Regression with Multi-Expert Deferral

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

Abstract

Learning to defer with multiple experts is a framework where the learner can choose to defer the prediction to several experts. While this problem has received significant attention in classification contexts, it presents unique challenges in regression due to the infinite and continuous nature of the label space. In this work, we introduce a novel framework of *regression with deferral*, which involves deferring the prediction to multiple experts. We present a comprehensive analysis for both the single-stage scenario, where there is simultaneous learning of predictor and deferral functions, and the two-stage scenario, which involves a pre-trained predictor with a learned deferral function. We introduce new surrogate loss functions for both scenarios and prove that they are supported by $\mathcal{H}$-consistency bounds. These bounds provide consistency guarantees that are stronger than Bayes consistency, as they are non-asymptotic and hypothesis set-specific. Our framework is versatile, applying to multiple experts, accommodating any bounded regression losses, addressing both instance-dependent and label-dependent costs, and supporting both single-stage and two-stage methods. Our single-stage formulation subsumes as a special case the recent *regression with abstention* (Cheng et al., 2023) framework, where only a single expert is considered, specifically for the squared loss and a label-independent cost. Minimizing our proposed loss functions directly leads to novel algorithms for regression with deferral. We report the results of extensive experiments showing the effectiveness of our proposed algorithms.

[1]Courant Institute of Mathematical Sciences, New York, NY; [2]Google Research, New York, NY. Correspondence to: Anqi Mao <aqmao@cims.nyu.edu>, Mehryar Mohri <mohri@google.com>, Yutao Zhong <yutao@cims.nyu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

The accuracy of learning algorithms can be greatly enhanced by redirecting uncertain predictions to experts or advanced pre-trained models. Experts can be individuals with specialized domain knowledge or more sophisticated, albeit costly, pre-trained models. The cost of an expert is important to consider, as it may capture the computational resources it requires or the quality of its performance. The cost can further be instance-dependent and label-dependent.

How can we effectively assign each input instance to the most suitable expert among a pool of several, considering both accuracy and cost? This is the challenge of *learning to defer in the presence of multiple experts*, which is prevalent in various domains, including natural language generation tasks, notably large language models (LLMs) (Wei et al., 2022; Bubeck et al., 2023), speech recognition, image annotation and classification, medical diagnosis, financial forecasting, natural language processing, computer vision, and many others.

This paper deals with the problem of learning to defer with multiple experts in the regression setting. While this problem has received significant attention in classification contexts (Hemmer et al., 2022; Keswani et al., 2021; Kerrigan et al., 2021; Straitouri et al., 2022; Benz & Rodriguez, 2022; Verma et al., 2023; Mao et al., 2023a; 2024a), it presents unique challenges in regression due to the infinite and continuous nature of the label space. In particular, the *score-based formulation* commonly used in classification is inapplicable here, since regression problems cannot be represented using multi-class scoring functions, with auxiliary labels corresponding to each expert.

Our approach involves defining prediction and deferral functions, consistent with previous studies in classification (Mao et al., 2023a; 2024a). We present a comprehensive analysis for both the single-stage scenario (simultaneous learning of predictor and deferral functions) (Section 3), and the two-stage scenario (pre-trained predictor with learned deferral function) (Section 4). We introduce new surrogate loss functions for both scenarios and prove that they are supported by $\mathcal{H}$-consistency bounds. These are consistency guarantees that are stronger than Bayes consistency, as they are non-asymptotic and hypothesis set-specific. Our framework is versatile, applying to multiple experts, accommodating

any bounded regression losses, addressing both instance-dependent and label-dependent costs, and supporting both single-stage and two-stage methods. We also instantiate our formulations in the special case of a single expert (Section 5), and demonstrate that our single-stage formulation includes the recent *regression with abstention* framework (Cheng et al., 2023) as a special case, where only a single expert, the squared loss and a label-independent cost are considered. In Section 6, we report the results of extensive experiments showing the effectiveness of our proposed algorithms.

Previous related work. The problem of learning to defer, or the special case of learning with abstention characterized by a single expert and constant cost, has received much attention in classification tasks. Previous work on this topic mainly includes the following formulations or methods: *confidence-based*, *predictor-rejector*, *score-based*, and *selective classification*.

In the *confidence-based formulation*, the rejection function r is based on the magnitude of the value of the predictor h (Chow, 1957; 1970; Bartlett & Wegkamp, 2008; Yuan & Wegkamp, 2010; 2011). This approach has been further extended to multi-class classification in (Ramaswamy et al., 2018; Ni et al., 2019), where the function r is based on the magnitude of the value of the probability (e.g., softmax) corresponding to the predictor h. This formulation becomes inapplicable in regression, since in this setting the prediction value cannot be interpreted as a measure of confidence.

The *score-based formulation* was proposed in the multi-class classification scenario, where the multi-class categories are augmented with additional labels corresponding to the experts, and the deferral is determined using the highest score (Mozannar & Sontag, 2020; Verma & Nalisnick, 2022; Cao et al., 2022; Mao et al., 2024c; Verma et al., 2023; Mao et al., 2024a). However, this formulation is also inapplicable in regression, since regression problems cannot be represented using multi-class scoring functions with auxiliary labels corresponding to each expert.

The approach of learning based on two distinct yet jointly learned functions h and r in this paper is commonly referred as the *predictor-rejector formulation* (Cortes et al., 2016b;a; Charoenphakdee et al., 2021; Cortes et al., 2023; Mohri et al., 2024; Mao et al., 2024b). We show that this method can be extended to the regression setting for deferral with multiple experts, which underscores its versatility and significance.

An alternative approach of *selective classification* (El-Yaniv et al., 2010; Wiener & El-Yaniv, 2011; El-Yaniv & Wiener, 2012; Wiener & El-Yaniv, 2012; 2015; Geifman & El-Yaniv, 2017; 2019) optimizes non-abstained sample generalization error under a fixed selection rate. However, this method does not apply to the deferral case where the cost depends on the label y and where there are multiple experts. Moreover, it has been reported to perform suboptimally compared to the predictor-rejector formulation in regression with abstention settings with constant cost and a single expert (Cheng et al., 2023).

More recently, a series of publications (Mao et al., 2023a; Mohri et al., 2024; Mao et al., 2024b) have explored the two-stage method of learning with deferral or abstention, wherein the predictor h is first learned and subsequently used in the learning process of the deferral function r. This scenario is crucial in practice because the predictor is often given and often cannot be retrained. This method also differs from post-hoc approaches (Okati et al., 2021; Narasimhan et al., 2022), which are not applicable to existing predictors trained in the standard classification scenario. In this work, we will study both the single-stage and two-stage methods for regression with deferral.

In the special case of regression with abstention (corresponding to a single expert and label-independent cost case), Wiener & El-Yaniv (2012) characterized the optimal selector for selective regression, Zaoui et al. (2020) studied non-parametric algorithms, Geifman & El-Yaniv (2019) and Jiang et al. (2020) explored the selective classification using neural network-based algorithms; Shah et al. (2022) used the selective classification with greedy algorithms; De et al. (2020) presented a method that is tailored specifically to ridge regression and small datasets. It proposed an approximate procedure for learning a linear hypothesis that determines which training instances should be deferred. They then used a nearest neighbor approach to defer on new instances; and Li et al. (2023) investigated a two-step no-rejection learning strategy. However, none of these previous publications studied surrogate losses for regression with abstention. This excludes (Cheng et al., 2023), who proposed a single-stage surrogate loss for learning the predictor-rejector pair. We will show that their method coincides with a special case of our single-stage regression with deferral surrogate losses, where there is a single expert and where the cost does not depend on the label y.

Another line of work has studied *dynamic classifier selection* or *dynamic ensemble selection* (Ko et al., 2008; Cruz et al., 2018; Ekanayake et al., 2023), which aims to select the most 'competent' classifiers or ensemble of classifiers in the local region where each instance is located. While these methods also consider how to select an expert from a pool of several, their primary mechanism involves dividing the feature space into distinct regions (a region is typically defined via clustering or nearest-neighbor techniques). In contrast, learning to defer with multiple experts aims to learn a deferral function by minimizing a surrogate loss that accounts for the accuracy and cost of each expert across

all instances. Also, no local region or local competence is considered.

It is worth noting that not all dynamic classifier selection methods consider a local region or local competence. For instance, the recent work by Ekanayake et al. (2023) learns a joint feature acquisition and classifier selection policy to identify the most relevant subset of features based on which classification should be performed, and the classifier to be used. In that sense, the policy for classifier selection essentially functions as a deferral mechanism, since it decides when to defer the decision to an expert. However, this method bases its decisions solely on accuracy, not on cost, whereas our focus is on regression, considering both accuracy and base (inference) cost.

Learning to defer with multiple experts in the classification setting has been studied in (Mao et al., 2023a; Verma et al., 2023; Mao et al., 2024a). Verma et al. (2023) and (Mao et al., 2024a) investigated the single-stage scenario with a score-based formulation, while Mao et al. (2023a) explored the two-stage scenario with both score-based and predictor-rejector formulations. However, as previously highlighted, the score-based formulation does not apply in the regression setting. Our new predictor-rejector formulation not only overcomes this limitation, but also provides the foundation for the design of new deferral algorithms for classification.

2. Preliminaries

Learning scenario of regression. We first describe the familiar problem of supervised regression and introduce our notation. Let $\mathcal{X}$ be the input space and $\mathcal{Y} \subseteq \mathbb{R}$ the label space. We write $\mathcal{D}$ to denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}_{\text{all}}$ be the family of all real-valued measurable functions $h\colon \mathcal{X} \to \mathcal{Y}$, and let $\mathcal{H} \subseteq \mathcal{H}_{\text{all}}$ be the hypothesis set adopted. The learning challenge in regression is to use the labeled sample to find a hypothesis $h \in \mathcal{H}$ with small expected loss or generalization error $\mathcal{E}_{\mathsf{L}}(h)$, with $\mathcal{E}_{\mathsf{L}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathsf{L}(h(x), y)]$, where $\mathsf{L}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function used to measure the magnitude of error in the regression. In the most common case, where L is the squared loss L_2 defined by $\mathsf{L}_2(y', y) = |y' - y|^2$, this represents the mean squared error. In the case where L is the L_1 loss defined by $\mathsf{L}_1(y', y) = |y' - y|$, this represents the mean absolute error. More generally, L can be an L_p loss, defined by $\mathsf{L}_p(y', y) = |y' - y|^p$ for all $y', y \in \mathcal{Y}$, for some $p \geq 1$. In this work, we will consider an arbitrary regression loss function L, subject to the boundedness assumption, that is $\mathsf{L}(y', y) \leq \bar{l}$ for some constant $\bar{l} > 0$ and for all $y, y' \in \mathcal{Y}$. This assumption is commonly adopted in the theoretical analysis of regression (Mohri et al., 2018).

Regression with deferral. We introduce a novel framework where a learner can defer predictions to multiple experts, $g_1, \ldots, g_{n_e}$. Each expert may represent a pre-trained model or a human expert. The learner's output is a pair (h, r), where $h\colon \mathcal{X} \to \mathcal{Y}$ is a prediction function and $r\colon \mathcal{X} \times \{0, 1, \ldots, n_e\} \to \mathbb{R}$ a deferral function. For any input x, $\mathsf{r}(x) = \operatorname{argmax}_{y \in [n_e]} r(x, y) = j$ is the expert deferred to when $j > 0$, no deferral if $j = 0$. The learner makes the prediction $h(x)$ when $\mathsf{r}(x) = 0$, or defers to g_j when $\mathsf{r}(x) = j > 0$. Deferral incurs the cost $\mathsf{L}(g_j(x), y) + \alpha_j$, where α_j is a base cost. The base cost can be the inference cost incurred when querying an expert, factoring in scenarios where engaging experts entails certain costs. Non-deferral incurs the cost $\mathsf{L}(h(x), y)$.

Let $\mathcal{H}_{\text{all}}$ and $\mathcal{R}_{\text{all}}$ denote the family of all measurable functions $h\colon \mathcal{X} \to \mathcal{Y}$ and $r\colon \mathcal{X} \times \{0, 1, \ldots, n_e\} \to \mathbb{R}$ respectively. Given a hypothesis set $\mathcal{H} \subset \mathcal{H}_{\text{all}}$ and a hypothesis set $\mathcal{R} \subset \mathcal{R}_{\text{all}}$, the goal of the regression with deferral problem consists of using the labeled sample to find a pair $(h, r) \in (\mathcal{H}, \mathcal{R})$ with small expected deferral loss $\mathcal{E}_{L_{\text{def}}}(h, r) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_{\text{def}}(h, r, x, y)]$, where L_{def} is defined for any $(h, r) \in \mathcal{H} \times \mathcal{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by

$$L_{\text{def}}(h, r, x, y) = \mathsf{L}(h(x), y)1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} c_j(x, y)1_{\mathsf{r}(x)=j} \quad (1)$$

and $c_j(x, y) > 0$ is a cost function, which can be typically chosen as $\alpha_j + \mathsf{L}(g_j(x), y)$ for an expert g_j and a base cost $\alpha_j > 0$ as mentioned before. Here, we adopt a general cost functions c_j for any j, and only require that the cost remains bounded: $c_j(x, y) \leq \bar{c}_j$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, for some constant $\bar{c}_j > 0$.

Learning with surrogate losses. As with most target losses in learning problems, such as the zero-one loss in classification (Zhang, 2004a; Bartlett et al., 2006; Zhang, 2004b; Tewari & Bartlett, 2007) and the classification with abstention loss (Bartlett & Wegkamp, 2008; Cortes et al., 2016b), directly minimizing the deferral loss L_{def} is computationally hard for most hypothesis sets due to its non-continuity and non-differentiability. Instead, surrogate losses are proposed and adopted in practice. Examples include the hinge loss in binary classification (Cortes & Vapnik, 1995), the (multinomial) logistic loss in multi-class classification (Verhulst, 1838; 1845; Berkson, 1944; 1951), and the predictor-rejector abstention loss in classification with abstention (Cortes et al., 2016b). We will derive surrogate losses for the deferral loss.

Given a surrogate loss $L\colon (h, r, x, y) \mapsto \mathbb{R}_+$, we denote by $\mathcal{E}_L(h, r)$ the generalization error of a pair (h, r), defined as

$$\mathcal{E}_L(h, r) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L(h, r, x, y)].$$

Let $\mathcal{E}_L(\mathcal{H}, \mathcal{R}) = \inf_{h\in\mathcal{H}, r\in\mathcal{R}} \mathcal{E}_L(h, r)$ be the best-in-class error within the family $\mathcal{H} \times \mathcal{R}$. One desired property for surrogate losses in this context is *Bayes-consistency* (Steinwart, 2007). This means that minimizing the expected surrogate

loss over the family of all measurable functions leads to minimizing the expected deferral loss over the same family. More precisely, for a surrogate loss $L\colon (h, r, x, y) \mapsto \mathbb{R}_+$, it is *Bayes-consistent* with respect to L_{def} if,

$$\mathcal{E}_L(h_n, r_n) - \mathcal{E}_L(\mathcal{H}, \mathcal{R}) \xrightarrow{n\to+\infty} 0$$
$$\Longrightarrow \mathcal{E}_{L_{\text{def}}}(h_n, r_n) - \mathcal{E}_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) \xrightarrow{n\to+\infty} 0$$

for all sequences $\{(h_n, r_n)\}_{n\in\mathbb{N}} \subset \mathcal{H} \times \mathcal{R}$ and all distributions. Recently, Awasthi et al. (2022a;b) (see also (Awasthi et al., 2021a;b; 2023a;b; Mao et al., 2023f;c;d; Zheng et al., 2023; Mao et al., 2023b;e; 2024e;d;f)) pointed out that Bayes-consistency does not take into account the hypothesis set $\mathcal{H}$ and is non-asymptotic. Thus, they proposed a stronger guarantee called $\mathcal{H}$-*consistency bounds*. In our context, a surrogate loss L is said to admit an $(\mathcal{H}, \mathcal{R})$-consistency bound with respect to L_{def} if, for all $(h, r) \in \mathcal{H} \times \mathcal{R}$ and all distributions, the following inequality holds:

$$f\Big(\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R})\Big) \leq \mathcal{E}_L(h, r) - \mathcal{E}^*_L(\mathcal{H}, \mathcal{R})$$

for some non-decreasing function $f\colon \mathcal{R}_+ \to \mathcal{R}_+$. In particular, when $(\mathcal{H}, \mathcal{R}) = (\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})$, the $(\mathcal{H}, \mathcal{R})$-consistency bound implies Bayes-consistency.

We will prove $(\mathcal{H}, \mathcal{R})$-consistency bounds for our proposed surrogate losses, which imply their Bayes-consistency. One key term in our bound is the *minimizability gap*, defined as $\mathcal{M}_L(\mathcal{H}, \mathcal{R}) = \mathcal{E}^*_L(\mathcal{H}, \mathcal{R}) - \mathbb{E}_x\, \mathbb{E}_{y|x}[L(h, r, x, y)]$. The minimizability gap characterizes the difference between the best-in-class error and the expected best-in-class pointwise error, and is non-negative. As shown by Mao et al. (2023f), the minimizability gap is upper bounded by the approximation error, satisfying $0 \leq \mathcal{M}_L(\mathcal{H}, \mathcal{R}) \leq \mathcal{E}^*_L(\mathcal{H}, \mathcal{R}) - \mathcal{E}^*_L(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})$ and is generally a finer quantity. The minimizability gap vanishes when $(\mathcal{H}, \mathcal{R}) = (\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})$, or, more generally, when $\mathcal{E}^*_L(\mathcal{H}, \mathcal{R}) = \mathcal{E}^*_L(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})$.

Given a loss function $\ell\colon (r, x, y) \mapsto \mathbb{R}_+$ that only depends on the hypothesis r, the notions of generalization error, best-in-class generalization error, and minimizability gaps, as well as Bayes-consistency and $\mathcal{R}$-consistency bounds, are similarly defined (Awasthi et al., 2022a;b).

In the next sections, we study the problem of learning a pair (h, r) in the framework of regression with deferral. We will derive a family of surrogate losses of L_{def}, starting from first principles. We will show that these loss functions benefit from strong consistency guarantees, which yield directly principled algorithms for our deferral problem. We will specifically distinguish two approaches: the single-stage surrogate losses, where the predictor h and the deferral function r are jointly learned, and the two-stage surrogate losses wherein the predictor h have been previously trained and is fixed and subsequently used in the learning process of the deferral function r.

3. Single-Stage Scenario

In this section, we derive single-stage surrogate losses for the deferral loss and prove their strong $(\mathcal{H}, \mathcal{R})$-consistency bounds guarantees. To do so, we first prove that the following alternative expression holds for L_{def}.

Lemma 3.1. *For any $(h, r) \in \mathcal{H} \times \mathcal{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss function L_{def} can be expressed as follows:*

$$\begin{aligned} L_{\text{def}}(h, r, x, y) = &\left[\sum_{j=1}^{n_e} c_j(x, y)\right] 1_{\mathsf{r}(x)\neq 0} \\ &+ \sum_{j=1}^{n_e}\left[\mathsf{L}(h(x), y) + \sum_{k=1}^{n_e} c_k(x, y) 1_{k\neq j}\right] 1_{\mathsf{r}(x)\neq j} \\ &- (n_e - 1)\left[\mathsf{L}(h(x), y) + \sum_{j=1}^{n_e} c_j(x, y)\right]. \end{aligned}$$

Let ℓ_{0-1} be the zero-one multi-class classification loss defined by $\ell_{0-1}(r, x, y) = 1_{\mathsf{r}(x)\neq y}$ for all $r \in \mathbb{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and let $\ell\colon \mathcal{R} \times \mathcal{X} \times [n_e] \to \mathbb{R}_+$ be a surrogate loss for ℓ_{0-1} such that $\ell \geq \ell_{0-1}$. ℓ may be chosen to be the logistic loss, for example. Since the last term $(n_e - 1)\sum_{j=1}^{n_e} c_j(x, y)$ in the expression of L_{def} in Lemma 3.1 does not depend on h and r, the following loss function L_ℓ defined for all $(h, r) \in \mathcal{H} \times \mathcal{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by

$$\begin{aligned} L_\ell(h, r, x, y) = &\left[\sum_{j=1}^{n_e} c_j(x, y)\right] \ell(r, x, 0) \\ &+ \sum_{j=1}^{n_e}\left[\mathsf{L}(h(x), y) + \sum_{j'\neq j}^{n_e} c_{j'}(x, y)\right] \ell(r, x, j) \\ &- (n_e - 1)\mathsf{L}(h(x), y), \end{aligned} \tag{2}$$

is a natural single-stage surrogate loss for L_{def}. We will show that when ℓ admits a strong $\mathcal{R}$-consistency bound with respect to ℓ_{0-1}, then L_ℓ admits an $(\mathcal{H}, \mathcal{R})$-consistency bound with respect to L_{def}.

Let us underscore the novelty of the surrogate loss formulation presented in equation (2) in the context of learning to defer with multiple experts. This formulation represents a substantial departure from the existing score-based approach prevalent in classification. As previously highlighted, the score-based formulation becomes inapplicable in regression. Our new predictor-rejector formulation not only overcomes this limitation, but also provides the foundation for the design of new deferral algorithms for classification.

We say that a hypothesis set $\mathcal{R}$ is *regular* if for any $x \in \mathcal{X}$, the predictions made by the hypotheses in $\mathcal{R}$ cover the complete set of possible classification labels: $\{\mathsf{r}(x)\colon r \in \mathcal{R}\} = \{0, 1, \ldots, n_e\}$. Widely used hypothesis sets such as linear hypotheses, neural networks, and of course the family of all measurable functions are all regular.

Recent studies by Awasthi et al. (2022b) and Mao et al. (2023f) demonstrate that common multi-class surrogate losses, such as constrained losses and comp-sum losses (including the logistic loss), admit strong $\mathcal{R}$-consistency bounds with respect to the multi-class zero-one loss ℓ_{0-1}, when using such regular hypothesis sets. The next result shows that, for multi-class loss functions ℓ, their corresponding deferral surrogate losses L_ℓ (Eq. (2)) also exhibit $(\mathcal{H}, \mathcal{R})$-consistency bounds with respect to the deferral loss (Eq. (1)).

Theorem 3.2. *Let $\mathcal{R}$ be a regular hypothesis set and ℓ a surrogate loss for the multi-class loss function ℓ_{0-1} upper-bounding ℓ_{0-1}. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, the following $(\mathcal{H}, \mathcal{R})$-consistency bound holds for all $h \in \mathcal{H}$, $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L_\ell}(h, r) - \mathcal{E}^*_{L_\ell}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{L_\ell}(\mathcal{H}, \mathcal{R})\Big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\big)^{1-\alpha} \beta t^\alpha\Big\}$.

The proof is given in Appendix B. As already mentioned, when the best-in-class error coincides with the Bayes error $\mathcal{E}^*_L(\mathcal{H}, \mathcal{R}) = \mathcal{E}^*_L(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})$ for $L = L_\ell$ and $L = L_{\text{def}}$, the minimizability gaps $\mathcal{M}_{L_\ell}(\mathcal{H}, \mathcal{R})$ and $\mathcal{M}_{L_{\text{def}}}(\mathcal{H}, \mathcal{R})$ vanish. In such cases, the $(\mathcal{H}, \mathcal{R})$-consistency bound guarantees that when the surrogate estimation error $\mathcal{E}_{L_\ell}(h, r) - \mathcal{E}^*_{L_\ell}(\mathcal{H}, \mathcal{R})$ is optimized up to ϵ, the estimation error of the deferral loss $\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R})$ is upper bounded by $\overline{\Gamma}(\epsilon)$.

In particular, when both $\mathcal{H}$ and $\mathcal{R}$ include all measurable functions, all the minimizability gap terms in Theorem 3.2 vanish, which yields the following result.

Corollary 3.3. *Given a multi-class loss function $\ell \geq \ell_{0-1}$. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following excess error bound holds for all $r \in \mathcal{R}_{\text{all}}$ and any distribution,*

$$\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}_{\text{all}}) \leq \Gamma(\mathcal{E}_\ell(r) - \mathcal{E}^*_\ell(\mathcal{R}_{\text{all}})).$$

Then, the following excess error bound holds for all $h \in \mathcal{H}_{\text{all}}$, $r \in \mathcal{R}_{\text{all}}$ and any distribution,

$$\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L_\ell}(h, r) - \mathcal{E}^*_{L_\ell}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})\Big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\big)^{1-\alpha} \beta t^\alpha\Big\}$.

In this case, as shown by Mao et al. (2023f), $\Gamma(t)$ can be expressed as $\sqrt{2t}$ for the logistic loss $\ell_{\log} \colon (r, x, y) \mapsto \log_2\big(\sum_{j=0}^{n_e} e^{r(x,j) - r(x,y)}\big)$. Then, by Corollary 3.3, we further obtain the following corollary.

Corollary 3.4. *For any $h \in \mathcal{H}_{\text{all}}$, $r \in \mathcal{R}_{\text{all}}$ and distribution,*

$$\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L_{\ell_{\log}}}(h, r) - \mathcal{E}^*_{L_{\ell_{\log}}}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})\Big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \sqrt{2n_e}\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)^{\frac{1}{2}} t^{\frac{1}{2}}\Big\}$.

By taking the limit on both sides, we derive the Bayes-consistency of these single-stage surrogate losses L_ℓ with respect to the deferral loss L_{def}. More generally, Corollary 3.3 shows that L_ℓ admits an excess error bound with respect to L_{def} when ℓ admits an excess error bound with respect to ℓ_{0-1}.

4. Two-Stage Scenario

In the single-stage scenario, we introduced a family of surrogate losses and resulting algorithms for effectively learning the pair (h, r). However, practical applications often encounter a *two-stage scenario*, where deferral decisions are based on a fixed, pre-trained predictor h. Retraining this predictor is often prohibitively expensive or time-consuming. Thus, this two-stage scenario (Mao et al., 2023a) requires a different approach to optimize deferral decisions controlled by r, while using the existing predictor h.

In this section, we will introduce a principled two-stage algorithm for regression with deferral, with favorable consistency guarantees. Remarkably, we show that the single-stage approach can be adapted for the two-stage scenario if we fix the predictor h and disregard constant terms.

Let h be a predictor learned by minimizing a regression loss L in a first stage. A deferral function r is then learned based on that predictor h and the following loss function L^h_ℓ in the second stage: for any $r \in \mathcal{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$L^h_\ell(r, x, y) = \left[\sum_{j=1}^{n_e} c_j(x, y)\right] \ell(r, x, 0) + \sum_{j=1}^{n_e} \left[\mathsf{L}(h(x), y) + \sum_{j' \neq j}^{n_e} c_{j'}(x, y)\right] \ell(r, x, j), \tag{3}$$

where ℓ is a surrogate loss in the standard multi-class classification. Equation (3) resembles (2), except for the constant term $(n_e - 1)\mathsf{L}(h(x), y)$. In (3), the predictor h remains fixed while only the deferral function r is optimized. In (2), both h and r are learned jointly.

Similarly, we define L^h_{def} as the deferral loss (1) with a fixed

predictor h as follows:

$$L^h_{\text{def}}(r, x, y) = \mathsf{L}(h(x), y)1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} c_j(x, y)1_{\mathsf{r}(x)=j}. \quad (4)$$

Here too, h is fixed in (4). Both L^h_ℓ and L^h_{def} are loss functions defined for deferral function r, while ℓ_ℓ and ℓ_{def} are loss functions defined for pairs $(h, r) \in (\mathcal{H}, \mathcal{R})$.

As with the proposed single-stage approach, the two-stage surrogate losses L^h_ℓ benefit from strong consistency guarantees. We show that in the second stage where a predictor h is fixed, the surrogate loss function L^h_ℓ benefits from $\mathcal{R}$-consistency bounds with respect to L^h_{def} when ℓ admits a strong $\mathcal{R}$-consistency bound with respect to the binary zero-one loss ℓ_{0-1}.

Theorem 4.1. *Given a hypothesis set $\mathcal{R}$, a multi-class loss function $\ell \geq \ell_{0-1}$ and a predictor h. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{L^h_{\text{def}}}(r) - \mathcal{E}^*_{L^h_{\text{def}}}(\mathcal{R}) + \mathcal{M}_{L^h_{\text{def}}}(\mathcal{R}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}) + \mathcal{M}_{L^h_\ell}(\mathcal{R})\Big),$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\Big)^{1-\alpha} \beta t^\alpha$.

The proof is given in Appendix C. When the best-in-class error coincides with the Bayes error, $\mathcal{E}^*_L(\mathcal{R}) = \mathcal{E}^*_L(\mathcal{R}_{\text{all}})$ for $L = L^h_\ell$ and $L = L^h_{\text{def}}$, the minimizability gaps $\mathcal{M}_{L^h_\ell}(\mathcal{R})$ and $\mathcal{M}_{L^h_{\text{def}}}(\mathcal{R})$ vanish. In that case, the $\mathcal{R}$-consistency bound guarantees that when the surrogate estimation error $\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R})$ is optimized up to ϵ, the target estimation error $\mathcal{E}_{L^h_{\text{def}}}(r) - \mathcal{E}^*_{L^h_{\text{def}}}(\mathcal{R})$ is upper bounded by $\overline{\Gamma}(\epsilon)$. In the special case where $\mathcal{H}$ and $\mathcal{R}$ are the family of all measurable functions, all the minimizability gap terms in Theorem 4.1 vanish. Thus, we obtain the following corollary.

Corollary 4.2. *Given a multi-class loss function $\ell \geq \ell_{0-1}$ and a predictor h. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following excess error bound holds for all $r \in \mathcal{R}$ and any distribution,*

$$\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}_{\text{all}}) \leq \Gamma(\mathcal{E}_\ell(r) - \mathcal{E}^*_\ell(\mathcal{R}_{\text{all}})).$$

Then, the following excess error bound holds for all $r \in \mathcal{R}_{\text{all}}$ and any distribution,

$$\mathcal{E}_{L^h_{\text{def}}}(r) - \mathcal{E}^*_{L^h_{\text{def}}}(\mathcal{R}_{\text{all}})) \leq \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}_{\text{all}})\Big), \quad (5)$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\Big)^{1-\alpha} \beta t^\alpha$.

Corollary 4.2 shows that L^h_ℓ admits an excess error bound with respect to L^h_{def} when ℓ admits an excess error bound with respect to ℓ_{0-1}.

We now establish $(\mathcal{H}, \mathcal{R})$-consistency bounds the entire two-stage approach with respect to the deferral loss function L_{def}. This result applies to any multi-class loss function ℓ that satisfies a strong $\mathcal{R}$-consistency bound with respect to the multi-class zero-one loss ℓ_{0-1}.

Theorem 4.3. *Given a hypothesis set $\mathcal{H}$, a regular hypothesis set $\mathcal{R}$ and a multi-class loss function $\ell \geq \ell_{0-1}$. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, the following $(\mathcal{H}, \mathcal{R})$-consistency bound holds for all $h \in \mathcal{H}$, $r \in \mathcal{R}$ and any distribution,

$$\begin{aligned}\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) \\ \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H}) \\ + \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}) + \mathcal{M}_{L^h_\ell}(\mathcal{R})\Big),\end{aligned} \quad (6)$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\Big)^{1-\alpha} \beta t^\alpha$.

The proof is presented in Appendix D. As before, when $\mathcal{H}$ and $\mathcal{R}$ are the family of all measurable functions, all the minimizability gap terms in Theorem 4.3 vanish. In particular, $\Gamma(t)$ can be expressed as $\sqrt{2t}$ for the logistic loss. Thus, we obtain the following on excess error bounds.

Corollary 4.4. *Given a multi-class loss function $\ell \geq \ell_{0-1}$. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following excess error bound holds for all $r \in \mathcal{R}_{\text{all}}$ and any distribution,*

$$\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}_{\text{all}}) \leq \Gamma(\mathcal{E}_\ell(r) - \mathcal{E}^*_\ell(\mathcal{R}_{\text{all}})).$$

Then, the following excess error bound holds for all $h \in \mathcal{H}_{\text{all}}$, $r \in \mathcal{R}_{\text{all}}$ and any distribution,

$$\begin{aligned}\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}}) \\ \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}_{\text{all}}) + \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}_{\text{all}})\Big),\end{aligned} \quad (7)$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)\Big)^{1-\alpha} \beta t^\alpha$. *In particular,* $\overline{\Gamma}(t) = \sqrt{2n_e}\big(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\big)^{\frac{1}{2}} t^{\frac{1}{2}}$ *for* $\ell = \ell_{\log}$.

Corollary 4.4 shows that our two-stage approach admits an excess error bound with respect to L_{def} when ℓ admits an excess error bound with respect to ℓ_{0-1}. More generally,

when the minimizability gaps are zero, as when the best-in-class errors coincide with the Bayes errors, the $(\mathcal{H}, \mathcal{R})$-consistency bound of Theorem 4.3 guarantees that the target estimation error, $\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R})$, is upper bounded by $\epsilon_1 + \overline{\Gamma}(\epsilon_2)$ provided that the surrogate estimation error in the first stage, $\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H})$, is reduced to ϵ_1 and the surrogate estimation error in the second stage, $\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R})$, reduced to ϵ_2.

Significance and novelty. The challenges in dealing with multiple experts in the theoretical analysis of learning to deferral in regression arise first from the need to formulate new surrogate losses that cannot be directly extended from previous work. Furthermore, proving theoretical guarantees requires analyzing the conditional regret for both the surrogate and target deferral loss, which becomes more complex with multiple experts. The novelty and significance of our work are rooted in these new surrogate losses and algorithmic solutions, which come with strong theoretical guarantees specifically tailored for this context. These enhancements are non-trivial and represent a substantial extension beyond the existing framework of regression with abstention, which is limited in scope to a single expert, squared loss, and label-independent cost.

5. Special Case of a Single Expert

In the special case of a single expert, $n_e = 1$, both the single-stage surrogate loss L_ℓ and the two-stage surrogate loss L^h_ℓ can be simplified as follows:

$$c(x, y)\ell(r, x, 0) + \mathsf{L}(h(x), y)\ell(r, x, 1).$$

Let $\ell(r, x, 0) = \Phi(r(x))$ and $\ell(r, x, 1) = \Phi(-r(x))$, where $\Phi\colon \mathbb{R} \to \mathbb{R}_+$ is a non-increasing auxiliary function upper bounding the indicator $u \mapsto 1_{u \leq 0}$. Here, $r\colon \mathcal{X} \to \mathbb{R}$ is a function whose sign determines if there is deferral, that is $r(x) \leq 0$:

$$\ell_{\text{def}}(h, r, x, y) = \mathsf{L}(h(x), y)1_{r(x)>0} + c(x, y)1_{r(x)\leq 0}.$$

As an example, Φ can be the auxiliary function that defines a margin-based loss in the binary classification. Thus, both the single-stage surrogate loss ℓ_Φ and the two-stage surrogate loss ℓ^h_Φ can be reformulated as follows:

$$c(x, y)\Phi(r(x)) + \mathsf{L}(h(x), y)\Phi(-r(x)). \tag{8}$$

Some common examples of Φ are listed in Table 2 in Appendix E. Note that (8) is a straightforward extension of the two-stage formulation given in (Mao et al., 2024b, Eq. (5)). In their formulation, the zero-one loss function replaces the regression loss and is tailored for the classification context. A special case of the straightforward extension (8) is one where the cost does not depend on the label y and the squared loss is considered. This coincides with the loss function (Cheng et al., 2023, Eq. (10)) in the context of regression with abstention. It is important to note that incorporating y as argument of the cost functions is crucial in the more general deferral setting, as each cost takes into account the accuracy of the corresponding expert.

Let the binary zero-one loss be $\ell^{\text{bi}}_{0-1}(r, x, y) = 1_{\text{sign}(r(x))\neq y}$, where $\text{sign}(\alpha) = 1_{\alpha>0} - 1_{\alpha\leq 0}$. We say that a hypothesis set $\mathcal{R}$ consists of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ is *regular*, if $\{\text{sign}(r(x))\colon r \in \mathcal{R}\} = \{+1, -1\}$ for any $x \in \mathcal{X}$.

Then, Theorems 3.2 and 4.3 can be reduced to Theorems 5.1 and 5.3 below respectively. We present these guarantees and their corresponding corollaries in the following sections.

5.1. Single-Stage Guarantees

Here, we present guarantees for the single-stage surrogate.

Theorem 5.1. *Given a hypothesis set $\mathcal{H}$, a regular hypothesis set $\mathcal{R}$ and a margin-based loss function Φ. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

$$\mathcal{E}_{\ell^{\text{bi}}_{0-1}}(r) - \mathcal{E}^*_{\ell^{\text{bi}}_{0-1}}(\mathcal{R}) + \mathcal{M}_{\ell^{\text{bi}}_{0-1}}(\mathcal{R}) \leq \Gamma(\mathcal{E}_\Phi(r) - \mathcal{E}^*_\Phi(\mathcal{R}) + \mathcal{M}_\Phi(\mathcal{R})).$$

Then, the following $(\mathcal{H}, \mathcal{R})$-consistency bound holds for all $h \in \mathcal{H}$, $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{\ell_{\text{def}}}(h, r) - \mathcal{E}^*_{\ell_{\text{def}}}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{\ell_{\text{def}}}(\mathcal{H}, \mathcal{R}) \leq \overline{\Gamma}\Big(\mathcal{E}_{\ell_\Phi}(h, r) - \mathcal{E}^*_{\ell_\Phi}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{\ell_\Phi}(\mathcal{H}, \mathcal{R})\Big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \big(\overline{l} + \overline{c}\big)^{1-\alpha}\beta t^\alpha\Big\}$.

In particular, when $\mathcal{H}$ and $\mathcal{R}$ are the family of all measurable functions, all the minimizability gap terms in Theorem 5.1 vanish. In this case, as shown by Awasthi et al. (2022a), $\Gamma(t)$ can be expressed as $\frac{t^2}{2}$ for exponential and logistic loss, t^2 for quadratic loss and t for hinge, sigmoid and ρ-margin losses. Thus, the following result holds.

Corollary 5.2. *Given a margin-based loss function Φ. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following excess error bound holds for all $r \in \mathcal{R}_{\text{all}}$ and any distribution,*

$$\mathcal{E}_{\ell^{\text{bi}}_{0-1}}(r) - \mathcal{E}^*_{\ell^{\text{bi}}_{0-1}}(\mathcal{R}_{\text{all}}) \leq \Gamma(\mathcal{E}_\Phi(r) - \mathcal{E}^*_\Phi(\mathcal{R}_{\text{all}})).$$

Then, the following excess error bound holds for all $h \in \mathcal{H}_{\text{all}}$, $r \in \mathcal{R}_{\text{all}}$ and any distribution,

$$\mathcal{E}_{\ell_{\text{def}}}(h, r) - \mathcal{E}^*_{\ell_{\text{def}}}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}}) \leq \overline{\Gamma}\Big(\mathcal{E}_{\ell_\Phi}(h, r) - \mathcal{E}^*_{\ell_\Phi}(\mathcal{H}_{\text{all}}, \mathcal{R}_{\text{all}})\Big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \big(\overline{l} + \overline{c}\big)^{1-\alpha}\beta t^\alpha\Big\}$. *In particular,* $\overline{\Gamma}(t) = \max\Big\{t, \frac{1}{2}(\overline{l} + \overline{c})^{\frac{1}{2}} t^{\frac{1}{2}}\Big\}$ *for* $\Phi = \Phi_{\exp}$ *and* $\Phi_{\log}$,

Table 1. System MSE of deferral with multiple experts: mean ± standard deviation over three runs.

Dataset	Base cost	Method	Base model	Single expert	Two experts	Three experts
Airfoil	✗	Single	—	18.98 ± 2.44	13.16 ± 0.93	**8.53 ± 1.57**
	✗	Two	23.35 ± 1.90	18.64 ± 1.96	13.33 ± 0.92	**8.81 ± 1.56**
	✓	Single	—	18.83 ± 2.14	13.79 ± 0.75	**8.64 ± 1.40**
	✓	Two	23.35 ± 1.90	19.15 ± 1.99	15.12 ± 0.62	**10.06 ± 1.54**
Housing	✗	Single	—	14.85 ± 5.40	14.75 ± 3.53	**12.43 ± 2.03**
	✗	Two	22.72 ± 7.68	16.26 ± 5.58	14.82 ± 3.60	**12.02 ± 1.97**
	✓	Single	—	15.17 ± 5.18	15.07 ± 2.88	**14.80 ± 3.48**
	✓	Two	22.72 ± 7.68	16.24 ± 4.64	15.62 ± 3.04	**14.87 ± 4.04**
Concrete	✗	Single	—	104.38 ± 5.55	41.08 ± 2.05	**37.83 ± 2.60**
	✗	Two	120.20 ± 8.09	114.73 ± 6.50	44.46 ± 5.34	**36.75 ± 1.76**
	✓	Single	—	105.01 ± 5.40	39.52 ± 2.81	**38.46 ± 1.79**
	✓	Two	120.20 ± 8.09	114.11 ± 5.34	39.93 ± 2.77	**37.51 ± 2.32**

$\overline{\Gamma}(t) = \max\left\{t, (\bar{l}+\bar{c})^{\frac{1}{2}} t^{\frac{1}{2}}\right\}$ *for* $\Phi = \Phi_{\mathrm{quad}}$, *and* $\overline{\Gamma}(t) = t$ *for* $\Phi = \Phi_{\mathrm{hinge}}$, Φ_{sig}, *and* Φ_ρ.

By taking the limit on both sides, we derive the Bayes-consistency and excess error bound of these single-stage surrogate losses ℓ_Φ with respect to the deferral loss ℓ_{def}. More generally, Corollary 5.2 shows that ℓ_Φ admits an excess error bound with respect to ℓ_{def} when Φ admits an excess error bound with respect to ℓ_{0-1}. Corollary 5.2 also include the theoretical guarantees in (Cheng et al., 2023, Theorems 7 and 8) as a special case where the cost does not depend on the label y and the squared loss is considered.

5.2. Two-Stage Guarantees

Here, we present guarantees for the two-stage surrogate.

Theorem 5.3. *Given a hypothesis set* $\mathcal{H}$, *a regular hypothesis set* $\mathcal{R}$ *and a margin-based loss function* Φ. *Assume that there exists a function* $\Gamma(t) = \beta t^\alpha$ *for some* $\alpha \in (0, 1]$ *and* $\beta > 0$, *such that the following* $\mathcal{R}$*-consistency bound holds for all* $r \in \mathcal{R}$ *and any distribution,*

$$\mathcal{E}_{\ell^{\mathrm{bi}}_{0-1}}(r) - \mathcal{E}^*_{\ell^{\mathrm{bi}}_{0-1}}(\mathcal{R}) + \mathcal{M}_{\ell^{\mathrm{bi}}_{0-1}}(\mathcal{R}) \leq \Gamma\left(\mathcal{E}_\Phi(r) - \mathcal{E}^*_\Phi(\mathcal{R}) + \mathcal{M}_\Phi(\mathcal{R})\right).$$

Then, the following $(\mathcal{H}, \mathcal{R})$*-consistency bound holds for all* $h \in \mathcal{H}$, $r \in \mathcal{R}$ *and any distribution,*

$$\begin{aligned}\mathcal{E}_{\ell_{\mathrm{def}}}(h, r) - \mathcal{E}^*_{\ell_{\mathrm{def}}}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{\ell_{\mathrm{def}}}(\mathcal{H}, \mathcal{R}) &\leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H}) \\ &\quad + \overline{\Gamma}\left(\mathcal{E}_{\ell^h_\Phi}(r) - \mathcal{E}^*_{\ell^h_\Phi}(\mathcal{R}) + \mathcal{M}_{\ell^h_\Phi}(\mathcal{R})\right),\end{aligned}$$

where $\overline{\Gamma}(t) = \left(\bar{l}+\bar{c}\right)^{1-\alpha} \beta t^\alpha$.

As before, when $\mathcal{H}$ and $\mathcal{R}$ include all measurable functions, all the minimizability gap terms in Theorem 4.3 vanish. In particular, $\Gamma(t)$ can be expressed as $\frac{t^2}{2}$ for exponential and logistic loss, t^2 for quadratic loss and t for hinge, sigmoid and ρ-margin losses (Awasthi et al., 2022a). Thus, we obtain the following result.

Corollary 5.4. *Given a margin-based loss function* Φ. *Assume that there exists a function* $\Gamma(t) = \beta t^\alpha$ *for some* $\alpha \in (0, 1]$ *and* $\beta > 0$, *such that the following excess error bound holds for all* $r \in \mathcal{R}_{\mathrm{all}}$ *and any distribution,*

$$\mathcal{E}_{\ell^{\mathrm{bi}}_{0-1}}(r) - \mathcal{E}^*_{\ell^{\mathrm{bi}}_{0-1}}(\mathcal{R}_{\mathrm{all}}) \leq \Gamma\left(\mathcal{E}_\Phi(r) - \mathcal{E}^*_\Phi(\mathcal{R}_{\mathrm{all}})\right).$$

Then, the following excess error bound holds for all $h \in \mathcal{H}_{\mathrm{all}}$, $r \in \mathcal{R}_{\mathrm{all}}$ *and any distribution,*

$$\begin{aligned}&\mathcal{E}_{\ell_{\mathrm{def}}}(h, r) - \mathcal{E}^*_{\ell_{\mathrm{def}}}(\mathcal{H}_{\mathrm{all}}, \mathcal{R}_{\mathrm{all}}) \\ &\quad \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}_{\mathrm{all}}) + \overline{\Gamma}\left(\mathcal{E}_{\ell^h_\Phi}(r) - \mathcal{E}^*_{\ell^h_\Phi}(\mathcal{R}_{\mathrm{all}})\right),\end{aligned} \tag{9}$$

where $\overline{\Gamma}(t) = \left(\bar{l}+\bar{c}\right)^{1-\alpha} \beta t^\alpha$. *In particular,* $\overline{\Gamma}(t) = \frac{1}{2}(\bar{l} + \bar{c})^{\frac{1}{2}} t^{\frac{1}{2}}$ *for* $\Phi = \Phi_{\exp}$ *and* $\Phi_{\log}$, $\overline{\Gamma}(t) = (\bar{l}+\bar{c})^{\frac{1}{2}} t^{\frac{1}{2}}$ *for* $\Phi = \Phi_{\mathrm{quad}}$, *and* $\overline{\Gamma}(t) = t$ *for* $\Phi = \Phi_{\mathrm{hinge}}$, Φ_{sig}, *and* Φ_ρ.

Corollary 5.4 shows that the proposed two-stage approach admits an excess error bound with respect to ℓ_{def} when Φ admits an excess error bound with respect to ℓ_{0-1}. More generally, in the cases where the minimizability gaps are zero (as when the best-in-class errors coincide with the Bayes errors), the $(\mathcal{H}, \mathcal{R})$-consistency bound in Theorem 5.3 guarantees that when the surrogate estimation error in the first stage $\mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H})$ is minimized up to ϵ_1 and the surrogate estimation error in the second stage $\mathcal{E}_{\ell^h_\Phi}(r) - \mathcal{E}^*_{\ell^h_\Phi}(\mathcal{R})$ is minimized up to ϵ_2, the target estimation error $\mathcal{E}_{\ell_{\mathrm{def}}}(h, r) - \mathcal{E}^*_{\ell_{\mathrm{def}}}(\mathcal{H}, \mathcal{R})$ is upper bounded by $\epsilon_1 + \overline{\Gamma}(\epsilon_2)$.

It is worth noting that while our theoretical results are general, they can be effectively applied to derive bounds for specific loss functions. Applicable regression loss functions are those with a boundedness assumption, and any classification loss function that benefits from $\mathcal{H}$-consistency bounds is suitable. Our theoretical analysis guides the selection of the loss function by considering several key factors: the functional form $\overline{\Gamma}$ of the bound, the approximation properties indicated by the minimizability gaps, the optimization

advantages of each loss function, and how favorably the bounds depend on the number of experts.

6. Experiments

In this section, we report the empirical results for our single-stage and two-stage algorithms for regression with deferral on three datasets from the UCI machine learning repository (Asuncion & Newman, 2007), the `Airfoil`, `Housing` and `Concrete`, which have also been studied in (Cheng et al., 2023).

Setup and Metrics. For each dataset, we randomly split it into a training set of 60% examples, a validation set of 20% examples and a test set of 20% examples. We report results averaged over three such random splits. We adopted linear models for both the predictor h and the deferral function r. We considered three experts g_1, g_2 and g_3, each trained by feedforward neural networks with ReLU activation functions (Nair & Hinton, 2010) with one, two, and three hidden layers, respectively. We used the Adam optimizer (Kingma & Ba, 2014) with a batch size of 256 and 2,000 training epochs. The learning rate for all datasets is selected from $\{0.01, 0.05, 0.1\}$. We adopted the squared loss as the regression loss ($\mathsf{L} = \mathsf{L}_2$). For our single-stage surrogate loss (2) and two-stage surrogate loss (3), we choose $\ell = \ell_{\log}$ as the logistic loss. In the experiments, we considered two types of costs: $c_j(x, y) = L(g_j(x), y)$ and $c_j(x, y) = L(g_j(x), y) + \alpha_j$, for $1 \leq j \leq n_e$. In the first case, the cost corresponds exactly to the expert's squared error. In the second case, the constant α_j is the base cost for deferring to expert g_j. We chose $(\alpha_1, \alpha_2, \alpha_3) = (4.0, 8.0, 12.0)$. We selected those values of α because they are empirically determined to encourage an optimal balance, ensuring a reasonable number of input instances are deferred to each expert. For evaluation, we compute the system mean squared error (MSE), that is the average squared difference between the target value and the prediction made by the predictor h or the expert selected by the deferral function r. We also report the empirical regression loss, $\frac{1}{n}\sum_{i=1}^{n} \mathsf{L}(h(x_i), y_i)$, of the base model used in the two-stage algorithm.

Results. In Table 1, we report the mean and standard deviation of the empirical regression loss of the base model, as well as the System MSE obtained by using a single expert g_1, two experts g_1 and g_2 and three experts g_1, g_2 and g_3, over three random splits of the dataset. Here, the base model is the predictor. We did not report its performance in the single-stage method because it is independently trained and exhibits varying accuracies across settings with single expert, two experts, and three experts, in contrast to the two-stage method where the base model is pre-learned and fixed. Additionally, in the single-stage method, the base model is always used in conjunction with deferral, rather than being used separately.

Table 1 shows that the performance of both our single-stage and two-stage algorithms improves as more experts are taken into account across the `Airfoil`, `Housing` and `Concrete` datasets. In particular, our algorithms are able to effectively defer difficult test instances to more suitable experts and outperform the base model. Table 1 also shows that the two-stage algorithm usually does not perform as well as the single-stage one in the regression setting, mainly due to the error accumulation in the two-stage process, particularly if the first-stage predictor has large errors. However, the two-stage algorithm is still useful when an existing predictor cannot be retrained due to cost or time constraints. In such cases, we can still improve its performance by learning a multi-expert deferral with our two-stage surrogate losses.

Note that since our work is the first to study regression with multi-expert deferral, there are no existing baselines for direct comparison. Nevertheless, for completeness, we include additional experimental results with three simple baselines in Appendix F, further demonstrating the effectiveness of our approach.

In real-life scenarios, "cancellation effects" might occur when using experts with similar expertise (Verma et al., 2023). However, the experts we have used do not exhibit such an effect. This is because, for each expert, there are specific input instances that they can predict correctly, which others cannot. Therefore, without base costs, the system's error after using deferral is lower than that of any individual expert. Furthermore, in our scenario, we operate under the assumption that the experts are predefined.

Our $\mathcal{H}$-consistency guarantees demonstrate that in both single- and two-stage scenarios, given a sufficient amount of data, our algorithms can approximate results close to the optimal deferral loss for the given experts. Our analysis and experiments do not directly address the process of selecting experts beforehand. Optimally selecting diverse and accurate experts is an interesting research question.

7. Conclusion

We introduced a novel and principled framework for regression with deferral, enhancing the accuracy and reliability of regression predictions through the strategic use of multiple experts. Our comprehensive analysis of this framework includes the formulation of novel surrogate losses for both single-stage and two-stage scenarios, and the proof of strong $\mathcal{H}$-consistency bounds. These theoretical guarantees lead to powerful algorithms that leverage multiple experts, providing a powerful tool for addressing the inherent challenges of regression problems. Empirical results validate the effectiveness of our approach, showcasing its practical significance and opening up new avenues for developing robust solutions across diverse regression tasks.

Acknowledgements

We thank the reviewers for suggesting useful baselines for our experiments and for bringing to our attention the literature on dynamic classifier selection, which, while distinct from the learning-to-defer framework studied in this paper, has somewhat relevant connections.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Awasthi, P., Frank, N., Mao, A., Mohri, M., and Zhong, Y. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, pp. 9804–9815, 2021a.

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

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pp. 1117–1174, 2022a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class H-consistency bounds. In *Advances in neural information processing systems*, pp. 782–795, 2022b.

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

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. DC-programming for neural network optimizations. *Journal of Global Optimization*, 2023b.

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

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

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

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

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

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

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

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

Cheng, X., Cao, Y., Wang, H., Wei, H., An, B., and Feng, L. Regression with cost-based rejection. In *Advances in Neural Information Processing Systems*, 2023.

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

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

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

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

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

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

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018.

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

Ekanayake, S. P., Zois, D.-S., and Chelmis, C. Sequential datum–wise feature acquisition and classifier selection. *IEEE Transactions on Artificial Intelligence*, 2023.

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

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

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

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

Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., and Satzger, G. Forming effective human-ai teams: Building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948*, 2022.

Jiang, W., Zhao, Y., and Wang, Z. Risk-controlled selective prediction for regression deep neural network models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

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

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

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

Ko, A. H., Sabourin, R., and Britto Jr, A. S. From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*, 41(5):1718–1731, 2008.

Li, X., Liu, S., Sun, C., and Wang, H. When no-rejection learning is optimal for regression with rejection. *arXiv preprint arXiv:2307.02932*, 2023.

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

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

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

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

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

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

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

Mao, A., Mohri, M., and Zhong, Y. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 822–867, 2024b.

Mao, A., Mohri, M., and Zhong, Y. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, pp. 4753–4761, 2024c.

Mao, A., Mohri, M., and Zhong, Y. H-consistency guarantees for regression. *arXiv preprint arXiv:2403.19480*, 2024d.

Mao, A., Mohri, M., and Zhong, Y. Top-k classification and cardinality-aware prediction. *arXiv preprint arXiv:2403.19625*, 2024e.

Mao, A., Mohri, M., and Zhong, Y. A universal growth rate for learning with smooth surrogate losses. *arXiv preprint arXiv:2405.05968*, 2024f.

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

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

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, pp. 29292–29304, 2022.

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

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

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

Shah, A., Bu, Y., Lee, J. K., Das, S., Panda, R., Sattigeri, P., and Wornell, G. W. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615, 2022.

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

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

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

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

Verhulst, P. F. 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.

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

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

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. Emergent abilities of large language models. *CoRR*, abs/2206.07682, 2022.

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

Wiener, Y. and El-Yaniv, R. Pointwise tracking the optimal regression function. *Advances in Neural Information Processing Systems*, 25, 2012.

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

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

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

Zaoui, A., Denis, C., and Hebiri, M. Regression with reject option and application to knn. In *Advances in Neural Information Processing Systems*, pp. 20073–20082, 2020.

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

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

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

Contents of Appendix

A. Useful lemmas

Lemma 3.1. *For any $(h, r) \in \mathcal{H} \times \mathcal{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss function L_{def} can be expressed as follows:*

$$L_{\text{def}}(h, r, x, y) = \left[\sum_{j=1}^{n_e} c_j(x, y)\right] 1_{\mathsf{r}(x)\neq 0} + \sum_{j=1}^{n_e}\left[\mathsf{L}(h(x), y) + \sum_{k=1}^{n_e} c_k(x, y) 1_{k\neq j}\right] 1_{\mathsf{r}(x)\neq j} - (n_e - 1)\left[\mathsf{L}(h(x), y) + \sum_{j=1}^{n_e} c_j(x, y)\right].$$

Proof. Observe that, for any $x \in \mathcal{X}$, since $\mathsf{r}(x) = 0$ if and only if $\mathsf{r}(x) \neq j$ for all $j \geq 1$, the following equality holds:

$$1_{\mathsf{r}(x)=0} = 1_{\wedge_{j=1}^{n_e}\{\mathsf{r}(x)\neq j\}} = \sum_{j=1}^{n_e} 1_{\mathsf{r}(x)\neq j} - (n_e - 1).$$

Similarly, since $\mathsf{r}(x) = j$ if and only if $\mathsf{r}(x) \neq k$ for $k \neq j$ and $\mathsf{r}(x) \neq 0$, the following equality holds:

$$1_{\mathsf{r}(x)=j} = 1_{\mathsf{r}(x)\neq 0} + \sum_{k=1}^{n_e} 1_{\mathsf{r}(x)\neq k} 1_{k\neq j} - (n_e - 1).$$

In view of these identities, starting from the definition of L_{def}, we can write:

$$\begin{aligned}
L_{\text{def}}(h, r, x, y) &= \mathsf{L}(h(x), y) 1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} c_j(x, y) 1_{\mathsf{r}(x)=j} \\
&= \mathsf{L}(h(x), y)\left[\sum_{j=1}^{n_e} 1_{\mathsf{r}(x)\neq j} - (n_e - 1)\right] + \sum_{j=1}^{n_e} c_j(x, y)\left[1_{\mathsf{r}(x)\neq 0} + \sum_{k=1}^{n_e} 1_{\mathsf{r}(x)\neq k} 1_{k\neq j} - (n_e - 1)\right] \\
&= \left[\sum_{j=1}^{n_e} c_j(x, y)\right] 1_{\mathsf{r}(x)\neq 0} + \sum_{j=1}^{n_e} \mathsf{L}(h(x), y) 1_{\mathsf{r}(x)\neq j} + \sum_{j=1}^{n_e}\sum_{k=1}^{n_e} c_j(x, y) 1_{k\neq j} 1_{\mathsf{r}(x)\neq k} \\
&\quad - (n_e - 1)\left[\mathsf{L}(h(x), y) + \sum_{j=1}^{n_e} c_j(x, y)\right] \\
&= \left[\sum_{j=1}^{n_e} c_j(x, y)\right] 1_{\mathsf{r}(x)\neq 0} + \sum_{j=1}^{n_e} \mathsf{L}(h(x), y) 1_{\mathsf{r}(x)\neq j} + \sum_{k=1}^{n_e}\sum_{j=1}^{n_e} c_k(x, y) 1_{k\neq j} 1_{\mathsf{r}(x)\neq j} \\
&\quad - (n_e - 1)\left[\mathsf{L}(h(x), y) + \sum_{j=1}^{n_e} c_j(x, y)\right] && \text{(change of variables } k \text{ and } j\text{)} \\
&= \left[\sum_{j=1}^{n_e} c_j(x, y)\right] 1_{\mathsf{r}(x)\neq 0} + \sum_{j=1}^{n_e}\left[\mathsf{L}(h(x), y) + \sum_{k=1}^{n_e} c_k(x, y) 1_{k\neq j}\right] 1_{\mathsf{r}(x)\neq j} - (n_e - 1)\left[\mathsf{L}(h(x), y) + \sum_{j=1}^{n_e} c_j(x, y)\right].
\end{aligned}$$

This completes the proof. □

Lemma A.1. *Assume that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, for any $p = (p_0, \ldots, p_{n_e}) \in \Delta^{n_e}$ and $x \in \mathcal{X}$, we have

$$\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right) \leq \Gamma\left(\sum_{j=0}^{n_e} p_j \ell(r, x, j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r, x, j)\right)\right).$$

Proof. For any $x \in \mathcal{X}$, consider a distribution δ_x that concentrates on that point. Let $p_j = \mathbb{P}(y = j \mid x)$, $j \in [n_e]$. Then, by definition, $\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{R})$ can be expressed as

$$\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{R}) = \sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right).$$

Similarly, $\mathcal{E}_\ell(r) - \mathcal{E}_\ell^*(\mathcal{R}) + \mathcal{M}_\ell(\mathcal{R})$ can be expressed as

$$\mathcal{E}_\ell(r) - \mathcal{E}_\ell^*(\mathcal{R}) + \mathcal{M}_\ell(\mathcal{R}) = \sum_{j=0}^{n_e} p_j \ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j)\right).$$

Since the $\mathcal{R}$-consistency bound holds by the assumption, we complete the proof. □

B. Proof of Theorem 3.2

Theorem 3.2. *Let $\mathcal{R}$ be a regular hypothesis set and ℓ a surrogate loss for the multi-class loss function ℓ_{0-1} upper-bounding ℓ_{0-1}. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0,1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

$$\mathcal{E}_{\ell_{0-1}}(r) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{R}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{R}) \le \Gamma(\mathcal{E}_\ell(r) - \mathcal{E}_\ell^*(\mathcal{R}) + \mathcal{M}_\ell(\mathcal{R})).$$

Then, the following $(\mathcal{H},\mathcal{R})$-consistency bound holds for all $h \in \mathcal{H}$, $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{L_{\text{def}}}(h,r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_{\text{def}}}(\mathcal{H},\mathcal{R}) \le \overline{\Gamma}\big(\mathcal{E}_{L_\ell}(h,r) - \mathcal{E}^*_{L_\ell}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_\ell}(\mathcal{H},\mathcal{R})\big),$$

where $\overline{\Gamma}(t) = \max\Big\{t, \big(n_e\big(\bar{l} + \sum_{j=1}^{n_e}\bar{c}_j\big)\big)^{1-\alpha}\beta t^\alpha\Big\}$.

Proof. The conditional error of the deferral loss can be expressed as

$$\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h,r,x,y)] = \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e}\mathop{\mathbb{E}}_{y|x}[c_j(x,y)]1_{\mathsf{r}(x)=j}. \tag{10}$$

Let $\bar{c}_0(x) = \inf_{h\in\mathcal{H}}\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]$ and $\bar{c}_j(x) = \mathbb{E}_{y|x}[c(x,y)]$. Thus, the best-in class conditional error of the deferral loss can be expressed as

$$\inf_{h\in\mathcal{H},r\in\mathcal{R}}\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h,r,x,y)] = \min_{j\in[n_e]}\bar{c}_j(x). \tag{11}$$

The conditional error of the surrogate loss can be expressed as

$$\begin{aligned}\mathop{\mathbb{E}}_{y|x}[\ell_\ell(h,r,x,y)] = &\left(\sum_{j=1}^{n_e}\mathop{\mathbb{E}}_{y|x}[c_j(x,y)]\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e}\mathop{\mathbb{E}}_{y|x}[c_{j'}(x,y)]\right)\ell(r,x,j)\\ &- (n_e-1)\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)].\end{aligned} \tag{12}$$

Note that the coefficient of term $\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]$ satisfies $\sum_{j=1}^{n_e}\ell(r,x,j) - (n_e-1) \ge 0$ since $\ell \ge \ell_{0-1}$. Thus, the best-in class conditional error of the surrogate loss can be expressed as

$$\begin{aligned}&\inf_{h\in\mathcal{H},r\in\mathcal{R}}\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)]\\ &= \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e}\bar{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\bar{c}_0(x) + \sum_{j'\neq j}^{n_e}\bar{c}_{j'}(x)\right)\ell(r,x,j)\right] - (n_e-1)\bar{c}_0(x).\end{aligned} \tag{13}$$

Next, we analyze four cases separately to show that the calibration gap of the surrogate loss can be lower bounded by that of the deferral loss.

Case I: $\mathsf{r}(x) = 0$ and $\bar{c}_0(x) \le \min_{j=1}^{n_e}\bar{c}_j(x)$. In this case, by (10) and (11), the calibration gap of the deferral loss can be expressed as

$$\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}}\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h,r,x,y)] = \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] - \inf_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)].$$

By (12) and (13), the calibration gap of the surrogate loss can be expressed as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] \\
&= \left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j) - (n_e-1)\mathop{\mathbb{E}}_{y|x}\mathsf{L}(h(x),y) \\
&\quad - \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\overline{c}_0(x) + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\right] + (n_e-1)\overline{c}_0(x).
\end{aligned}$$

Since $\ell \geq \ell_{0-1}$, we have $\ell(r,x,j) \geq 1$ for $j \neq 0$. By eliminating the infimum over $\mathcal{R}$ from the final line, and consequently canceling the terms related to $\overline{c}_j(x)$ for $j \neq 0$, the calibration gap of the surrogate loss can be lower bounded as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] \\
&\geq \left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] - \inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)]\right)\left(\sum_{j=1}^{n_e}\ell(r,x,j) - n_e + 1\right) \\
&\geq \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] - \inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] \qquad \left(\textstyle\sum_{j=1}^{n_e}\ell(r,x,j) - n_e + 1 \geq 1\right) \\
&= \mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)].
\end{aligned}$$

Case II: $\overline{c}_0(x) > \min_{j=1}^{n_e} \overline{c}_j(x)$**.** In this case, by (10) and (11), the calibration gap of the deferral loss can be expressed as

$$\mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] = \overline{c}_{\mathsf{r}(x)}(x) - \min_{j=1}^{n_e} \overline{c}_j(x).$$

By (12) and (13), the calibration gap of the surrogate loss can be expressed as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] \\
&= \left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j) - (n_e-1)\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] \\
&\quad - \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\overline{c}_0(x) + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\right] + (n_e-1)\overline{c}_0(x).
\end{aligned}$$

Using the fact that $\overline{c}_0(x) = \inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] \leq \mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]$, the calibration gap of the surrogate loss can be lower bounded as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] \\
&\geq \left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j) \\
&\quad - \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\right] \\
&= n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)\left[\sum_{j=0}^{n_e} p_j\ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j\ell(r,x,j)\right)\right]
\end{aligned}$$

where we let $p_0 = \frac{\sum_{j=1}^{n_e} \overline{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]+\sum_{j=1}^{n_e} \overline{c}_j(x)\right)}$ and $p_j = \frac{\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]+\sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]+\sum_{j=1}^{n_e} \overline{c}_j(x)\right)}$, $j = \{1,\ldots,n_e\}$ in the last equality. By Lemma A.1, we have

$$
\begin{aligned}
&\sum_{j=0}^{n_e} p_j \ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j)\right)\\
&\geq \Gamma^{-1}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right)\right)\\
&= \Gamma^{-1}\left(\max_{j\in[n_e]} p_j - p_{\mathsf{r}(x)}\right)\\
&= \Gamma^{-1}\left(\frac{\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] - \min_{j=1}^{n_e} \overline{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)}\right)
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)]\\
&\geq n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)\Gamma^{-1}\left(\frac{\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] - \min_{j=1}^{n_e} \overline{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)}\right)\\
&\geq n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)\Gamma^{-1}\left(\frac{\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)]}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)}\right)\\
&\geq \frac{1}{\beta^{\frac{1}{\alpha}}}\frac{\left(\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)]\right)^{\frac{1}{\alpha}}}{\left(n_e(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j)\right)^{\frac{1}{\alpha}-1}}
\end{aligned}
$$

where we use the fact that $\Gamma(t) = \beta t^\alpha$, $\alpha \in (0,1]$, $\beta > 0$, $\mathsf{L} \leq \overline{l}$ and $c_j \leq \overline{c}_j$, $j = \{1,\ldots,n_e\}$ in the last inequality.

Case III: $\mathsf{r}(x) > 0$ and $\overline{c}_0(x) \leq \min_{j=1}^{n_e} \overline{c}_j(x)$. In this case, by (10) and (11), the calibration gap of the deferral loss can be expressed as

$$
\mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] = \overline{c}_{\mathsf{r}(x)}(x) - \overline{c}_0(x).
$$

By (12) and (13), the calibration gap of the surrogate loss can be expressed as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)]\\
&= \left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j) - (n_e-1)\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)]\\
&\quad - \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\overline{c}_0(x) + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\right] + (n_e-1)\overline{c}_0(x).
\end{aligned}
$$

Using the fact that $\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] \geq \inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] = \overline{c}_0(x)$, the calibration gap of the surrogate loss can be lower bounded as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)]\\
&\geq \left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\overline{c}_0(x) + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\\
&\quad - \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \overline{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\overline{c}_0(x) + \sum_{j'\neq j}^{n_e} \overline{c}_{j'}(x)\right)\ell(r,x,j)\right]\\
&= n_e\left(\sum_{j=0}^{n_e} \overline{c}_j(x)\right)\left[\sum_{j=0}^{n_e} p_j\ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j\ell(r,x,j)\right)\right]
\end{aligned}$$

where we let $p_0 = \frac{\sum_{j=1}^{n_e}\overline{c}_j(x)}{n_e\left(\sum_{j=0}^{n_e}\overline{c}_j(x)\right)}$ and $p_j = \frac{\overline{c}_0(x)+\sum_{j'\neq j}^{n_e}\overline{c}_{j'}(x)}{n_e\left(\sum_{j=0}^{n_e}\overline{c}_j(x)\right)}$, $j = \{1,\ldots,n_e\}$ in the last equality. By Lemma A.1, we have

$$\begin{aligned}
&\sum_{j=0}^{n_e} p_j\ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j\ell(r,x,j)\right)\\
&\geq \Gamma^{-1}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right)\right)\\
&= \Gamma^{-1}\left(\max_{j\in[n_e]} p_j - p_{\mathsf{r}(x)}\right)\\
&= \Gamma^{-1}\left(\frac{\overline{c}_{\mathsf{r}(x)}(x) - \overline{c}_0(x)}{n_e\left(\sum_{j=0}^{n_e}\overline{c}_j(x)\right)}\right)
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(h,r,x,y)]\\
&\geq n_e\left(\sum_{j=0}^{n_e} \overline{c}_j(x)\right)\Gamma^{-1}\left(\frac{\overline{c}_{\mathsf{r}(x)}(x) - \overline{c}_0(x)}{n_e\left(\sum_{j=0}^{n_e}\overline{c}_j(x)\right)}\right)\\
&= n_e\left(\sum_{j=0}^{n_e} \overline{c}_j(x)\right)\Gamma^{-1}\left(\frac{\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}}\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)]}{n_e\left(\sum_{j=0}^{n_e}\overline{c}_j(x)\right)}\right)\\
&\geq \frac{1}{\beta^{\frac{1}{\alpha}}}\frac{\left(\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)] - \inf_{h\in\mathcal{H},r\in\mathcal{R}}\mathbb{E}_{y|x}[L_{\mathrm{def}}(h,r,x,y)]\right)^{\frac{1}{\alpha}}}{\left(n_e(\overline{l} + \sum_{j=1}^{n_e}\overline{c}_j)\right)^{\frac{1}{\alpha}-1}}
\end{aligned}$$

where we use the fact that $\Gamma(t) = \beta t^\alpha$, $\alpha \in (0,1]$, $\beta > 0$, $\mathsf{L} \leq \overline{l}$ and $c_j \leq \overline{c}_j$, $j = \{1,\ldots,n_e\}$ in the last inequality.

Overall, by taking the expectation of the deferral and surrogate calibration gaps and using Jensen's inequality in each case, we obtain

$$\mathcal{E}_{L_{\mathrm{def}}}(h,r) - \mathcal{E}^*_{L_{\mathrm{def}}}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_{\mathrm{def}}}(\mathcal{H},\mathcal{R}) \leq \overline{\Gamma}\left(\mathcal{E}_{L_\ell}(h,r) - \mathcal{E}^*_{L_\ell}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_\ell}(\mathcal{H},\mathcal{R})\right).$$

where $\overline{\Gamma}(t) = \max\left\{t, \left(n_e\left(\overline{l} + \sum_{j=1}^{n_e}\overline{c}_j\right)\right)^{1-\alpha}\beta\, t^\alpha\right\}$. □

C. Proof of Theorem 4.1

Theorem 4.1. *Given a hypothesis set $\mathcal{R}$, a multi-class loss function $\ell \geq \ell_{0-1}$ and a predictor h. Assume that there exists a function $\Gamma(t) = \beta\, t^\alpha$ for some $\alpha \in (0,1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{L^h_{\text{def}}}(r) - \mathcal{E}^*_{L^h_{\text{def}}}(\mathcal{R}) + \mathcal{M}_{L^h_{\text{def}}}(\mathcal{R}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}) + \mathcal{M}_{L^h_\ell}(\mathcal{R})\Big),$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\bar{l} + \sum_{j=1}^{n_e} \bar{c}_j\big)\Big)^{1-\alpha} \beta\, t^\alpha$.

Proof. Given a hypothesis set $\mathcal{R}$, a multi-class loss function ℓ and a predictor h. For any $r \in \mathcal{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the conditional error of L^h_ℓ and L^h_{def} can be written as

$$\begin{aligned}
\mathop{\mathbb{E}}_{y|x}\Big[L^h_{\text{def}}(r,x,y)\Big] &= \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x,y)]1_{\mathsf{r}(x)=j} \\
\mathop{\mathbb{E}}_{y|x}\Big[L^h_\ell(r,x,y)\Big] &= \left(\sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x,y)]\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_{j'}(x,y)]\right)\ell(r,x,j).
\end{aligned} \tag{14}$$

Let $\bar{c}_0(x) = \inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\mathsf{L}(h(x),y)]$ and $\bar{c}_j(x) = \mathbb{E}_{y|x}[c(x,y)]$. Thus, the best-in class conditional error of of L^h_ℓ and L^h_{def} can be expressed as

$$\begin{aligned}
\inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}\Big[L^h_{\text{def}}(r,x,y)\Big] &= \min_{j\in[n_e]} \bar{c}_j(x) \\
\inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}\Big[L^h_\ell(r,x,y)\Big] &= \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \bar{c}_j(x)\right)\ell(r,x,0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \bar{c}_{j'}(x)\right)\ell(r,x,j)\right]
\end{aligned} \tag{15}$$

Let $p_0 = \frac{\sum_{j=1}^{n_e} \bar{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}$ and $p_j = \frac{\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \bar{c}_{j'}(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}$, $j = \{1,\ldots,n_e\}$. Then, the calibration gap of L^h_ℓ can be written as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}\Big[L^h_\ell(r,x,y)\Big] - \inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}\Big[L^h_\ell(r,x,y)\Big] \\
&= n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)\left[\sum_{j=0}^{n_e} p_j \ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j)\right)\right]
\end{aligned}$$

By Lemma A.1, we have

$$\begin{aligned}
\sum_{j=0}^{n_e} p_j \ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j)\right) &\geq \Gamma^{-1}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right)\right) \\
&= \Gamma^{-1}\left(\max_{j\in[n_e]} p_j - p_{\mathsf{r}(x)}\right) \\
&= \Gamma^{-1}\left(\frac{\bar{c}_{\mathsf{r}(x)}(x) - \min_{j\in[n_e]} \bar{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}\right).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_\ell(r,x,y)] - \inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_\ell(r,x,y)] \\
&\geq n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)\Gamma^{-1}\left(\frac{\bar{c}_{\mathsf{r}(x)}(x) - \min_{j\in[n_e]} \bar{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}\right) \\
&\geq \frac{1}{\beta^{\frac{1}{\alpha}}} \frac{\left(\mathbb{E}_{y|x}\Big[L^h_{\text{def}}(r,x,y)\Big] - \inf_{r\in\mathcal{R}} \mathbb{E}_{y|x}\Big[L^h_{\text{def}}(r,x,y)\Big]\right)^{\frac{1}{\alpha}}}{\left(n_e(\bar{l} + \sum_{j=1}^{n_e} \bar{c}_j)\right)^{\frac{1}{\alpha}-1}}
\end{aligned}$$

where we use the fact that $\Gamma(t) = \beta t^\alpha$, $\alpha \in (0,1]$, $\beta > 0$, $\mathsf{L} \leq \bar{l}$ and $c_j \leq \bar{c}_j$, $j = \{1,\ldots,n_e\}$ in the last inequality. Taking the expectation on both sides and using Jensen's inequality, we obtain

$$\mathcal{E}_{L^h_{\text{def}}}(r) - \mathcal{E}^*_{L^h_{\text{def}}}(\mathcal{R}) + \mathcal{M}_{L^h_{\text{def}}}(\mathcal{R}) \leq \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}) + \mathcal{M}_{L^h_\ell}(\mathcal{R})\Big).$$

where $\overline{\Gamma}(t) = \Big(n_e\big(\bar{l} + \sum_{j=1}^{n_e} \bar{c}_j\big)\Big)^{1-\alpha} \beta\, t^\alpha$. □

D. Proof of Theorem 4.3

Theorem 4.3. *Given a hypothesis set $\mathcal{H}$, a regular hypothesis set $\mathcal{R}$ and a multi-class loss function $\ell \geq \ell_{0-1}$. Assume that there exists a function $\Gamma(t) = \beta t^{\alpha}$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{R}$-consistency bound holds for all $r \in \mathcal{R}$ and any distribution,*

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

Then, the following $(\mathcal{H}, \mathcal{R})$-consistency bound holds for all $h \in \mathcal{H}$, $r \in \mathcal{R}$ and any distribution,

$$\mathcal{E}_{L_{\text{def}}}(h, r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) + \mathcal{M}_{L_{\text{def}}}(\mathcal{H}, \mathcal{R}) \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H}) + \overline{\Gamma}\Big(\mathcal{E}_{L^h_\ell}(r) - \mathcal{E}^*_{L^h_\ell}(\mathcal{R}) + \mathcal{M}_{L^h_\ell}(\mathcal{R})\Big),$$

where $\overline{\Gamma}(t) = \Big(n_e\Big(\bar{l} + \sum_{j=1}^{n_e} \bar{c}_j\Big)\Big)^{1-\alpha} \beta t^{\alpha}$.

Proof. The conditional error of the deferral loss can be expressed as

$$\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h, r, x, y)] = \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x, y)]1_{\mathsf{r}(x)=j}.$$

Let $\bar{c}_0(x) = \inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\mathsf{L}(h(x), y)]$ and $\bar{c}_j(x) = \mathbb{E}_{y|x}[c(x, y)]$. Thus, the best-in class conditional error of the deferral loss can be expressed as

$$\inf_{h\in\mathcal{H}, r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h, r, x, y)] = \min_{j\in[n_e]} \bar{c}_j(x).$$

Thus, by introducing the term $\min\big\{\mathbb{E}_{y|x}[\mathsf{L}(h(x), y)], \min_{j=1}^{n_e} \bar{c}_j(x)\big\}$ and subsequently subtracting it after rearranging, the conditional regret of the deferral loss L_{def} can be written as follows

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h, r, x, y)] - \inf_{h\in\mathcal{H}, r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}[L_{\text{def}}(h, r, x, y)] \\
&= \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x, y)]1_{\mathsf{r}(x)=j} - \min_{j\in[n_e]} \bar{c}_j(x) \\
&= \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x, y)]1_{\mathsf{r}(x)=j} - \min_{j=1}^{n_e} \bar{c}_j(x) + \left(\min_{j=1}^{n_e} \bar{c}_j(x) - \min_{j\in[n_e]} \bar{c}_j(x)\right).
\end{aligned} \tag{16}$$

Note that by the property of the minimum, the second term can be upper bounded as

$$\min_{j=1}^{n_e} \bar{c}_j(x) - \min_{j\in[n_e]} \bar{c}_j(x) \leq \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)] - \inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)].$$

Next, we will upper bound the first term. Note that the conditional error and the best-in class conditional error of L^h_ℓ can be expressed as

$$\begin{aligned}
L^h_\ell(r, x, y) &= \left(\sum_{j=1}^{n_e} \bar{c}_j(x)\right)\ell(r, x, 0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)] + \sum_{j'\neq j}^{n_e} \bar{c}_{j'}(x)\right)\ell(r, x, j) \\
\inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}\big[L^h_\ell(r, x, y)\big] &= \inf_{r\in\mathcal{R}}\left[\left(\sum_{j=1}^{n_e} \bar{c}_j(x)\right)\ell(r, x, 0) + \sum_{j=1}^{n_e}\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)] + \sum_{j'\neq j}^{n_e} \bar{c}_{j'}(x)\right)\ell(r, x, j)\right]
\end{aligned} \tag{17}$$

Let $p_0 = \frac{\sum_{j=1}^{n_e} \bar{c}_j(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}$ and $p_j = \frac{\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j'\neq j}^{n_e} \bar{c}_{j'}(x)}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)}$, $j = \{1, \ldots, n_e\}$. Then, the first term can be rewritten as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x, y)]1_{\mathsf{r}(x)=j} - \min_{j=1}^{n_e} \bar{c}_j(x) \\
&= n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x), y)] + \sum_{j=1}^{n_e} \bar{c}_j(x)\right)\left[\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq 0} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right)\right].
\end{aligned}$$

By Lemma A.1, we have

$$\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right) \leq \Gamma\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j) - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j \ell(r,x,j)\right)\right)$$
$$= \Gamma\left(\frac{L_\ell^h(r,x,y) - \inf_{r\in\mathcal{R}} \mathbb{E}_{y|x}\left[L_\ell^h(r,x,y)\right]}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)}\right).$$

Therefore, the first term can be upper bounded as

$$\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)]1_{\mathsf{r}(x)=0} + \sum_{j=1}^{n_e} \mathop{\mathbb{E}}_{y|x}[c_j(x,y)]1_{\mathsf{r}(x)=j} - \min_{j=1}^{n_e} \overline{c}_j(x)$$
$$= n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)\left[\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq 0} - \inf_{r\in\mathcal{R}}\left(\sum_{j=0}^{n_e} p_j 1_{\mathsf{r}(x)\neq j}\right)\right]$$
$$\leq n_e\left(\mathop{\mathbb{E}}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)\Gamma\left(\frac{L_\ell^h(r,x,y) - \inf_{r\in\mathcal{R}} \mathbb{E}_{y|x}\left[L_\ell^h(r,x,y)\right]}{n_e\left(\mathbb{E}_{y|x}[\mathsf{L}(h(x),y)] + \sum_{j=1}^{n_e} \overline{c}_j(x)\right)}\right)$$
$$\leq \left(n_e\left(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\right)\right)^{1-\alpha} \beta\left(L_\ell^h(r,x,y) - \inf_{r\in\mathcal{R}} \mathop{\mathbb{E}}_{y|x}\left[L_\ell^h(r,x,y)\right]\right)^{\alpha}$$

where we use the fact that $\Gamma(t) = \beta t^\alpha$, $\alpha \in (0,1]$, $\beta > 0$, $\mathsf{L} \leq \overline{l}$ and $c_j \leq \overline{c}_j$, $j = \{1,\ldots,n_e\}$ in the last inequality. After upper bounding the first term and the second term in (16) as above, taking the expectation on both sides and using Jensen's inequality, we obtain

$$\mathcal{E}_{L_{\text{def}}}(h,r) - \mathcal{E}^*_{L_{\text{def}}}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_{\text{def}}}(\mathcal{H},\mathcal{R}) \leq \mathcal{E}_{\mathsf{L}}(h) - \mathcal{E}_{\mathsf{L}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}}(\mathcal{H})$$
$$+ \overline{\Gamma}\left(\mathcal{E}_{L_\ell^h}(r) - \mathcal{E}^*_{L_\ell^h}(\mathcal{R}) + \mathcal{M}_{L_\ell^h}(\mathcal{R})\right),$$

where $\overline{\Gamma}(t) = \left(n_e\left(\overline{l} + \sum_{j=1}^{n_e} \overline{c}_j\right)\right)^{1-\alpha} \beta\, t^\alpha$. □

E. Common margin-based losses and corresponding deferral surrogate losses

Table 2. Common margin-based losses and corresponding deferral surrogate losses.

Name	$\Phi(u)$	Deferral surrogate loss ℓ_Φ
Exponential	$\Phi_{\exp}(u) = e^{-u}$	$\mathsf{L}(h(x),y)e^{r(x)} + c(x,y)e^{-r(x)}$
Logistic	$\Phi_{\log}(u) = \log(1+e^{-u})$	$\mathsf{L}(h(x),y)\log\left(1+e^{r(x)}\right) + c(x,y)\log\left(1+e^{-r(x)}\right)$
Quadratic	$\Phi_{\text{quad}}(u) = \max\{1-u,0\}^2$	$\mathsf{L}(h(x),y)\Phi_{\text{quad}}(-r(x)) + c(x,y)\Phi_{\text{quad}}(r(x))$
Hinge	$\Phi_{\text{hinge}}(u) = \max\{1-u,0\}$	$\mathsf{L}(h(x),y)\Phi_{\text{hinge}}(-r(x)) + c(x,y)\Phi_{\text{hinge}}(r(x))$
Sigmoid	$\Phi_{\text{sig}}(u) = 1-\tanh(ku), k>0$	$\mathsf{L}(h(x),y)\Phi_{\text{sig}}(-r(x)) + c(x,y)\Phi_{\text{sig}}(r(x))$
ρ-Margin	$\Phi_\rho(u) = \min\left\{1,\max\left\{0,1-\frac{u}{\rho}\right\}\right\}, \rho>0$	$\mathsf{L}(h(x),y)\Phi_\rho(-r(x)) + c(x,y)\Phi_\rho(r(x))$

Table 3. Comparison of our proposed method with three simple baselines.

Baseline 1			Baseline 2			Baseline 3			Ours		
EXP 1	EXP 2	EXP 3	EXP 1	EXP 1, 2	EXP 1, 2, 3	EXP 1	EXP 2	EXP 3	EXP 1	EXP 1, 2	EXP 1, 2, 3
17.37 ± 4.80	15.07 ± 3.03	12.72 ± 2.30	17.77 ± 5.12	15.43 ± 2.83	12.92 ± 2.45	16.26 ± 5.58	15.44 ± 2.25	12.36 ± 3.32	$\mathbf{16.26 \pm 5.58}$	$\mathbf{14.82 \pm 3.60}$	$\mathbf{12.02 \pm 1.97}$

F. Additional experiments

Here, we report additional experimental results with three simple baselines:

- Baseline 1: The accuracy of the expert.
- Baseline 2: Always defer to one expert (random or not random) with probability a%.
- Baseline 3: Single-expert formulation using only expert 1 (or 2, or 3).

In Table 3, we report the empirical results of our two-stage method without base cost on the Housing dataset alongside the corresponding baselines, which further demonstrates our approach's effectiveness. For our method, the single-expert deferral ratio is 91%, the two-expert deferral rate is 8% for the first expert and 85% for the second expert, and the three-expert deferral rate is 4% for the first expert, 35% for the second expert, and 60% for the third expert. We use the same deferral rate for randomly deferring to experts in Baseline 2. The error of the base model is 22.72 ± 7.68. EXP represents the expert used, and system MSE values are reported. Clearly, our method outperforms all three baselines.