Principled Approaches for Learning to Defer with Multiple Experts

Anqi Mao[1], **Mehryar Mohri**[1,2], **Yutao Zhong**[1]

[1]Courant Institute
[2]Google Research
aqmao@cims.nyu.edu, mohri@google.com, yutao@cims.nyu.edu

Abstract

We present a study of surrogate losses and algorithms for the general problem of *learning to defer with multiple experts*. We first introduce a new family of surrogate losses specifically tailored for the multiple-expert setting, where the prediction and deferral functions are learned simultaneously. We then prove that these surrogate losses benefit from strong $\mathcal{H}$-consistency bounds. We illustrate the application of our analysis through several examples of practical surrogate losses, for which we give explicit guarantees. These loss functions readily lead to the design of new learning to defer algorithms based on their minimization. While the main focus of this work is a theoretical analysis, we also report the results of several experiments on SVHN and CIFAR-10 datasets.

1 Introduction

In many real-world applications, expert decisions can complement or significantly enhance existing models. These experts may consist of humans possessing domain expertise or more sophisticated albeit expensive models. For instance, contemporary language models and dialog-based text generation systems have exhibited susceptibility to generating erroneous information, often referred to as *hallucinations*. Thus, their response quality can be substantially improved by deferring uncertain predictions to more advanced or domain-specific pre-trained models. This particular issue has been recognized as a central challenge for large language models (LLMs) (Wei et al. 2022; Bubeck et al. 2023). Similar observations apply to other generation systems, including image or video generation, as well as learning models used in various applications such as image classification, image annotation, and speech recognition. Thus, the problem of *learning to defer with multiple experts* has become increasingly critical in applications.

The concept of *learning to defer* can be traced back to the original work on *learning with rejection* or *abstention* based on confidence thresholds (Chow 1957, 1970; Herbei and Wegkamp 2005; Bartlett and Wegkamp 2008; Grandvalet et al. 2008; Yuan and Wegkamp 2010, 2011; Ramaswamy, Tewari, and Agarwal 2018; Ni et al. 2019), rejection or abstention functions (Cortes, DeSalvo, and Mohri 2016b,a; Charoenphakdee et al. 2021; Cao et al. 2022), or *selective classification* (El-Yaniv et al. 2010; Wiener and El-Yaniv 2011; Geifman and El-Yaniv 2017; Gangrade, Kag, and Saligrama 2021; Geifman and El-Yaniv 2019), and other methods (Kalai, Kanade, and Mansour 2012; Ziyin et al. 2019; Acar, Gangrade, and Saligrama 2020). In these studies, either the cost of abstention is not explicit or it is chosen to be a constant.

However, a constant cost does not fully capture all the relevant information in the deferral scenario. It is important to take into account the quality of the expert, whose prediction we rely on. These may be human experts as in several critical applications (Kamar, Hacker, and Horvitz 2012; Tan et al. 2018; Kleinberg et al. 2018; Bansal et al. 2021). To address this gap, Madras et al. (2018) incorporated the human expert's decision into the cost and proposed the first *learning to defer (L2D)* framework, which has also been examined in (Raghu et al. 2019; Wilder, Horvitz, and Kamar 2021; Pradier et al. 2021; Keswani, Lease, and Kenthapadi 2021). Mozannar and Sontag (2020) proposed the first *Bayes-consistent* (Zhang 2004; Bartlett, Jordan, and McAuliffe 2006; Steinwart 2007) surrogate loss for L2D, and subsequent work (Raman and Yee 2021; Liu, Gallego, and Barbieri 2022) further improved upon it. Another Bayes-consistent surrogate loss in L2D is the one-versus-all loss proposed by Verma and Nalisnick (2022) that is also studied in (Charusaie et al. 2022) as a special case of a general family of loss functions. An additional line of research investigated post-hoc methods (Okati, De, and Rodriguez 2021; Narasimhan et al. 2022), where Okati, De, and Rodriguez (2021) proposed an alternative optimization method between the predictor and rejector, and Narasimhan et al. (2022) provided a correction to the surrogate losses in (Mozannar and Sontag 2020; Verma and Nalisnick 2022) when they are underfitting. Finally, L2D or its variants have been adopted or studied in various other scenarios (De et al. 2020; Straitouri et al. 2021; Zhao et al. 2021; Joshi, Parbhoo, and Doshi-Velez 2021; Gao et al. 2021; Mozannar, Satyanarayan, and Sontag 2022; Liu, Gallego, and Barbieri 2022; Hemmer et al. 2023; Narasimhan et al. 2023).

All the studies mentioned so far mainly focused on learning to defer with a single expert. Most recently, Verma, Barrejón, and Nalisnick (2023) highlighted the significance of *learning to defer with multiple experts* (Hemmer et al. 2022; Keswani, Lease, and Kenthapadi 2021; Kerrigan, Smyth, and Steyvers 2021; Straitouri et al. 2022; Benz and Rodriguez 2022) and extended the surrogate loss in (Verma and Nalisnick 2022; Mozannar and Sontag 2020) to accommodate the multiple-

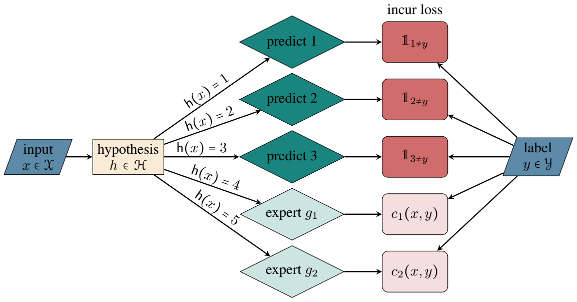

Figure 1: Illustration of the scenario of learning to defer with multiple experts ($n = 3$ and $n_e = 2$).

expert setting, which is currently the only work to propose Bayes-consistent surrogate losses in this scenario. They further showed that a mixture of experts (MoE) approach to multi-expert L2D proposed in (Hemmer et al. 2022) is not consistent.

Meanwhile, recent work by Awasthi et al. (2022a,b) introduced new consistency guarantees, called $\mathcal{H}$-consistency bounds, which they argued are more relevant to learning than Bayes-consistency since they are hypothesis set-specific and non-asymptotic. $\mathcal{H}$-consistency bounds are also stronger guarantees than Bayes-consistency. They established $\mathcal{H}$-consistent bounds for common surrogate losses in standard classification (see also (Zheng et al. 2023)). This naturally raises the question: can we design deferral surrogate losses that benefit from these more significant consistency guarantees?

Our contributions. We study the general framework of learning to defer with multiple experts. We first introduce a new family of surrogate losses specifically tailored for the multiple-expert setting, where the prediction and deferral functions are learned simultaneously (Section 3). Next, we prove that these surrogate losses benefit from $\mathcal{H}$-consistency bounds (Section 4). This implies, in particular, their Bayes-consistency. We illustrate the application of our analysis through several examples of practical surrogate losses, for which we give explicit guarantees. These loss functions readily lead to the design of new learning to defer algorithms based on their minimization. Our $\mathcal{H}$-consistency bounds incorporate a crucial term known as the *minimizability gap*. We show that this makes them more advantageous guarantees than bounds based on the approximation error (Section 5). We further demonstrate that our $\mathcal{H}$-consistency bounds can be used to derive generalization bounds for the minimizer of a surrogate loss expressed in terms of the minimizability gaps (Section 6). While the main focus of this work is a theoretical analysis, we also report the results of several experiments with SVHN and CIFAR-10 datasets (Section 7).

We give a more detailed discussion of related work in Appendix A. We start with the introduction of preliminary definitions and notation needed for our discussion of the problem of learning to defer with multiple experts.

2 Preliminaries

We consider the standard multi-class classification setting with an input space $\mathcal{X}$ and a set of $n \geq 2$ labels $\mathcal{Y} = [n]$, where we use the notation $[n]$ to denote the set $\{1, \ldots, n\}$. We study the scenario of *learning to defer with multiple experts*, where the label set $\mathcal{Y}$ is augmented with n_e additional labels $\{n+1, \ldots, n+n_e\}$ corresponding to n_e pre-defined experts $g_1, \ldots, g_{n_e}$, which are a series of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. In this scenario, the learner has the option of returning a label $y \in \mathcal{Y}$, which represents the category predicted, or a label $y = n + j$, $1 \leq j \leq n_e$, in which case it is *deferring* to expert g_j.

We denote by $\overline{\mathcal{Y}} = [n + n_e]$ the augmented label set and consider a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X} \times \overline{\mathcal{Y}}$ to $\mathbb{R}$. The prediction associated by $h \in \mathcal{H}$ to an input $x \in \mathcal{X}$ is denoted by $\mathsf{h}(x)$ and defined as the element in $\overline{\mathcal{Y}}$ with the highest score, $\mathsf{h}(x) = \operatorname{argmax}_{y \in [n+n_e]} h(x,y)$, with an arbitrary but fixed deterministic strategy for breaking ties. We denote by $\mathcal{H}_{\text{all}}$ the family of all measurable functions.

The *deferral loss function* L_{def} is defined as follows for any $h \in \mathcal{H}$ and $(x,y) \in \mathcal{X} \times \mathcal{Y}$:

$$\mathsf{L}_{\text{def}}(h,x,y) = \mathbb{1}_{\mathsf{h}(x) \neq y} \mathbb{1}_{\mathsf{h}(x) \in [n]} + \sum_{j=1}^{n_e} c_j(x,y) \mathbb{1}_{\mathsf{h}(x) = n+j} \tag{1}$$

Thus, the loss incurred coincides with the standard zero-one classification loss when $\mathsf{h}(x)$, the label predicted, is in $\mathcal{Y}$. Otherwise, when $\mathsf{h}(x)$ is equal to $n + j$, the loss incurred is $c_j(x,y)$, the cost of deferring to expert g_j. We give an illustration of the scenario of learning to defer with three classes and two experts ($n = 3$ and $n_e = 2$) in Figure 1. We will denote by $\underline{c}_j \geq 0$ and $\overline{c}_j \leq 1$ finite lower and upper bounds on the cost c_j, that is $c_j(x,y) \in [\underline{c}_j, \overline{c}_j]$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. There are many possible choices for these costs. Our analysis is general and requires no assumption other than their boundedness. One natural choice is to define cost c_j as a function relying on expert g_j's accuracy, for example $c_j(x,y) = \alpha_j \mathbb{1}_{\mathsf{g}_j(x) \neq y} + \beta_j$, with $\alpha_j, \beta_j > 0$, where $\mathsf{g}_j(x) = \operatorname{argmax}_{y \in [n]} g_j(x,y)$ is the prediction made by expert g_j for input x.

Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we will denote by $\mathcal{E}_{\mathsf{L}_{\text{def}}}(h)$ the expected deferral loss of a hypothesis $h \in \mathcal{H}$,

$$\mathcal{E}_{\mathsf{L}_{\text{def}}}(h) = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\mathsf{L}_{\text{def}}(h,x,y)], \tag{2}$$

and by $\mathcal{E}^*_{\mathsf{L}_{\text{def}}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{E}_{\mathsf{L}_{\text{def}}}(h)$ its infimum or best-in-class expected loss. We will adopt similar definitions for any surrogate loss function L:

$$\mathcal{E}_{\mathsf{L}}(h) = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\mathsf{L}(h,x,y)], \quad \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{E}_{\mathsf{L}}(h). \tag{3}$$

3 General surrogate losses

In this section, we introduce a new family of surrogate losses specifically tailored for the multiple-expert setting starting from first principles.

The scenario we consider is one where the prediction (first n scores) and deferral functions (last n_e scores) are learned simultaneously. Consider a hypothesis $h \in \mathcal{H}$. Note that, for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$, if the learner chooses to defer to an expert, $\mathsf{h}(x) \in \{n+1, \ldots, n+n_e\}$, then it does not make a prediction of the category, and thus $\mathsf{h}(x) \neq y$. This implies that the following identity holds:

$$\mathbb{1}_{\mathsf{h}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\in\{n+1,\ldots,n+n_e\}} = \mathbb{1}_{\mathsf{h}(x)\in\{n+1,\ldots,n+n_e\}}.$$

Using this identity and $\mathbb{1}_{\mathsf{h}(x)\in[n]} = 1 - \mathbb{1}_{\mathsf{h}(x)\in\{n+1,\ldots,n+n_e\}}$, we can write the first term of (1) as $\mathbb{1}_{\mathsf{h}(x)\neq y} - \mathbb{1}_{\mathsf{h}(x)\in\{n+1,\ldots,n+n_e\}}$. Note that deferring occurs if and only if one of the experts is selected, that is $\mathbb{1}_{\mathsf{h}(x)\in\{n+1,\ldots,n+n_e\}} = \sum_{j=1}^{n_e}\mathbb{1}_{\mathsf{h}(x)=n+j}$. Therefore, the deferral loss function can be written in the following form for any $h \in \mathcal{H}$ and $(x,y) \in \mathcal{X}\times\mathcal{Y}$:

$$\begin{aligned}
&\mathsf{L}_{\mathrm{def}}(h,x,y)\\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} - \sum_{j=1}^{n_e}\mathbb{1}_{\mathsf{h}(x)=n+j} + \sum_{j=1}^{n_e}c_j(x,y)\mathbb{1}_{\mathsf{h}(x)=n+j}\\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} + \sum_{j=1}^{n_e}(c_j(x,y)-1)\mathbb{1}_{\mathsf{h}(x)=n+j}\\
&= \mathbb{1}_{\mathsf{h}(x)\neq y} + \sum_{j=1}^{n_e}(1-c_j(x,y))\mathbb{1}_{\mathsf{h}(x)\neq n+j} + \sum_{j=1}^{n_e}(c_j(x,y)-1).
\end{aligned}$$

In light of this expression, since the last term $\sum_{j=1}^{n_e}(c_j(x,y)-1)$ does not depend on h, if ℓ is a surrogate loss for the zero-one multi-class classification loss over the augmented label set $\overline{\mathcal{Y}}$, then L, defined as follows for any $h\in\mathcal{H}$ and $(x,y)\in\mathcal{X}\times\mathcal{Y}$, is a natural surrogate loss for $\mathsf{L}_{\mathrm{def}}$:

$$\mathsf{L}(h,x,y) = \ell(h,x,y) + \sum_{j=1}^{n_e}(1-c_j(x,y))\,\ell(h,x,n+j). \tag{4}$$

We will study the properties of the general family of surrogate losses L thereby defined. Note that in the special case where ℓ is the logistic loss and $n_e = 1$, that is where there is only one pre-defined expert, L coincides with the surrogate loss proposed in (Mozannar and Sontag 2020; Cao et al. 2022). However, even for that special case, our derivation of the surrogate loss from first principle is new and it is this analysis that enables us to define a surrogate loss for the more general case of multiple experts and other ℓ loss functions. Our formulation also recovers the softmax surrogate loss in (Verma, Barrejón, and Nalisnick 2023) when $\ell = \ell_{\log}$ and $c_j(x,y) = 1_{\mathsf{g}_j(x)\neq y}$.

4 $\mathcal{H}$-consistency bounds for surrogate losses

Here, we prove strong consistency guarantees for a surrogate deferral loss L of the form described in the previous section, provided that the loss function ℓ it is based upon admits a similar consistency guarantee with respect to the standard zero-one classification loss.

$\mathcal{H}$-consistency bounds. To do so, we will adopt the notion of *$\mathcal{H}$-consistency bounds* recently introduced by Awasthi, Mao, Mohri, and Zhong (2022a,b). These are guarantees that, unlike Bayes-consistency or excess error bound, take into account the specific hypothesis set $\mathcal{H}$ and do not assume $\mathcal{H}$ to be the family of all measurable functions. Moreover, in contrast with Bayes-consistency, they are not just asymptotic guarantees. In this context, they have the following form: $\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq f(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))$, where f is a non-decreasing function, typically concave. Thus, when the surrogate estimation loss $(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))$ is reduced to ϵ, the deferral estimation loss $\left(\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})\right)$ is guaranteed to be at most $f(\epsilon)$.

Minimizability gaps. A key quantity appearing in these bounds is the *minimizability gap* $\mathcal{M}_\ell(\mathcal{H})$ which, for a loss function ℓ and hypothesis set $\mathcal{H}$, measures the difference of the best-in-class expected loss and the expected pointwise infimum of the loss:

$$\mathcal{M}_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}) - \mathbb{E}_x\Big[\inf_{h\in\mathcal{H}}\mathbb{E}_{y|x}[\ell(h,x,y)]\Big].$$

By the super-additivity of the infimum, since $\mathcal{E}^*_\ell(\mathcal{H}) = \inf_{h\in\mathcal{H}}\mathbb{E}_x\big[\mathbb{E}_{y|x}[\ell(h,x,y)]\big]$, the minimizability gap is always non-negative.

When the loss function ℓ only depends on $h(x,\cdot)$ for all h, x, and y, that is $\ell(h,x,y) = \Psi(h(x,1),\ldots,h(x,n),y)$, for some function Ψ, then it is not hard to show that the minimizability gap vanishes for the family of all measurable functions: $\mathcal{M}_\ell(\mathcal{H}_{\mathrm{all}}) = 0$ (Steinwart 2007)[lemma 2.5]. It is also null when $\mathcal{E}^*_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}_{\mathrm{all}})$, that is when the Bayes-error coincides with the best-in-class error. In general, however, the minimizabiliy gap is non-zero for a restricted hypothesis set $\mathcal{H}$ and is therefore important to analyze. In Section 5, we will discuss in more detail minimizability gaps for a relatively broad case and demonstrate that $\mathcal{H}$-consistency bounds with minimizability gaps can often be more favorable than excess error bounds based on the approximation error.

The following theorem is the main result of this section.

Theorem 1 (**$\mathcal{H}$-consistency bounds for score-based surrogates**). *Assume that ℓ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1}. Thus, there exists a non-decreasing concave function Γ with $\Gamma(0) = 0$ such that, for any distribution $\mathcal{D}$ and for all $h\in\mathcal{H}$, we have*

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

Then, L admits the following $\mathcal{H}$-consistency bound with respect to $\mathsf{L}_{\mathrm{def}}$: for all $h\in\mathcal{H}$,

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})\\
&\leq \left(n_e + 1 - \sum_{j=1}^{n_e}\underline{c}_j\right)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e}\overline{c}_j}\right).
\end{aligned} \tag{5}$$

Furthermore, constant factors $\left(n_e + 1 - \sum_{j=1}^{n_e}\underline{c}_j\right)$ and $\frac{1}{n_e+1-\sum_{j=1}^{n_e}\overline{c}_j}$ can be removed when Γ is linear.

The proof is given in Appendix C.4. It consists of first analyzing the conditional regret of the deferral loss and that of a surrogate loss. Next, we show how the former can be upper bounded in terms of the latter by leveraging the $\mathcal{H}$-consistency bound of ℓ with respect to the zero-one loss with an appropriate conditional distribution that we construct. This, combined with the results of Awasthi et al. (2022b), proves our $\mathcal{H}$-consistency bounds.

Let us emphasize that the theorem is broadly applicable and that there are many choices for the surrogate loss ℓ meeting the assumption of the theorem: Awasthi et al. (2022b) showed that a variety of surrogate loss functions ℓ admit an $\mathcal{H}$-consistency bound with respect to the zero-one loss for common hypothesis sets such as linear models and multi-layer neural networks, including *sum losses* (Weston and Watkins 1998), *constrained losses* (Lee, Lin, and Wahba 2004), and, as shown more recently by Mao, Mohri, and Zhong (2023) (see also (Zheng et al. 2023)), *comp-sum losses*, which include the logistic loss (Verhulst 1838, 1845; Berkson 1944, 1951), the *sum-exponential loss* and many other loss functions.

Thus, the theorem gives a strong guarantee for a broad family of surrogate losses L based upon such loss functions ℓ. The presence of the minimizability gaps in these bounds is important. In particular, while the minimizability gap can be upper bounded by the approximation error $\mathcal{A}_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathbb{E}_{y|x}[\ell(h,x,y)]\big] = \mathcal{E}^*_\ell(\mathcal{H}) - \mathcal{E}^*_\ell(\mathcal{H}_{\mathrm{all}})$, it is a finer quantity than the approximation error and can lead to more favorable guarantees.

Note that when the Bayes-error coincides with the best-in-class error, $\mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}_{\mathrm{all}})$, we have $\mathcal{M}_{\mathsf{L}}(\mathcal{H}) \leq \mathcal{A}_{\mathsf{L}}(\mathcal{H}) = 0$. This leads to the following corollary, using the non-negativity property of the minimizability gap.

Corollary 2. *Assume that ℓ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1}. Then, for all $h \in \mathcal{H}$ and any distribution such that $\mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) = \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}_{\mathrm{all}})$, the following bound holds:*

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right) \Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right),$$

Furthermore, constant factors $\big(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\big)$ and $\frac{1}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}$ can be removed when Γ is linear.

Thus, when the estimation error of the surrogate loss, $\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})$, is reduced to ϵ, the estimation error of the deferral loss, $\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})$, is upper bounded by

$$\left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right) \Gamma\left(\epsilon \Big/ \left(n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j\right)\right).$$

Moreover, $\mathcal{H}$-consistency holds since $\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) \to 0$ implies $\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \to 0$.

Table 1 shows several examples of surrogate deferral losses and their corresponding $\mathcal{H}$-consistency bounds, using the multi-class $\mathcal{H}$-consistency bounds known for comp-sum losses ℓ with respect to the zero-one loss (Mao, Mohri, and Zhong 2023, Theorem 1). The bounds have been simplified here using the inequalities $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$. See Appendix D.1 for a more detailed derivation.

Similarly, Table 2 and Table 3 show several examples of surrogate deferral losses with sum losses or constrained losses adopted for ℓ and their corresponding $\mathcal{H}$-consistency bounds, using the multi-class $\mathcal{H}$-consistency bounds in (Awasthi et al. 2022b, Table 2) and (Awasthi et al. 2022b, Table 3) respectively. Here too, we present the simplified bounds by using the inequalities $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$. See Appendix D.2 and Appendix D.3 for a more detailed derivation.

5 Benefits of minimizability gaps

As already pointed out, the minimizabiliy gap can be upper bounded by the approximation error $\mathcal{A}_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathbb{E}_{y|x}[\ell(h,x,y)]\big] = \mathcal{E}^*_\ell(\mathcal{H}) - \mathcal{E}^*_\ell(\mathcal{H}_{\mathrm{all}})$. It is however a finer quantity than the approximation error and can thus lead to more favorable guarantees. More precisely, 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 can be rewritten as

$$\begin{aligned}&\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{A}_{\ell_2}(\mathcal{H})\\ &\qquad\qquad \leq \Gamma\big(\mathcal{E}_{\ell_1}(h) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) + \mathcal{A}_{\ell_1}(\mathcal{H})\big),\end{aligned}$$

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:

$$\begin{aligned}&\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H})\\ &\qquad\qquad \leq \Gamma\big(\mathcal{E}_{\ell_1}(h) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})\big).\end{aligned}$$

For a target loss function ℓ_2 with discrete outputs, such as the zero-one loss or the deferral loss, we have $\mathbb{E}_x\big[\inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\ell_2(h,x,y)]\big] = \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathbb{E}_{y|x}[\ell_2(h,x,y)]\big]$ when the hypothesis set generates labels that cover all possible outcomes for each input (See (Awasthi et al. 2022b, Lemma 3), Lemma 4 in Appendix C.1). Consequently, we have $\mathcal{M}_{\ell_2}(\mathcal{H}) = \mathcal{A}_{\ell_2}(\mathcal{H})$. For a surrogate loss function ℓ_1, the minimizability gap is upper bounded by the approximation error, $\mathcal{M}_{\ell_1}(\mathcal{H}) \leq \mathcal{A}_{\ell_1}(\mathcal{H})$, and is generally finer.

Consider a simple binary classification example with the conditional distribution denoted as $\eta(x) = D(Y = 1 | X = x)$. Let $\mathcal{H}$ be a family of functions h such that $|h(x)| \leq \Lambda$ for all $x \in \mathcal{X}$, for some $\Lambda > 0$, and such that all values in the range $[-\Lambda, +\Lambda]$ can be achieved. For the exponential-based margin loss, defined as $\ell(h,x,y) = e^{-yh(x)}$, we have

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

It can be observed that the infimum over all measurable functions can be written as follows, for all x:

$$\inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathop{\mathbb{E}}_{y|x}[\ell(h,x,y)] = 2\sqrt{\eta(x)(1-\eta(x))},$$

while the infimum over $\mathcal{H}$, $\inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\ell(h,x,y)]$, depends on Λ. That infimum over $\mathcal{H}$ is achieved by

$$h(x) = \begin{cases}\min\Big\{\frac{1}{2}\log\frac{\eta(x)}{1-\eta(x)}, \Lambda\Big\} & \eta(x) \geq 1/2\\ \max\Big\{\frac{1}{2}\log\frac{\eta(x)}{1-\eta(x)}, -\Lambda\Big\} & \text{otherwise.}\end{cases}$$

Thus, in the deterministic case, we can explicitly compute the difference between the approximation error and the minimizability gap:

$$\begin{aligned}&\mathcal{A}_\ell(\mathcal{H}) - \mathcal{M}_\ell(\mathcal{H})\\ &\quad = \mathop{\mathbb{E}}_x\Big[\inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{y|x}[\ell(h,x,y)] - \inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathop{\mathbb{E}}_{y|x}[\ell(h,x,y)]\Big] = e^{-\Lambda}.\end{aligned}$$

ℓ	L	$\mathcal{H}$-consistency bounds
ℓ_{exp}	$\sum_{y' \neq y} e^{h(x,y') - h(x,y)} + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} e^{h(x,y') - h(x,n+j)}$	$\sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}$
ℓ_{log}	$-\log\left(\frac{e^{h(x,y)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}}\right) - \sum_{j=1}^{n_e} (1 - c_j(x,y)) \log\left(\frac{e^{h(x,n+j)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}}\right)$	$\sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}$
ℓ_{gce}	$\frac{1}{\alpha}\left[1 - \left[\frac{e^{h(x,y)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}}\right]^{\alpha}\right] + \frac{1}{\alpha} \sum_{j=1}^{n_e} (1 - c_j(x,y)) \left[1 - \left[\frac{e^{h(x,n+j)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}}\right]^{\alpha}\right]$	$\sqrt{2n^{\alpha}}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}$
ℓ_{mae}	$1 - \frac{e^{h(x,y)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}} + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \left(1 - \frac{e^{h(x,n+j)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}}\right)$	$n(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))$

Table 1: Examples of the deferral surrogate loss (4) with comp-sum losses adopted for ℓ and their associated $\mathcal{H}$-consistency bounds provided by Corollary 2 (with only the surrogate portion displayed).

ℓ	L	$\mathcal{H}$-consistency bounds
$\Phi_{\text{sq}}^{\text{sum}}$	$\sum_{y' \neq y} \Phi_{\text{sq}}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\text{sq}}(\Delta_h(x,n+j,y'))$	$(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}$
$\Phi_{\text{exp}}^{\text{sum}}$	$\sum_{y' \neq y} \Phi_{\text{exp}}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\text{exp}}(\Delta_h(x,n+j,y'))$	$\sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}$
Φ_{ρ}^{sum}	$\sum_{y' \neq y} \Phi_{\rho}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\rho}(\Delta_h(x,n+j,y'))$	$\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})$

Table 2: Examples of the deferral surrogate loss (4) with sum losses adopted for ℓ and their associated $\mathcal{H}$-consistency bounds provided by Corollary 2 (with only the surrogate portion displayed), where $\Delta_h(x,y,y') = h(x,y) - h(x,y')$, $\Phi_{\text{sq}}(t) = \max\{0, 1-t\}^2$, $\Phi_{\text{exp}}(t) = e^{-t}$, and $\Phi_{\rho}(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$.

As the parameter Λ decreases, the hypothesis set $\mathcal{H}$ becomes more restricted and the difference between the approximation error and the minimizability gap increases. In summary, an $\mathcal{H}$-consistency bound can be more favorable than the excess error bound as $\mathcal{M}_{\ell_2}(\mathcal{H}) = \mathcal{A}_{\ell_2}(\mathcal{H})$ when ℓ_2 represents the zero-one loss or deferral loss, and $\mathcal{M}_{\ell_1}(\mathcal{H}) \leq \mathcal{A}_{\ell_1}(\mathcal{H})$. Moreover, we will show in the next section that our $\mathcal{H}$-consistency bounds can lead to learning bounds for the deferral loss and a hypothesis set $\mathcal{H}$ with finite samples.

6 Learning bounds

For a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn from $\mathcal{D}^m$, we will denote by $\widehat{h}_S$ the empirical minimizer of the empirical loss within $\mathcal{H}$ with respect to the surrogate loss function L: $\widehat{h}_S = \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathsf{L}(h, x_i, y_i)$. Given an $\mathcal{H}$-consistency bound in the form of (5), we can further use it to derive a learning bound for the deferral loss by upper bounding the surrogate estimation error $\mathcal{E}_{\mathsf{L}}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})$ with the complexity (e.g. the Rademacher complexity) of the family of functions associated with L and $\mathcal{H}$: $\mathcal{H}_{\mathsf{L}} = \{(x, y) \mapsto \mathsf{L}(h, x, y) : h \in \mathcal{H}\}$.

We denote by $\mathfrak{R}_m^{\mathsf{L}}(\mathcal{H})$ the Rademacher complexity of $\mathcal{H}_{\mathsf{L}}$ and by B_{L} an upper bound of the surrogate loss L. Then, we obtain the following learning bound for the deferral loss based on (5).

Theorem 3 (**Learning bound**). *Under the same assumptions as Theorem 1, 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 deferral loss estimation bound holds for $\widehat{h}_S$:*

$$\mathcal{E}_{\mathsf{L}_{\text{def}}}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_{\text{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})$$
$$\leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right) \Gamma\left(\frac{4\mathfrak{R}_m^{\mathsf{L}}(\mathcal{H}) + 2B_{\mathsf{L}}\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right).$$

The proof is presented in Appendix E. To the best of our knowledge, Theorem 3 provides the first finite-sample guarantee for the estimation error of the minimizer of a surrogate deferral loss L defined for multiple experts. The proof exploits our $\mathcal{H}$-consistency bounds with respect to the deferral loss, as well as standard Rademacher complexity guarantees.

When $\overline{c}_j = 0$ and $\overline{c}_j = 1$ for any $j \in [n_e]$, the right-hand side of the bound admits the following simpler form:

$$(n_e + 1)\,\Gamma\left(4\mathfrak{R}_m^{\mathsf{L}}(\mathcal{H}) + 2B_{\mathsf{L}}\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}}(\mathcal{H})\right).$$

The dependency on the number of experts n_e makes this bound less favorable. There is a trade-off however since, on the other hand, more experts can help us achieve a better accuracy overall and reduce the best-in-class deferral loss. These learning bounds take into account the minimizability gap, which varies as a function of the upper bound Λ on the magnitude of the scoring functions. Thus, both the minimizability gaps and the Rademacher complexity term suggest a regularization controlling the complexity of the hypothesis set and the magnitude of the scores.

Adopting different loss functions ℓ in the definition of our deferral surrogate loss (4) will lead to a different functional form Γ, which can make the bound more or less favorable. For example, a linear form of Γ is in general more favorable than a square-root form modulo a constant. But, the dependency on the number of classes n appearing in Γ (e.g., $\ell = \ell_{\text{gce}}$ or $\ell = \ell_{\text{mae}}$) is also important to take into account since a larger value of n tends to negatively impact the guarantees. We already discussed the dependency on the number of experts n_e in Γ (e.g., $\ell = \ell_{\text{gce}}$ or $\ell = \ell_{\text{exp}}$) and the associated trade-off, which is also important to consider.

Note that the bound of Theorem 3 is expressed in terms of the global complexity of the prediction and deferral scoring functions $\mathcal{H}$. One can however derive a finer bound distinguishing the complexity of the deferral scoring functions and

ℓ	L	$\mathcal{H}$-consistency bounds
$\Phi_{\text{hinge}}^{\text{cstnd}}$	$\sum_{y' \neq y} \Phi_{\text{hinge}}(-h(x,y')) + \sum_{j=1}^{n_e}(1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\text{hinge}}(-h(x,y'))$	$\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}^*(\mathcal{H})$
$\Phi_{\text{sq}}^{\text{cstnd}}$	$\sum_{y' \neq y} \Phi_{\text{sq}}(-h(x,y')) + \sum_{j=1}^{n_e}(1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\text{sq}}(-h(x,y'))$	$(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}^*(\mathcal{H}))^{\frac{1}{2}}$
$\Phi_{\text{exp}}^{\text{cstnd}}$	$\sum_{y' \neq y} \Phi_{\text{exp}}(-h(x,y')) + \sum_{j=1}^{n_e}(1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\text{exp}}(-h(x,y'))$	$\sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}^*(\mathcal{H}))^{\frac{1}{2}}$
$\Phi_{\rho}^{\text{cstnd}}$	$\sum_{y' \neq y} \Phi_{\rho}(-h(x,y')) + \sum_{j=1}^{n_e}(1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\rho}(-h(x,y'))$	$\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}^*(\mathcal{H})$

Table 3: Examples of the deferral surrogate loss (4) with constrained losses adopted for ℓ and their associated $\mathcal{H}$-consistency bounds provided by Corollary 2 (with only the surrogate portion displayed), where $\Phi_{\text{hinge}}(t) = \max\{0, 1 - t\}$, $\Phi_{\text{sq}}(t) = \max\{0, 1 - t\}^2$, $\Phi_{\text{exp}}(t) = e^{-t}$, and $\Phi_{\rho}(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$ with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$.

that of the prediction scoring functions following a similar proof and analysis.

Recall that for a surrogate loss L, the minimizability gap $\mathcal{M}_{\mathsf{L}}(\mathcal{H})$ is in general finer than the approximation error $\mathcal{A}_{\mathsf{L}}(\mathcal{H})$, while for the deferral loss, for common hypothesis sets, these two quantities coincide. Thus, our bound can be rewritten as follows for common hypothesis sets:

$$\mathcal{E}_{\mathsf{L}_{\text{def}}}(\widehat{h}_S) - \mathcal{E}_{\mathsf{L}_{\text{def}}}^*(\mathcal{H}_{\text{all}}) \\ \leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right) \Gamma\left(\frac{4\mathfrak{R}_m^{\mathsf{L}}(\mathcal{H}) + 2B_{\mathsf{L}}\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right).$$

This is more favorable and more relevant than a similar excess loss bound where $\mathcal{M}_{\mathsf{L}}(\mathcal{H})$ is replaced with $\mathcal{A}_{\mathsf{L}}(\mathcal{H})$, which could be derived from a generalization bound for the surrogate loss.

7 Experiments

In this section, we examine the empirical performance of our proposed surrogate loss in the scenario of learning to defer with multiple experts. More specifically, we aim to compare the overall system accuracy for the learned predictor and deferral pairs, considering varying numbers of experts. This comparison provides valuable insights into the performance of our algorithm under different expert configurations. We explore three different scenarios:

- Only a single expert is available, specifically where a larger model than the base model is chosen as the deferral option.
- Two experts are available, consisting of one small model and one large model as the deferral options.
- Three experts are available, including one small model, one medium model, and one large model as the deferral options.

By comparing these scenarios, we evaluate the impact of varying the number and type of experts on the overall system accuracy.

Type of cost. We carried out experiments with two types of cost functions. For the first type, we selected the cost function to be exactly the misclassification error of the expert: $c_j(x,y) = \mathbb{1}_{\mathsf{g}_j(x) \neq y}$, where $\mathsf{g}_j(x) = \operatorname{argmax}_{y \in [n]} g_j(x,y)$ is the prediction made by expert g_j for input x. In this scenario, the cost incurred for deferring is determined solely based on the expert's accuracy. For the second type, we chose a cost function admitting the form $c_j(x,y) = \mathbb{1}_{\mathsf{g}_j(x) \neq y} + \beta_j$, where an additional non-zero base cost β_j is assigned to each expert. Deferring to a larger model then tends to incur a higher inference cost and hence, the corresponding β_j value for a larger model is higher as well. In addition to the base cost, each expert also incurs a misclassification error, as with the first type. Experimental setup and additional experiments (see Table 6) are included in Appendix B.

	Single expert	Two experts	Three experts
SVHN	92.08 ± 0.15%	93.18 ± 0.18%	93.46 ± 0.12%
CIFAR-10	73.31 ± 0.21%	77.12 ± 0.34%	78.71 ± 0.43%

Table 4: Overall system accuracy with the first type of cost functions.

	Single expert	Two experts	Three experts
SVHN	92.36 ± 0.22%	93.23 ± 0.21%	93.36 ± 0.11%
CIFAR-10	73.70 ± 0.40%	76.29 ± 0.41%	76.43 ± 0.55%

Table 5: Overall system accuracy with the second type of cost functions.

Experimental Results. In Table 4 and Table 5, we report the mean and standard deviation of the system accuracy over three runs with different random seeds. We noticed a positive correlation between the number of experts and the overall system accuracy. Specifically, as the number of experts increases, the performance of the system in terms of accuracy improves. This observation suggests that incorporating multiple experts in the learning to defer framework can lead to better predictions and decision-making. The results also demonstrate the effectiveness of our proposed surrogate loss for deferral with multiple experts.

8 Conclusion

We presented a comprehensive study of surrogate losses for the core challenge of learning to defer with multiple experts. Through our study, we established theoretical guarantees, strongly endorsing the adoption of the loss function family we introduced. This versatile family of loss functions can effectively facilitate the learning to defer algorithms across a wide range of applications. Our analysis offers great flexibility by accommodating diverse cost functions, encouraging exploration and evaluation of various options in real-world scenarios. We encourage further research into the theoretical properties of different choices and their impact on the overall performance to gain deeper insights into their effectiveness.

A Related work

The concept of *learning to defer* has its roots in research on abstention, particularly in binary classification scenarios with a constant cost function. Early work by (Chow 1957) and Chow (1970) focused on rejection and set the foundation for subsequent studies on learning with abstention. These studies explored different approaches such as *confidence-based methods* (Herbei and Wegkamp 2005; Bartlett and Wegkamp 2008; Grandvalet et al. 2008; Yuan and Wegkamp 2010), the *predictor-rejector framework* (Cortes, DeSalvo, and Mohri 2016b,a), or *selective classification* (El-Yaniv et al. 2010; Yuan and Wegkamp 2011; Wiener and El-Yaniv 2011)

Cortes, DeSalvo, and Mohri (2016b,a) showed that the confidence-based approach could fail to determine the optimal rejection region when the predictor did not match the Bayes solution. Instead, they proposed a novel *predictor-rejector* framework, for which they gave both Bayes-consistent and *realizable $\mathcal{H}$-consistent* surrogate losses (Long and Servedio 2013; Kuznetsov, Mohri, and Syed 2014; Zhang and Agarwal 2020), which achieve state-of-the-art performance in the binary setting.

El-Yaniv et al. (2010); Wiener and El-Yaniv (2011) introduced and studied a selective classification based on a predictor and a selector and explored the trade-off between classifier coverage and accuracy, drawing connections to active learning in their analysis.

The confidence-based and predictor-rejector frameworks have been both further analyzed in the context of *multi-class classification*. Ramaswamy, Tewari, and Agarwal (2018); Ni et al. (2019); Geifman and El-Yaniv (2017); Acar, Gangrade, and Saligrama (2020); Gangrade, Kag, and Saligrama (2021) extended the confidence-based method to multi-class settings, while Ni et al. (2019) noted that deriving a Bayes-consistent surrogate loss under the *predictor-rejector* framework is quite challenging and left it as an open problem. In response to this challenge, Mozannar and Sontag (2020) formulated a different *score-based* approach to learn the predictor and rejector simultaneously, by introducing an additional scoring function corresponding to rejection. This method has been further explored in a subsequent work (Cao et al. 2022). The surrogate losses derived under this framework are currently the state-of-the-art Mozannar and Sontag (2020); Cao et al. (2022).

Geifman and El-Yaniv (2019) proposed a new neural network architecture for abstention in the selective classification framework for multi-class classification. They did not derive consistent surrogate losses for this formulation. Ziyin et al. (2019) defined a loss function for the predictor-selector framework based on the doubling rate of gambling that requires almost no modification to the model architecture.

Another line of research studied multi-class abstention using an *implicit criterion* (Kalai, Kanade, and Mansour 2012; Acar, Gangrade, and Saligrama 2020; Gangrade, Kag, and Saligrama 2021; Charoenphakdee et al. 2021), by directly modeling regions with high confidence.

However, a constant cost does not fully capture all the relevant information in the deferral scenario. It is important to take into account the quality of the expert, whose prediction we rely on. These may be human experts as in several critical applications (Kamar, Hacker, and Horvitz 2012; Tan et al. 2018; Kleinberg et al. 2018; Bansal et al. 2021). To address this gap, Madras et al. (2018) incorporated the human expert's decision into the cost and proposed the first *learning to defer (L2D)* framework, which has also been examined in (Raghu et al. 2019; Wilder, Horvitz, and Kamar 2021; Pradier et al. 2021; Keswani, Lease, and Kenthapadi 2021). Mozannar and Sontag (2020) proposed the first *Bayes-consistent* (Zhang 2004; Bartlett, Jordan, and McAuliffe 2006; Steinwart 2007) surrogate loss for L2D, and subsequent work (Raman and Yee 2021; Liu, Gallego, and Barbieri 2022) further improved upon it. Another Bayes-consistent surrogate loss in L2D is the one-versus-all loss proposed by Verma and Nalisnick (2022) that is also studied in (Charusaie et al. 2022) as a special case of a general family of loss functions. An additional line of research investigated post-hoc methods (Okati, De, and Rodriguez 2021; Narasimhan et al. 2022), where Okati, De, and Rodriguez (2021) proposed an alternative optimization method between the predictor and rejector, and Narasimhan et al. (2022) provided a correction to the surrogate losses in (Mozannar and Sontag 2020; Verma and Nalisnick 2022) when they are underfitting. Finally, L2D or its variants have been adopted or studied in various other scenarios (De et al. 2020; Straitouri et al. 2021; Zhao et al. 2021; Joshi, Parbhoo, and Doshi-Velez 2021; Gao et al. 2021; Mozannar, Satyanarayan, and Sontag 2022; Liu, Gallego, and Barbieri 2022; Hemmer et al. 2023; Narasimhan et al. 2023).

All the studies mentioned so far mainly focused on learning to defer with a single expert. Most recently, Verma, Barrejón, and Nalisnick (2023) highlighted the significance of *learning to defer with multiple experts* (Hemmer et al. 2022; Keswani, Lease, and Kenthapadi 2021; Kerrigan, Smyth, and Steyvers 2021; Straitouri et al. 2022; Benz and Rodriguez 2022) and extended the surrogate loss in (Verma and Nalisnick 2022; Mozannar and Sontag 2020) to accommodate the multiple-expert setting, which is currently the only work to propose Bayes-consistent surrogate losses in this scenario. They further showed that a mixture of experts (MoE) approach to multi-expert L2D proposed in (Hemmer et al. 2022) is not consistent.

Meanwhile, recent work by Awasthi et al. (2022a,b) introduced new consistency guarantees, called $\mathcal{H}$-consistency bounds, which they argued are more relevant to learning than Bayes-consistency since they are hypothesis set-specific and non-asymptotic. $\mathcal{H}$-consistency bounds are also stronger guarantees than Bayes-consistency. They established $\mathcal{H}$-consistent bounds for common surrogate losses in standard classification (see also (Zheng et al. 2023)).

In this work, we study the general framework of learning to defer with multiple experts. Furthermore, we design deferral surrogate losses that benefit from these more significant consistency guarantees, namely, $\mathcal{H}$-consistency bounds, in the general multiple-expert setting.

B Experimental details

Experimental setup. For our experiments, we used two popular datasets: CIFAR-10 (Krizhevsky 2009) and SVHN (Street View House Numbers) (Netzer et al. 2011). CIFAR-10 consists of 60,000 color images in 10 different classes, with 6,000 images per class. The dataset is split into 50,000 training images and 10,000 test images. SVHN contains images of house numbers captured from Google Street View. It consists of 73,257 images for training and 26,032 images for testing. We trained for 50 epochs on CIFAR-10 and 15 epochs on SVHN without any data augmentation.

In our experiments, we adopted the ResNet (He et al. 2016) architecture for the base model and selected various sizes of ResNet models as experts in each scenario. Throughout all three scenarios, we used ResNet-4 for both the predictor and the deferral models. In the first scenario, we chose ResNet-10 as the expert model. In the second scenario, we included ResNet-10 and ResNet-16 as expert models. The third scenario involves ResNet-10, ResNet-16, and ResNet-28 as expert models with increasing complexity. The expert models are pre-trained on the training data of SVHN and CIFAR-10 respectively.

During the training process, we simultaneously trained the predictor ResNet-4 and the deferral model ResNet-4. We adopted the Adam optimizer (Kingma and Ba 2014) with a batch size of 128 and a weight decay of 1×10^{-4}. We used our proposed deferral surrogate loss (4) with the generalized cross-entropy loss being adopted for ℓ. As suggested by Zhang and Sabuncu (2018), we set the parameter α to 0.7.

For the second type of cost functions, we set the base costs as follows: $\beta_1 = 0.1$, $\beta_2 = 0.12$ and $\beta_3 = 0.14$ for the SVHN dataset and $\beta_1 = 0.3$, $\beta_2 = 0.32$, $\beta_3 = 0.34$ for the CIFAR-10 dataset, where β_1 corresponds to the cost associated with the smallest expert model, ResNet-10, β_2 to that of the medium model, ResNet-16, and β_3 to that of the largest expert model, ResNet-28. A base cost value that is not too far from the misclassification error of expert models encourages in practice a reasonable amount of input instances to be deferred. We observed that the performance remains close for other neighboring values of base costs.

Additional experiments. Here, we share additional experimental results in an intriguing setting where multiple experts are available and each of them has a clear domain of expertise. We report below the empirical results of our proposed deferral surrogate loss and the one-vs-all (OvA) surrogate loss proposed in recent work (Verma, Barrejón, and Nalisnick 2023), which is the state-of-the-art surrogate loss for learning to defer with multiple experts, on CIFAR-10. In this setting, the two experts have a clear domain of expertise. The expert 1 is always correct on the first three classes, 0 to 2, and predicts uniformly at random for other classes; the expert 2 is always correct on the next three classes, 3 to 5, and generates random predictions otherwise. We train a ResNet-16 for the predictor/deferral model.

As shown in Table 6, our method achieves comparable system accuracy with OvA. Among the images in classes 0 to 2, only 3.57% is deferred to expert 2 which predicts uniformly at random. Similarly, among the images in classes 3 to 5, only 3.33% is deferred to expert 1. For the rest of the images in classes 6 to 9, the predictor decides to learn to classify them by itself and actually makes 92.88% of the final predictions. This illustrates that our proposed surrogate loss is effective and comparable to the baseline.

C Proof of $\mathcal{H}$-consistency bounds for deferral surrogate losses

To prove $\mathcal{H}$-consistency bounds for our deferral surrogate loss functions, we will show how the *conditional regret* of the deferral loss can be upper bounded in terms of the *conditional regret* of the surrogate loss. The general theorems proven by Awasthi et al. (2022b, Theorem 4, Theorem 5) then guarantee our $\mathcal{H}$-consistency bounds.

For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let $p(x, y)$ denote the conditional probability of $Y = y$ given $X = x$ for any $y \in \mathcal{Y}$. Then, for any $x \in \mathcal{X}$, the *conditional* $\mathsf{L}_{\mathrm{def}}$*-loss* $\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x)$ and *conditional regret* (or *calibration gap*) $\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x)$ of a hypothesis $h \in \mathcal{H}$ are defined by

$$\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x) = \mathop{\mathbb{E}}_{y|x}[\mathsf{L}_{\mathrm{def}}(h, x, y)] = \sum_{y \in \mathcal{Y}} p(x, y)\mathsf{L}_{\mathrm{def}}(h, x, y)$$
$$\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x) = \mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x) - \mathcal{C}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}, x),$$

where $\mathcal{C}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}, x) = \inf_{h \in \mathcal{H}} \mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x)$. Similar definitions hold for the surrogate loss L. To bound $\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h, x)$ in terms of $\Delta\mathcal{C}_{\mathsf{L}}(h, x)$, we first give more explicit expressions for these conditional regrets.

To do so, it will be convenient to use the following definition for any $x \in \mathcal{X}$ and $y \in [n + n_e]$:

$$q(x, y) = \begin{cases} p(x, y) & y \in \mathcal{Y} \\ 1 - \sum_{y \in \mathcal{Y}} p(x, y) c_j(x, y) & n + 1 \leq y \leq n + n_e. \end{cases}$$

Note that $q(x, y)$ is non-negative but, in general, these quantities do not sum to one. We denote by $\overline{q}(x, y) = \frac{q(x,y)}{Q}$ their normalized counterparts which represent probabilities, where $Q = \sum_{y \in [n+n_e]} q(x, y)$.

For any $x \in \mathcal{X}$, we will denote by $\mathsf{H}(x)$ the set of labels generated by hypotheses in $\mathcal{H}$: $\mathsf{H}(x) = \{\mathsf{h}(x) \colon h \in \mathcal{H}\}$. We denote by $y_{\max} \in [n + n_e]$ the label associated by q to an input $x \in \mathcal{X}$, defined as $y_{\max} = \operatorname{argmax}_{y \in [n+n_e]} q(x, y)$, with the same deterministic strategy for breaking ties as that of $\mathsf{h}(x)$.

C.1 Conditional regret of the deferral loss

With these definitions, we can now express the conditional loss and regret of the deferral loss.

Lemma 4. *For any $x \in \mathcal{X}$, the minimal conditional $\mathsf{L}_{\mathrm{def}}$-loss and the calibration gap for $\mathsf{L}_{\mathrm{def}}$ can be expressed as follows:*

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

Proof. The conditional $\mathsf{L}_{\mathrm{def}}$-risk of h can be expressed as

Method	System accuracy (%)	Ratio of deferral (%)											
		all the classes			classes 0 to 2			classes 3 to 5			classes 6 to 9		
		predictor	expert 1	expert 2	predictor	expert 1	expert 2	predictor	expert 1	expert 2	predictor	expert 1	expert 2
Ours	92.19	61.43	17.38	21.19	46.77	49.67	3.57	33.60	3.33	63.07	92.88	3.43	3.70
OvA	91.39	59.72	16.78	23.50	48.63	47.67	3.70	27.87	2.47	69.67	92.73	3.50	3.78

Table 6: Comparison of our proposed deferral surrogate loss with the one-vs-all (OvA) surrogate loss in an intriguing setting where multiple experts are available and each of them has a clear domain of expertise.

follows:

$$
\begin{aligned}
&\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h,x)\\
&= \mathop{\mathbb{E}}_{y|x}\left[\mathsf{L}_{\mathrm{def}}(h,x,y)\right]\\
&= \mathop{\mathbb{E}}_{y|x}\left[\mathbb{1}_{\mathsf{h}(x)\neq y}\right]\mathbb{1}_{\mathsf{h}(x)\in[n]} + \sum_{j=1}^{n_e}\mathop{\mathbb{E}}_{y|x}\left[c_j(x,y)\right]\mathbb{1}_{\mathsf{h}(x)=n+j}\\
&= \sum_{y\in\mathcal{Y}} q(x,y)\mathbb{1}_{\mathsf{h}(x)\neq y}\mathbb{1}_{\mathsf{h}(x)\in[n]} + \sum_{j=1}^{n_e}(1-q(x,n+j))\mathbb{1}_{\mathsf{h}(x)=n+j}\\
&= (1-q(x,\mathsf{h}(x)))\mathbb{1}_{\mathsf{h}(x)\in[n]} + \sum_{j=1}^{n_e}(1-q(x,\mathsf{h}(x)))\mathbb{1}_{\mathsf{h}(x)=n+j}\\
&= 1-q(x,\mathsf{h}(x)).
\end{aligned}
$$

Then, the minimal conditional $\mathsf{L}_{\mathrm{def}}$-risk is given by

$$\mathcal{C}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H},x) = 1 - \max_{y\in\mathsf{H}(x)} q(x,y),$$

and the calibration gap can be expressed as follows:

$$
\begin{aligned}
\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}},\mathcal{H}}(h,x) &= \mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h,x) - \mathcal{C}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H},x)\\
&= \max_{y\in\mathsf{H}(x)} q(x,y) - q(x,\mathsf{h}(x)),
\end{aligned}
$$

which completes the proof. □

C.2 Conditional regret of a surrogate deferral loss

Lemma 5. *For any $x\in\mathcal{X}$, the conditional surrogate L-loss and regret can be expressed as follows:*

$$
\begin{aligned}
\mathcal{C}_{\mathsf{L}}(h,x) &= \sum_{y\in[n+n_e]} q(x,y)\ell(h,x,y)\\
\Delta\mathcal{C}_{\mathsf{L}}(h,x) &= \sum_{y\in[n+n_e]} q(x,y)\ell(h,x,y)\\
&\quad - \inf_{h\in\mathcal{H}} \sum_{y\in[n+n_e]} q(x,y)\ell(h,x,y).
\end{aligned}
$$

Proof. By definition, $\mathcal{C}_{\mathsf{L}}(h,x)$ is the conditional-L loss can be expressed as follows:

$$
\begin{aligned}
&\mathcal{C}_{\mathsf{L}}(h,x)\\
&= \mathop{\mathbb{E}}_{y}\left[\mathsf{L}(h,x,y)\right]\\
&= \mathop{\mathbb{E}}_{y}\left[\ell(h,x,y)\right] + \sum_{j=1}^{n_e}\mathop{\mathbb{E}}_{y|x}\left[(1-c_j(x,y))\right]\ell(h,x,n+j) \qquad (6)\\
&= \sum_{y\in\mathcal{Y}} q(x,y)\ell(h,x,y) + \sum_{j=1}^{n_e} q(x,n+j)\ell(h,x,n+j)\\
&= \sum_{y\in[n+n_e]} q(x,y)\ell(h,x,y),
\end{aligned}
$$

which ends the proof. □

C.3 Conditional regret of zero-one loss

We will also make use of the following result for the zero-one loss $\ell_{0-1}(h,x,y) = \mathbb{1}_{\mathsf{h}(x)\neq y}$ with label space $[n+n_e]$ and the conditional probability vector $\overline{q}(x,\cdot)$, which characterizes the minimal conditional ℓ_{0-1}-loss and the corresponding calibration gap (Awasthi et al. 2022b, Lemma 3).

Lemma 6. *For any $x\in\mathcal{X}$, the minimal conditional ℓ_{0-1}-loss and the calibration gap for ℓ_{0-1} can be expressed as follows:*

$$
\begin{aligned}
\mathcal{C}^*_{\ell_{0-1}}(x) &= 1 - \max_{y\in\mathsf{H}(x)} \overline{q}(x,y)\\
\Delta\mathcal{C}_{\ell_{0-1}}(h,x) &= \max_{y\in\mathsf{H}(x)} \overline{q}(x,y) - \overline{q}(x,\mathsf{h}(x)).
\end{aligned}
$$

C.4 Proof of $\mathcal{H}$-consistency bounds for deferral surrogate losses (Theorem 1)

Theorem 1 (**$\mathcal{H}$-consistency bounds for score-based surrogates**)**.** *Assume that ℓ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss ℓ_{0-1}. Thus, there exists a non-decreasing concave function Γ with $\Gamma(0)=0$ such that, for any distribution $\mathcal{D}$ and for all $h\in\mathcal{H}$, we have*

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

Then, L admits the following $\mathcal{H}$-consistency bound with respect to $\mathsf{L}_{\mathrm{def}}$: for all $h\in\mathcal{H}$,

$$
\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})\\
&\leq \left(n_e + 1 - \sum_{j=1}^{n_e}\underline{c}_j\right)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e}\overline{c}_j}\right). \qquad (5)
\end{aligned}
$$

Furthermore, constant factors $\left(n_e + 1 - \sum_{j=1}^{n_e}\underline{c}_j\right)$ and $\frac{1}{n_e+1-\sum_{j=1}^{n_e}\overline{c}_j}$ can be removed when Γ is linear.

Proof. We denote the normalization factor as $Q = \sum_{y\in[n+n_e]} q(x,y) = n_e + 1 - \mathbb{E}_y[c_j(x,y)]$, which is a constant that ensures the sum of $\overline{q}(x,y) = \frac{q(x,y)}{Q}$ is equal to 1. By Lemma 4, the calibration gap of $\mathsf{L}_{\mathrm{def}}$ can be expressed and upper-bounded as follows:

$$\begin{aligned}
&\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h,x)\\
&= \max_{y\in\mathsf{H}(x)} q(x,y) - q(x,\mathsf{h}(x)) && \text{(Lemma 4)}\\
&= Q\left(\max_{y\in\mathsf{H}(x)} \overline{q}(x,y) - \overline{q}(x,\mathsf{h}(x))\right)\\
&= Q\Delta\mathcal{C}_{\ell_{0-1}}(h,x) && \text{(Lemma 6)}\\
&\leq Q\Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)) && \text{($\mathcal{H}$-consistency bound of ℓ)}\\
&= Q\Gamma\left(\sum_{y\in[n+n_e]} \overline{q}(x,y)\ell(h,x,y)\right.\\
&\qquad \left. - \inf_{h\in\mathcal{H}} \sum_{y\in[n+n_e]} \overline{q}(x,y)\ell(h,x,y)\right)\\
&= Q\Gamma\left(\sum_{y\in[n+n_e]} \frac{q(x,y)}{Q}\ell(h,x,y)\right.\\
&\qquad \left. - \inf_{h\in\mathcal{H}} \sum_{y\in[n+n_e]} \frac{q(x,y)}{Q}\ell(h,x,y)\right)\\
&= Q\Gamma\left(\frac{1}{Q}\Delta\mathcal{C}_{\mathsf{L}}(h,x)\right). && \text{(Lemma 5)}
\end{aligned}$$

Thus, taking expectations gives:

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})\\
&= \mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\mathsf{L}_{\mathrm{def}}}(h,x)]\\
&\leq \mathop{\mathbb{E}}_{X}\left[Q\Gamma\left(\frac{1}{Q}\Delta\mathcal{C}_{\mathsf{L}}(h,x)\right)\right]\\
&\leq Q\Gamma\left(\frac{1}{Q}\mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\mathsf{L}}(h,x)]\right)\\
&\qquad\qquad \text{(concavity of Γ and Jensen's ineq.)}\\
&= Q\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{Q}\right)\\
&= \left(n_e + 1 - \mathop{\mathbb{E}}_{y}[c_j(x,y)]\right)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \mathbb{E}_y[c_j(x,y)]}\right)\\
&\leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right)\Gamma\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right)\\
&\qquad\qquad (\underline{c}_j \leq c_j(x,y) \leq \overline{c}_j, \forall j \in [n_e])
\end{aligned}$$

and $\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \Gamma(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H}))$ when Γ is linear, which completes the proof. □

D Examples of deferral surrogate losses and their $\mathcal{H}$-consistency bounds

D.1 ℓ being adopted as comp-sum losses

Example: $\ell = \ell_{\exp}$. Plug in $\ell = \ell_{\exp} = \sum_{y'\neq y} e^{h(x,y')-h(x,y)}$ in (4), we obtain

$$\begin{aligned}
\mathsf{L} &= \sum_{y'\neq y} e^{h(x,y')-h(x,y)}\\
&\qquad + \sum_{j=1}^{n_e}(1 - c_j(x,y)) \sum_{y'\neq n+j} e^{h(x,y')-h(x,n+j)}.
\end{aligned}$$

By Mao, Mohri, and Zhong (2023, Theorem 1), $\ell_{\exp}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{2t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2}\left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right)\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right)^{\frac{1}{2}}.$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \ell_{\log}$. Plug in $\ell = \ell_{\log} = -\log\left[\frac{e^{h(x,y)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right]$ in (4), we obtain

$$\begin{aligned}
\mathsf{L} &= -\log\left(\frac{e^{h(x,y)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right)\\
&\qquad - \sum_{j=1}^{n_e}(1 - c_j(x,y)) \log\left(\frac{e^{h(x,n+j)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right).
\end{aligned}$$

By Mao, Mohri, and Zhong (2023, Theorem 1), $\ell_{\log}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{2t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2}\left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right)\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right)^{\frac{1}{2}}.$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \ell_{\mathrm{gce}}$. Plug in $\ell = \ell_{\mathrm{gce}} == \frac{1}{\alpha}\left[1 - \left[\frac{e^{h(x,y)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right]^{\alpha}\right]$ in (4), we obtain

$$\begin{aligned}
\mathsf{L} &= \frac{1}{\alpha}\left[1 - \left[\frac{e^{h(x,y)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right]^{\alpha}\right]\\
&\qquad + \frac{1}{\alpha}\sum_{j=1}^{n_e}(1 - c_j(x,y))\left[1 - \left[\frac{e^{h(x,n+j)}}{\sum_{y'\in\overline{\mathcal{Y}}} e^{h(x,y')}}\right]^{\alpha}\right].
\end{aligned}$$

By Mao, Mohri, and Zhong (2023, Theorem 1), ℓ_{gce} admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{2n^{\alpha}t}$, using Corollary 2, we obtain

$$\begin{aligned}
&\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H})\\
&\qquad \leq \sqrt{2n^{\alpha}}\left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right)\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right)^{\frac{1}{2}}.
\end{aligned}$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2n^{\alpha}}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \ell_{\mathrm{mae}}$. Plug in $\ell = \ell_{\mathrm{mae}} = 1 - \frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}$ in (4), we obtain

$$\mathsf{L} = 1 - \frac{e^{h(x,y)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}} + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \left(1 - \frac{e^{h(x,n+j)}}{\sum_{y' \in \overline{\mathcal{Y}}} e^{h(x,y')}} \right).$$

By Mao, Mohri, and Zhong (2023, Theorem 1), ℓ_{mae} admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = nt$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq n(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})).$$

D.2 ℓ being adopted as sum losses

Example: $\ell = \Phi^{\mathrm{sum}}_{\mathrm{sq}}$. Plug in $\ell = \Phi^{\mathrm{sum}}_{\mathrm{sq}} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}(h(x,y) - h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\mathrm{sq}}(\Delta_h(x, n+j, y')),$$

where $\Delta_h(x,y,y') = h(x,y) - h(x,y')$ and $\Phi_{\mathrm{sq}}(t) = \max\{0, 1-t\}^2$. By Awasthi et al. (2022b, Table 2), $\Phi^{\mathrm{sum}}_{\mathrm{sq}}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \right) \left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j} \right)^{\frac{1}{2}}.$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq (n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \Phi^{\mathrm{sum}}_{\mathrm{exp}}$. Plug in $\ell = \Phi^{\mathrm{sum}}_{\mathrm{exp}} = \sum_{y' \neq y} \Phi_{\mathrm{exp}}(h(x,y) - h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\mathrm{exp}}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\mathrm{exp}}(\Delta_h(x, n+j, y')),$$

where $\Delta_h(x,y,y') = h(x,y) - h(x,y')$ and $\Phi_{\mathrm{exp}}(t) = e^{-t}$. By Awasthi et al. (2022b, Table 2), $\Phi^{\mathrm{sum}}_{\mathrm{exp}}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{2t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2} \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \right) \left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j} \right)^{\frac{1}{2}}.$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \Phi^{\mathrm{sum}}_{\rho}$. Plug in $\ell = \Phi^{\mathrm{sum}}_{\rho} = \sum_{y' \neq y} \Phi_{\rho}(h(x,y) - h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\rho}(\Delta_h(x,y,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\rho}(\Delta_h(x, n+j, y')),$$

where $\Delta_h(x,y,y') = h(x,y) - h(x,y')$ and $\Phi_{\rho}(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$. By Awasthi et al. (2022b, Table 2), $\Phi^{\mathrm{sum}}_{\rho}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = t$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}).$$

D.3 ℓ being adopted as constrained losses

Example: $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{hinge}}$. Plug in $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{hinge}} = \sum_{y' \neq y} \Phi_{\mathrm{hinge}}(-h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\mathrm{hinge}}(-h(x,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\mathrm{hinge}}(-h(x,y')),$$

where $\Phi_{\mathrm{hinge}}(t) = \max\{0, 1-t\}$ with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$. By Awasthi et al. (2022b, Table 3), $\Phi^{\mathrm{cstnd}}_{\mathrm{hinge}}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = t$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}).$$

Example: $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{sq}}$. Plug in $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{sq}} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}(-h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}(-h(x,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\mathrm{sq}}(-h(x,y')),$$

where $\Phi_{\mathrm{sq}}(t) = \max\{0, 1-t\}^2$ with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$. By Awasthi et al. (2022b, Table 3), $\Phi^{\mathrm{cstnd}}_{\mathrm{sq}}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \right) \left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j} \right)^{\frac{1}{2}}.$$

Since $1 \leq n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \leq n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \leq n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \leq (n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{exp}}$. Plug in $\ell = \Phi^{\mathrm{cstnd}}_{\mathrm{exp}} = \sum_{y' \neq y} \Phi_{\mathrm{exp}}(-h(x,y'))$ in (4), we obtain

$$\mathsf{L} = \sum_{y' \neq y} \Phi_{\mathrm{exp}}(-h(x,y')) + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_{\mathrm{exp}}(-h(x,y')),$$

where $\Phi_{\exp}(t) = e^{-t}$ with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$. By Awasthi et al. (2022b, Table 3), $\Phi^{\mathrm{cstnd}}_{\exp}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = \sqrt{2t}$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \le \sqrt{2}\left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right)\left(\frac{\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right)^{\frac{1}{2}}.$$

Since $1 \le n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j \le n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j \le n_e + 1$, the bound can be simplified as

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \le \sqrt{2}(n_e + 1)(\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}))^{\frac{1}{2}}.$$

Example: $\ell = \Phi^{\mathrm{cstnd}}_{\rho}$**.** Plug in $\ell = \Phi^{\mathrm{cstnd}}_{\rho} = \sum_{y' \neq y} \Phi_\rho(-h(x,y'))$ in (4), we obtain

$$\begin{aligned} \mathsf{L} &= \sum_{y' \neq y} \Phi_\rho(-h(x,y')) \\ &\quad + \sum_{j=1}^{n_e} (1 - c_j(x,y)) \sum_{y' \neq n+j} \Phi_\rho(-h(x,y')), \end{aligned}$$

where $\Phi_\rho(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$ with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$. By Awasthi et al. (2022b, Table 3), $\Phi^{\mathrm{cstnd}}_{\rho}$ admits an $\mathcal{H}$-consistency bound with respect to ℓ_{0-1} with $\Gamma(t) = t$, using Corollary 2, we obtain

$$\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) \le \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}^*_{\mathsf{L}}(\mathcal{H}).$$

E Proof of learning bounds for deferral surrogate losses (Theorem 3)

Theorem 3 (**Learning bound**). *Under the same assumptions as Theorem 1, 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 deferral loss estimation bound holds for $\widehat{h}_S$:*

$$\begin{aligned} &\mathcal{E}_{\mathsf{L}_{\mathrm{def}}}(\widehat{h}_S) - \mathcal{E}^*_{\mathsf{L}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H}) \\ &\le \left(n_e + 1 - \sum_{j=1}^{n_e} \underline{c}_j\right) \Gamma\left(\frac{4\mathfrak{R}^{\mathsf{L}}_m(\mathcal{H}) + 2B_{\mathsf{L}}\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \mathcal{M}_{\mathsf{L}}(\mathcal{H})}{n_e + 1 - \sum_{j=1}^{n_e} \overline{c}_j}\right). \end{aligned}$$

Proof. By using the standard Rademacher complexity bounds (Mohri, Rostamizadeh, and Talwalkar 2018), for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

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

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

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

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

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

Plugging in this inequality in the bound (5) completes the proof. □

References

Acar, D. A. E.; Gangrade, A.; and Saligrama, V. 2020. Budget learning via bracketing. In *International Conference on Artificial Intelligence and Statistics*, 4109–4119.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2022a. $\mathcal{H}$-Consistency Bounds for Surrogate Loss Minimizers. In *International Conference on Machine Learning*.

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

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

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473): 138–156.

Bartlett, P. L.; and Wegkamp, M. H. 2008. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research*, 9(8).

Benz, N. L. C.; and Rodriguez, M. G. 2022. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, 453–463. PMLR.

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

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

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cao, Y.; Cai, T.; Feng, L.; Gu, L.; Gu, J.; An, B.; Niu, G.; and Sugiyama, M. 2022. Generalizing Consistent Multi-Class Classification with Rejection to be Compatible with Arbitrary Losses. In *Advances in neural information processing systems*.

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

Charusaie, M.-A.; Mozannar, H.; Sontag, D.; and Samadi, S. 2022. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, 2972–3005.

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

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

Cortes, C.; DeSalvo, G.; and Mohri, M. 2016a. Boosting with Abstention. In *Advances in Neural Information Processing Systems*, 1660–1668.

Cortes, C.; DeSalvo, G.; and Mohri, M. 2016b. Learning with Rejection. In *International Conference on Algorithmic Learning Theory*, 67–82.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2611–2620.

El-Yaniv, R.; et al. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(5).

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

Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021. Human-ai collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614*.

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

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

Grandvalet, Y.; Rakotomamonjy, A.; Keshet, J.; and Canu, S. 2008. Support vector machines with a reject option. In *Advances in neural information processing systems*.

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

Hemmer, P.; Schellhammer, S.; Vössing, M.; Jakubik, J.; and Satzger, G. 2022. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. *arXiv preprint arXiv:2206.07948*.

Hemmer, P.; Thede, L.; Vössing, M.; Jakubik, J.; and Kühl, N. 2023. Learning to Defer with Limited Expert Predictions. *arXiv preprint arXiv:2304.07306*.

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

Joshi, S.; Parbhoo, S.; and Doshi-Velez, F. 2021. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*.

Kalai, A. T.; Kanade, V.; and Mansour, Y. 2012. Reliable agnostic learning. *Journal of Computer and System Sciences*, 78(5): 1481–1495.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, 467–474.

Kerrigan, G.; Smyth, P.; and Steyvers, M. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34: 4421–4434.

Keswani, V.; Lease, M.; and Kenthapadi, K. 2021. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 154–165.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293.

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

Kuznetsov, V.; Mohri, M.; and Syed, U. 2014. Multi-Class Deep Boosting. In *Advances in Neural Information Processing Systems*, 2501–2509.

Lee, Y.; Lin, Y.; and Wahba, G. 2004. 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.

Liu, J.; Gallego, B.; and Barbieri, S. 2022. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific reports*, 12(1): 1762.

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

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In *International Conference on Machine Learning*.

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

Mozannar, H.; Satyanarayan, A.; and Sontag, D. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5323–5331.

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

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

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

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *Advances in Neural Information Processing Systems*.

Ni, C.; Charoenphakdee, N.; Honda, J.; and Sugiyama, M. 2019. On the Calibration of Multiclass Classification with Rejection. In *Advances in Neural Information Processing Systems*, 2582–2592.

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

Pradier, M. F.; Zazo, J.; Parbhoo, S.; Perlis, R. H.; Zazzi, M.; and Doshi-Velez, F. 2021. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Summits on Translational Science Proceedings*, 2021: 525.

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

Raman, N.; and Yee, M. 2021. Improving Learning-to-Defer Algorithms Through Fine-Tuning. *arXiv preprint arXiv:2112.10768*.

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

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

Straitouri, E.; Singla, A.; Meresht, V. B.; and Gomez-Rodriguez, M. 2021. Reinforcement learning under algorithmic triage. *arXiv preprint arXiv:2109.11328*.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2022. Provably improving expert predictions with conformal prediction. *arXiv preprint arXiv:2201.12006*.

Tan, S.; Adebayo, J.; Inkpen, K.; and Kamar, E. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123*.

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

Verhulst, P. F. 1845. 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.

Verma, R.; Barrejón, D.; and Nalisnick, E. 2023. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In *International Conference on Artificial Intelligence and Statistics*, 11415–11434.

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

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. *CoRR*, abs/2206.07682.

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

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

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

Yuan, M.; and Wegkamp, M. 2010. Classification Methods with Reject Option Based on Convex Risk Minimization. *Journal of Machine Learning Research*, 11(1).

Yuan, M.; and Wegkamp, M. 2011. SVMs with a Reject Option. In *Bernoulli*.

Zhang, M.; and Agarwal, S. 2020. Bayes Consistency vs. H-Consistency: The Interplay between Surrogate Loss Functions and the Scoring Function Class. In *Advances in Neural Information Processing Systems*.

Zhang, T. 2004. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1): 56–85.

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

Zhao, J.; Agrawal, M.; Razavi, P.; and Sontag, D. 2021. Directing human attention in event localization for clinical timeline creation. In *Machine Learning for Healthcare Conference*, 80–102.

Zheng, C.; Wu, G.; Bao, F.; Cao, Y.; Li, C.; and Zhu, J. 2023. Revisiting Discriminative vs. Generative Classifiers: Theory and Implications. *arXiv preprint arXiv:2302.02334*.

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