\left(\max_{y\in\mathsf{H}(x)} \overline{p}(x,y) - \overline{p}(x,\mathsf{h}(x))\right) && (\text{Let } \overline{p}(x,y) = \tfrac{p(x,y)}{2-c}\mathbb{1}_{y\in\mathcal{Y}} + \tfrac{1-c}{2-c}\mathbb{1}_{y=n+1})\\
&= (2-c)\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && (\text{By (Awasthi et al., 2022b, Lemma 3)})\\
&\leq (2-c)\Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)) && (\text{By } \mathcal{H}\text{-consistency bound of } \ell)\\
&= (2-c)\Gamma\left(\sum_{y\in\mathcal{Y}\bigcup\{n+1\}} \overline{p}(x,y)\ell(h,x,y) - \inf_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}\bigcup\{n+1\}} \overline{p}(x,y)\ell(h,x,y)\right)\\
&= (2-c)\Gamma\left(\sum_{y\in\mathcal{Y}} \frac{p(x,y)}{2-c}\ell(h,x,y) + \frac{1-c}{2-c}\ell(h,x,n+1) - \inf_{h\in\mathcal{H}}\left(\sum_{y\in\mathcal{Y}} \frac{p(x,y)}{2-c}\ell(h,x,y) + \frac{1-c}{2-c}\ell(h,x,n+1)\right)\right)\\
& && (\text{Plug in } \overline{p}(x,y) = \tfrac{p(x,y)}{2-c}\mathbb{1}_{y\in\mathcal{Y}} + \tfrac{1-c}{2-c}\mathbb{1}_{y=n+1})\\
&= (2-c)\Gamma\left(\frac{1}{2-c}\left[\sum_{y\in\mathcal{Y}} p(x,y)\mathsf{L}(h,x,y) - \inf_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} p(x,y)\mathsf{L}(h,x,y)\right]\right)\\
&= (2-c)\Gamma\left(\frac{1}{2-c}\Delta\mathcal{C}_{\mathsf{L},\mathcal{H}}(h,x)\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) &= \mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\mathsf{L}_{\text{abs}},\mathcal{H}}(h,x)]\\
&\leq \mathop{\mathbb{E}}_{X}\left[(2-c)\Gamma\left(\frac{1}{2-c}\Delta\mathcal{C}_{\mathsf{L},\mathcal{H}}(h,x)\right)\right]\\
&\leq (2-c)\Gamma\left(\frac{1}{2-c}\mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\mathsf{L},\mathcal{H}}(h,x)]\right) && (\Gamma \text{ is concave})\\
&= (2-c)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{2-c}\right),
\end{aligned}$$

which completes the proof. □

C.4 Proof of $\mathcal{H}$-Consistency Bounds for Two-Stage Surrogates (Theorem 4)

Theorem 4 (**$\mathcal{H}$-consistency bounds for two-stage surrogates**). *Given a hypothesis set* $\mathcal{H} = \mathcal{H}_{\mathcal{Y}} \times \mathcal{H}_{n+1}$*. Assume that* ℓ *admits an* $\mathcal{H}_{\mathcal{Y}}$*-consistency bound with respect to the multi-class zero-one classification loss* ℓ_{0-1} *and that* Φ *admits an* $\mathcal{H}^{\tau}_{n+1}$*-consistency bound with respect to the binary zero-one classification loss* $\ell^{\text{binary}}_{0-1}$ *for any* $\tau \in \mathbb{R}$*. Thus, there are non-decreasing concave functions* Γ_1 *and* Γ_2 *such that, for all* $h_{\mathcal{Y}} \in \mathcal{H}_{\mathcal{Y}}$*,* $h^{\tau}_{n+1} \in \mathcal{H}^{\tau}_{n+1}$ *and* $\tau \in \mathbb{R}$*, we have*

$$\begin{aligned}
&\mathcal{E}_{\ell_{0-1}}(h_{\mathcal{Y}}) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}_{\mathcal{Y}}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}_{\mathcal{Y}})\\
&\quad\leq \Gamma_1(\mathcal{E}_{\ell}(h_{\mathcal{Y}}) - \mathcal{E}^*_{\ell}(\mathcal{H}_{\mathcal{Y}}) + \mathcal{M}_{\ell}(\mathcal{H}_{\mathcal{Y}}))\\
&\mathcal{E}_{\ell^{\text{binary}}_{0-1}}(h^{\tau}_{n+1}) - \mathcal{E}^*_{\ell^{\text{binary}}_{0-1}}(\mathcal{H}^{\tau}_{n+1}) + \mathcal{M}_{\ell^{\text{binary}}_{0-1}}(\mathcal{H}^{\tau}_{n+1})\\
&\quad\leq \Gamma_2(\mathcal{E}_{\Phi}(h^{\tau}_{n+1}) - \mathcal{E}^*_{\Phi}(\mathcal{H}^{\tau}_{n+1}) + \mathcal{M}_{\Phi}(\mathcal{H}^{\tau}_{n+1})).
\end{aligned}$$

Then, the following holds for all $h = (h_y, h_{n+1}) \in \mathcal{H}$:

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \\
&\leq \Gamma_1(\mathcal{E}_\ell(h_y) - \mathcal{E}^*_\ell(\mathcal{H}_y) + \mathcal{M}_\ell(\mathcal{H}_y)) \\
&+ (1+c)\Gamma_2\left(\frac{\mathcal{E}_{\ell_{h_y}}(h_{n+1}) - \mathcal{E}^*_{\ell_{h_y}}(\mathcal{H}_{n+1}) + \mathcal{M}_{\ell_{h_y}}(\mathcal{H}_{n+1})}{c}\right),
\end{aligned}$$

where the constant factors $(1+c)$ and $\frac{1}{c}$ can be removed when Γ_2 is linear.

Proof. For any $h = (h_y, h_{n+1})$, we can rewrite $\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H})$ as

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\text{abs}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\text{abs}}}(\mathcal{H}) \\
&= \mathop{\mathbb{E}}_X\left[\mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H},x)\right] \\
&= \mathop{\mathbb{E}}_X\left[\mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) + \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H},x)\right] \\
&= \mathop{\mathbb{E}}_X\left[\mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x)\right] + \mathop{\mathbb{E}}_X\left[\inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\text{abs}}}(\mathcal{H},x)\right]
\end{aligned} \tag{6}$$

By the assumptions, we have

$$\begin{aligned}
&\mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) - \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\text{abs}}}(h,x) \\
&= \sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1} - \inf_{h_{n+1}\in\mathcal{H}_{n+1}}\left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1}\right) \\
&= \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c\right)\times\left[\eta(x)\ell^{\text{binary}}_{0\text{-}1}\left(h_{n+1} - \max_{y\in\mathcal{Y}} h_y(x,y), x, +1\right) + (1-\eta(x))\ell^{\text{binary}}_{0\text{-}1}\left(h_{n+1} - \max_{y\in\mathcal{Y}} h_y(x,y), x, -1\right)\right. \\
&\quad \left. - \inf_{h_{n+1}\in\mathcal{H}_{n+1}}\left(\eta(x)\ell^{\text{binary}}_{0\text{-}1}\left(h_{n+1} - \max_{y\in\mathcal{Y}} h_y(x,y), x, +1\right) + (1-\eta(x))\ell^{\text{binary}}_{0\text{-}1}\left(h_{n+1} - \max_{y\in\mathcal{Y}} h_y(x,y), x, -1\right)\right)\right] \\
&\qquad\qquad \left(\text{Let } \eta(x) = \tfrac{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y}}{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c}\right) \\
&\leq \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c\right)\Gamma_2\left[\eta(x)\Phi\left(h_{n+1}(x) - \max_{y\in\mathcal{Y}} h_y(x,y)\right) + (1-\eta(x))\Phi\left(\max_{y\in\mathcal{Y}} h_y(x,y) - h_{n+1}(x)\right)\right. \\
&\quad \left. - \inf_{h_{n+1}\in\mathcal{H}_{n+1}}\left(\eta(x)\Phi\left(h_{n+1}(x) - \max_{y\in\mathcal{Y}} h_y(x,y)\right) + (1-\eta(x))\Phi\left(\max_{y\in\mathcal{Y}} h_y(x,y) - h_{n+1}(x)\right)\right)\right] \\
&\qquad\qquad \left(\text{By } \mathcal{H}^\tau_{n+1}\text{-consistency bounds of } \Phi \text{ under assumption, } \tau = \max_{y\in\mathcal{Y}} h_y(x,y)\right) \\
&= \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c\right)\Gamma_2\left(\frac{\sum_{y\in\mathcal{Y}} p(x,y)\ell_{h_y}(h_{n+1},x,y) - \inf_{h_{n+1}\in\mathcal{H}_{n+1}}\sum_{y\in\mathcal{Y}} p(x,y)\ell_{h_y}(h_{n+1},x,y)}{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c}\right) \\
&\qquad\qquad \left(\eta(x) = \tfrac{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y}}{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c} \text{ and formulation (4)}\right) \\
&= \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c\right)\Gamma_2\left(\frac{\mathcal{C}_{\ell_{h_y}}(h_{n+1},x) - \mathcal{C}^*_{\ell_{h_y}}(\mathcal{H}_{n+1},x)}{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c}\right) \\
&\leq \begin{cases}\Gamma_2\left(\mathcal{C}_{\ell_{h_y}}(h_{n+1},x) - \mathcal{C}^*_{\ell_{h_y}}(\mathcal{H}_{n+1},x)\right) & \text{when } \Gamma_2 \text{ is linear} \\ (1+c)\Gamma_2\left(\frac{\mathcal{C}_{\ell_{h_y}}(h_{n+1},x) - \mathcal{C}^*_{\ell_{h_y}}(\mathcal{H}_{n+1},x)}{c}\right) & \text{otherwise}\end{cases} \\
&\qquad\qquad \left(c \leq \sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_y(x)\neq y} + c \leq 1 + c \text{ and } \Gamma_2 \text{ is non-decreasing}\right) \\
&= \begin{cases}\Gamma_2\left(\Delta\mathcal{C}_{\ell_{h_y},\mathcal{H}_{n+1}}(h_{n+1},x)\right) & \text{when } \Gamma_2 \text{ is linear} \\ (1+c)\Gamma_2\left(\frac{\Delta\mathcal{C}_{\ell_{h_y},\mathcal{H}_{n+1}}(h_{n+1},x)}{c}\right) & \text{otherwise}\end{cases}
\end{aligned}$$

and

$$
\begin{aligned}
&\inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\rm abs}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\rm abs}}(\mathcal{H},x)\\
&= \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\rm abs}}(h,x) - \inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}},h_{n+1}\in\mathcal{H}_{n+1}} \mathcal{C}_{\mathsf{L}_{\rm abs}}(h,x)\\
&= \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1}\right) - \inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}},h_{n+1}\in\mathcal{H}_{n+1}}\left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1}\right)\\
&= \inf_{h_{n+1}\in\mathcal{H}_{n+1}} \left(\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1}\right) - \inf_{h_{n+1}\in\mathcal{H}_{n+1}}\left(\inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}}\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\neq n+1} + c\mathbb{1}_{\mathsf{h}(x)=n+1}\right)\\
&= \min\left\{\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}, c\right\} - \min\left\{\inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}}\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}, c\right\}\\
&\leq \sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y} - \inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}}\sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}_{\mathcal{Y}}(x)\neq y}\\
&= \mathcal{C}_{\ell_{0-1}}(h_{\mathcal{Y}},x) - \mathcal{C}^*_{\ell_{0-1}}(\mathcal{H}_{\mathcal{Y}},x)\\
&= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x)\\
&\leq \Gamma_1(\Delta\mathcal{C}_{\ell,\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x)). && \text{(By } \mathcal{H}_{\mathcal{Y}}\text{-consistency bounds of } \ell \text{ under assumption)}
\end{aligned}
$$

Therefore, by (6), we obtain

$$
\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\rm abs}}(h) - \mathcal{E}^*_{\mathsf{L}_{\rm abs}}(\mathcal{H}_{\mathcal{Y}}) + \mathcal{M}_{\mathsf{L}_{\rm abs}}(\mathcal{H}_{\mathcal{Y}})\\
&\leq \begin{cases} \mathbb{E}_X\left[\Gamma_2\left(\Delta\mathcal{C}_{\ell_{h_{\mathcal{Y}}},\mathcal{H}_{n+1}}(h_{n+1},x)\right)\right] + \mathbb{E}_X[\Gamma_1(\Delta\mathcal{C}_{\ell,\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x))] & \text{when } \Gamma_2 \text{ is linear}\\ (1+c)\,\mathbb{E}_X\left[\Gamma_2\left(\frac{\Delta\mathcal{C}_{\ell_{h_{\mathcal{Y}}},\mathcal{H}_{n+1}}(h_{n+1},x)}{c}\right)\right] + \mathbb{E}_X[\Gamma_1(\Delta\mathcal{C}_{\ell,\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x))] & \text{otherwise}\end{cases}\\
&\leq \begin{cases} \Gamma_2\left(\mathbb{E}_X\left[\Delta\mathcal{C}_{\ell_{h_{\mathcal{Y}}},\mathcal{H}_{n+1}}(h_{n+1},x)\right]\right) + \Gamma_1(\mathbb{E}_X[\Delta\mathcal{C}_{\ell,\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x)]) & \text{when } \Gamma_2 \text{ is linear}\\ (1+c)\Gamma_2\left(\frac{1}{c}\,\mathbb{E}_X\left[\Delta\mathcal{C}_{\ell_{h_{\mathcal{Y}}},\mathcal{H}_{n+1}}(h_{n+1},x)\right]\right) + \Gamma_1(\mathbb{E}_X[\Delta\mathcal{C}_{\ell,\mathcal{H}_{\mathcal{Y}}}(h_{\mathcal{Y}},x)]) & \text{otherwise}\end{cases}\\
& && (\Gamma_1 \text{ and } \Gamma_2 \text{ are concave})\\
&= \begin{cases} \Gamma_1(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}_{\mathcal{Y}}) + \mathcal{M}_\ell(\mathcal{H}_{\mathcal{Y}})) + \Gamma_2\left(\mathcal{E}_{\ell_{h_{\mathcal{Y}}}}(h_{n+1}) - \mathcal{E}^*_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1}) + \mathcal{M}_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1})\right) & \text{when } \Gamma_2 \text{ is linear}\\ (\Gamma_1(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}_{\mathcal{Y}}) + \mathcal{M}_\ell(\mathcal{H}_{\mathcal{Y}})) + (1+c)\Gamma_2\left(\frac{\mathcal{E}_{\ell_{h_{\mathcal{Y}}}}(h_{n+1}) - \mathcal{E}^*_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1}) + \mathcal{M}_{\ell_{h_{\mathcal{Y}}}}(\mathcal{H}_{n+1})}{c}\right) & \text{otherwise,}\end{cases}
\end{aligned}
$$

which completes the proof.

□

C.5 Proof of Realizable $\mathcal{H}$-Consistency for Two-Stage Surrogates (Theorem 8)

Definition 7 (Realizable $\mathcal{H}$-consistency). *Let $\hat{h}$ denote a hypothesis attaining the infimum of the expected surrogate loss, $\mathcal{E}_{\mathsf{L}}(\hat{h}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})$. A score-based abstention surrogate loss L is said to be* realizable $\mathcal{H}$-consistent *with respect to the abstention loss $\mathsf{L}_{\rm abs}$ if, for any distribution in which an optimal hypothesis h^* exists in $\mathcal{H}$ with an abstention loss of zero (i.e., $\mathcal{E}_{\mathsf{L}_{\rm abs}}(h^*) = 0$), we have $\mathcal{E}_{\mathsf{L}_{\rm abs}}(\hat{h}) = 0$.*

Next, we demonstrate that our proposed two-stage score-based surrogate losses are not only Bayes-consistent, as previously established in Section 4, but also realizable $\mathcal{H}$-consistent, which will be shown in Theorem 8. This effectively addresses the open question posed by Mozannar et al. (2023) in the context of score-based multi-class abstention and highlights the benefits of the two-stage formulation.

Theorem 8 (Realizable $\mathcal{H}$-consistency for two-stage surrogates). *Given a hypothesis set $\mathcal{H} = \mathcal{H}_{\mathcal{Y}} \times \mathcal{H}_{n+1}$ that is closed under scaling. Let Φ be a function that satisfies the condition $\lim_{t\to+\infty}\Phi(t) = 0$ and $\Phi(t) \geq 1_{t\leq 0}$ for any $t \in \mathbb{R}$. Assume that $\hat{h} = (\hat{h}_{\mathcal{Y}}, \hat{h}_{n+1}) \in \mathcal{H}$ attains the infimum of the expected surrogate loss, $\mathcal{E}_\ell(\hat{h}_{\mathcal{Y}}) = \inf_{h_{\mathcal{Y}}\in\mathcal{H}_{\mathcal{Y}}}\mathcal{E}_\ell(h_{\mathcal{Y}})$ and $\mathcal{E}_{\ell_{\hat{h}_{\mathcal{Y}}}}(\hat{h}_{n+1}) = \inf_{h\in\mathcal{H}}\mathcal{E}_{\ell_{h_{\mathcal{Y}}}}(h_{n+1})$. Then, for any distribution in which an optimal hypothesis $h^* = (h^*_{\mathcal{Y}}, h^*_{n+1})$ exists in $\mathcal{H}$ with $\mathcal{E}_{\mathsf{L}_{\rm abs}}(h^*) = 0$, we have $\mathcal{E}_{\mathsf{L}_{\rm abs}}(\hat{h}) = 0$.*

Proof. By the assumptions, ℓ_{h_y} serves as an upper bound for L_{abs} and thus $\mathcal{E}_{\mathsf{L}_{\text{abs}}}(\hat{h}) \leq \mathcal{E}_{\ell_{\hat{h}_y}}(\hat{h}_{n+1})$. If abstention happens, that is $h^*_{n+1}(x) > \max_{y\in\mathcal{Y}} h^*_y(x,y)$ for some point x, then we must have $c = 0$ by the realizability assumption. Therefore, there exists an optimal h^{**} such that $h^{**}_{n+1}(x) > \max_{y\in\mathcal{Y}} h^{**}_y(x,y)$ for all $x \in \mathcal{X}$ without incurring any cost. Then, by the Lebesgue dominated convergence theorem and the assumption that $\mathcal{H}$ is closed under scaling,

$$\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\text{abs}}}(\hat{h}) &\leq \mathcal{E}_{\ell_{\hat{h}_y}}(\hat{h}_{n+1})\\
&\leq \lim_{\alpha\to+\infty} \mathcal{E}_{\ell_{\alpha h^{**}_y}}(\alpha h^{**}_{n+1})\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\Big[\ell_{\alpha h^{**}_y}(\alpha h^{**}_{n+1}, x, y)\Big]\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\left[\mathbb{1}_{\mathsf{h}^{**}_y(x)\neq y}\Phi\left(\alpha\left(h^{**}_{n+1}(x) - \max_{y\in\mathcal{Y}} h^{**}_y(x,y)\right)\right) + c\Phi\left(\alpha\left(\max_{y\in\mathcal{Y}} h^{**}_y(x,y) - h^{**}_{n+1}(x)\right)\right)\right]\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\left[\mathbb{1}_{\mathsf{h}^{**}_y(x)\neq y}\Phi\left(\alpha\left(h^{**}_{n+1}(x) - \max_{y\in\mathcal{Y}} h^{**}_y(x,y)\right)\right)\right] && (c = 0)\\
&= 0. && (\text{using } \lim_{t\to+\infty}\Phi(t) = 0 \text{ and the Lebesgue dominated convergence theorem})
\end{aligned}$$

If abstention does not happen, that is $h^*_{n+1}(x) - \max_{y\in\mathcal{Y}} h^*_y(x,y) < 0$ for all $x \in \mathcal{X}$, then we must have $h^*_y(x,y) - \max_{y'\neq y} h^*_y(x,y') > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ by the realizability assumption. Then, by the Lebesgue dominated convergence theorem and the assumption that $\mathcal{H}$ is closed under scaling,

$$\begin{aligned}
\mathcal{E}_{\mathsf{L}_{\text{abs}}}(\hat{h}) &\leq \mathcal{E}_{\ell_{\hat{h}_y}}(\hat{h}_{n+1})\\
&\leq \lim_{\alpha\to+\infty} \mathcal{E}_{\ell_{\alpha h^{*}_y}}(\alpha h^{*}_{n+1})\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\Big[\ell_{\alpha h^{*}_y}(\alpha h^{*}_{n+1}, x, y)\Big]\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\left[\mathbb{1}_{\mathsf{h}^{*}_y(x)\neq y}\Phi\left(\alpha\left(h^{*}_{n+1}(x) - \max_{y\in\mathcal{Y}} h^{*}_y(x,y)\right)\right) + c\Phi\left(\alpha\left(\max_{y\in\mathcal{Y}} h^{*}_y(x,y) - h^{*}_{n+1}(x)\right)\right)\right]\\
&= \lim_{\alpha\to+\infty} \mathbb{E}\left[c\Phi\left(\alpha\left(\max_{y\in\mathcal{Y}} h^{*}_y(x,y) - h^{*}_{n+1}(x)\right)\right)\right] && (h^*_y(x,y) - \max_{y'\neq y} h^*_y(x,y') > 0)\\
&= 0. && (\text{using } \lim_{t\to+\infty}\Phi(t) = 0 \text{ and the Lebesgue dominated convergence theorem})
\end{aligned}$$

By combining the above two analysis, we conclude the proof. □

D Significance of $\mathcal{H}$-consistency bounds with minimizability gaps

As previously highlighted, the minimizabiliy gap can be upper bounded by the approximation error $\mathcal{A}_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\text{all}}} \mathbb{E}_y[\mathsf{L}(h,X,y) \mid X = x]\big] = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}_{\text{all}})$. However, it is a finer quantity than the approximation error, and as such, it can potentially provide more significant guarantees. To elaborate, as shown by (Awasthi et al., 2022a,b), for a target loss function L_2 and a surrogate loss function L_1, the excess error bound $\mathcal{E}_{\mathsf{L}_2}(h) - \mathcal{E}^*_{\mathsf{L}_2}(\mathcal{H}_{\text{all}}) \leq \Gamma\big(\mathcal{E}_{\mathsf{L}_1}(h) - \mathcal{E}^*_{\mathsf{L}_1}(\mathcal{H}_{\text{all}})\big)$ can be reformulated as

$$\mathcal{E}_{\mathsf{L}_2}(h) - \mathcal{E}^*_{\mathsf{L}_2}(\mathcal{H}) + \mathcal{A}_{\mathsf{L}_2}(\mathcal{H}) \leq \Gamma\big(\mathcal{E}_{\mathsf{L}_1}(h) - \mathcal{E}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{A}_{\mathsf{L}_1}(\mathcal{H})\big),$$

where Γ is typically linear or the square-root function modulo constants. On the other hand, an $\mathcal{H}$-consistency bound can be expressed as follows:

$$\mathcal{E}_{\mathsf{L}_2}(h) - \mathcal{E}^*_{\mathsf{L}_2}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_2}(\mathcal{H}) \leq \Gamma\big(\mathcal{E}_{\mathsf{L}_1}(h) - \mathcal{E}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_1}(\mathcal{H})\big).$$

For a target loss function L_2 with discrete outputs, such as the zero-one loss or the deferral loss, we have $\mathbb{E}_x\big[\inf_{h\in\mathcal{H}} \mathbb{E}_y[\mathsf{L}_2(h,x,y) \mid X = x]\big] = \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\text{all}}} \mathbb{E}_y[\mathsf{L}_2(h,x,y) \mid X = x]\big]$ when the hypothesis set generates labels that cover all possible outcomes for each input (See (Awasthi et al., 2022b, Lemma 3), Lemma 5 in Appendix C). Consequently, we have $\mathcal{M}_{\mathsf{L}_2}(\mathcal{H}) = \mathcal{A}_{\mathsf{L}_2}(\mathcal{H})$. However, for a surrogate loss function L_1, the minimizability gap is upper bounded by the approximation error, $\mathcal{M}_{\mathsf{L}_1}(\mathcal{H}) \leq \mathcal{A}_{\mathsf{L}_1}(\mathcal{H})$, and is generally finer.

Let us consider a straightforward binary classification example where the conditional distribution is denoted as $\eta(x) = D(Y = 1|X = x)$. We will define $\mathcal{H}$ as a set of functions h, such that $|h(x)| \leq \Lambda$ for all $x \in \mathcal{X}$, for some $\Lambda > 0$,

and it is also possible to achieve any value in the range $[-\Lambda, +\Lambda]$. For the exponential-based margin loss, which we define as $\mathsf{L}(h, x, y) = e^{-yh(x)}$, we obtain the following equation:

$$\mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, y) \mid X = x] = \eta(x)e^{-h(x)} + (1 - \eta(x))e^{h(x)}.$$

Upon observing this, it becomes apparent that the infimum over all measurable functions can be expressed in the following way, for all x:

$$\inf_{h \in \mathcal{H}_{\text{all}}} \mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, y) \mid X = x] = 2\sqrt{\eta(x)(1 - \eta(x))},$$

while the infimum over $\mathcal{H}$, $\inf_{h \in \mathcal{H}} \mathbb{E}_y[\mathsf{L}(h, x, y) \mid X = x]$, depends on Λ and can be expressed as

$$\inf_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, y) \mid X = x] = \begin{cases} \max\{\eta(x), 1 - \eta(x)\}e^{-\Lambda} + \min\{\eta(x), 1 - \eta(x)\}e^{\Lambda} & \Lambda < \frac{1}{2}\left|\log \frac{\eta(x)}{1-\eta(x)}\right| \\ 2\sqrt{\eta(x)(1 - \eta(x))} & \text{otherwise.} \end{cases}$$

Thus, in the deterministic scenario, the discrepancy between the approximation error $\mathcal{A}_{\mathsf{L}}(\mathcal{H})$ and the minimizability gap $\mathcal{M}_{\mathsf{L}}(\mathcal{H})$ is:

$$\mathcal{A}_{\mathsf{L}}(\mathcal{H}) - \mathcal{M}_{\mathsf{L}}(\mathcal{H}) = \mathop{\mathbb{E}}_{x}\left[\inf_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, y) \mid X = x] - \inf_{h \in \mathcal{H}_{\text{all}}} \mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, y) \mid X = x]\right] = e^{-\Lambda}.$$

Therefore, for a surrogate loss, the minimizability gap can be strictly less than the approximation error. In summary, an $\mathcal{H}$-consistency bound can be more significant than the excess error bound as $\mathcal{M}_{\mathsf{L}_2}(\mathcal{H}) = \mathcal{A}_{\mathsf{L}_2}(\mathcal{H})$ when L_2 represents the zero-one loss or deferral loss, and $\mathcal{M}_{\mathsf{L}_1}(\mathcal{H}) \leq \mathcal{A}_{\mathsf{L}_1}(\mathcal{H})$. They can also be directly used to derive finite sample estimation bounds for a surrogate loss minimizer, which are more favorable and relevant than a similar finite sample guarantee that could be derived from an excess error bound (see Section 6).