Adversarial Robustness in One-Stage Learning-to-Defer

Yannis Montreuil[*,1,4,5] **Letian Yu**[*,1] **Axel Carlier**[2,4]

Lai Xing Ng[3,4] **Wei Tsang Ooi**[1,4]

[1]School of Computing, National University of Singapore, Singapore
[2]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France
[3]Institute for Infocomm Research, A*STAR , Singapore
[4]IPAL, IRL 2955, Singapore
[5]CNRS@CREATE LTD, 1 Create Way, Singapore

Abstract

Learning-to-Defer (L2D) enables hybrid decision-making by routing inputs either to a predictor or to external experts. While promising, L2D is highly vulnerable to adversarial perturbations, which can not only flip predictions but also manipulate deferral decisions. Prior robustness analyses focus solely on two-stage settings, leaving open the end-to-end (one-stage) case where predictor and allocation are trained jointly. We introduce the first framework for adversarial robustness in one-stage L2D, covering both classification and regression. Our approach formalizes attacks, proposes cost-sensitive adversarial surrogate losses, and establishes theoretical guarantees including $\mathcal{H}$, $(\mathcal{R}, \mathcal{F})$, and Bayes consistency. Experiments on benchmark datasets confirm that our methods improve robustness against untargeted and targeted attacks while preserving clean performance.

1 INTRODUCTION

With the growing adoption of routing-based methods, *Learning-to-Defer* (L2D) has emerged as a principled framework for hybrid decision-making (Madras et al., 2018; Mozannar and Sontag, 2020). An L2D system may either predict directly or defer to an expert, thereby trading off predictive performance against selective reliance on external decision-makers. This framework is particularly relevant in safety-critical applications (Joshi et al., 2023; Strong et al., 2025; Palomba et al., 2025) and also offers a unifying perspective on a broad class of routing problems (Chen et al., 2024; Jitkrittum et al., 2026), where the central challenge is to design allocation policies that distribute decisions optimally across multiple agents.

Yet, despite its promise, L2D inherits the adversarial vulnerabilities of standard machine learning models (Goodfellow et al., 2014; Madry et al., 2017; Jia and Liang, 2017) while introducing new ones unique to the routing setting. Recent work (Montreuil et al., 2025a) shows that L2D systems can be even more fragile to adversarial perturbations than conventional systems. An adversary may not only induce misclassification but also manipulate whether, and to whom, the system defers—for example, bypassing a reliable expert or forcing deferral to one known to perform poorly. Such attacks threaten both predictive performance and the reliability of decision allocation, raising pressing concerns for deployment in safety-critical environments.

Existing robustness studies focus exclusively on the simplified *two-stage* L2D setting, where the predictor and experts are trained offline and only the allocation policy is learned (Narasimhan et al., 2022; Mao et al., 2023a, 2024c; Montreuil et al., 2025b,a, 2026a). This formulation is convenient but fails to capture the complexity of end-to-end (*one-stage*) L2D, in which the predictor and allocation policy must be optimized jointly. We address two problems: (i) characterizing attacks that target one-stage L2D systems, and (ii) designing one-stage L2D methods with guarantees for the proposed adversarial losses.

In this work, we develop a framework for adversarial robustness in one-stage L2D, encompassing both clas-

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

sification and regression. Our contributions are threefold:

1. We introduce both untargeted and targeted attacks that reveal how adversaries can manipulate prediction and deferral simultaneously;
2. We propose new outcome-wise adversarial surrogate deferral losses with tractable relaxations and establish theoretical guarantees for the proposed adversarial losses, including $\mathcal{H}$-consistency for classification and $(\mathcal{R}, \mathcal{F})$-consistency for regression;
3. We demonstrate empirically, on image classification and tabular regression benchmarks, that our methods substantially improve robustness against both attack types while maintaining competitive clean accuracy.

Together, these results provide a theoretically grounded and practically effective approach to adversarially robust one-stage L2D.

2 RELATED WORK

One-Stage L2D. Madras et al. (2018) introduced the first formal framework for Learning-to-Defer (L2D), incorporating expert predictions via a defer-to-expert mechanism. A key advance was made by Mozannar and Sontag (2020), who proposed a *score-based* formulation of L2D. In this approach, the classifier is augmented with a shared scoring function that jointly determines both prediction and allocation, thereby unifying two decisions that had previously been treated separately.

Subsequent work has pursued several directions: strengthening theoretical guarantees such as $\mathcal{H}$-consistency and realizability (Mozannar et al., 2023; Mao et al., 2024a,b, 2025); broadening the class of statistically sound surrogate losses (Charusaie et al., 2022; Mao et al., 2024a); improving estimation methods (Verma et al., 2022, 2023; Cao et al., 2024); and extending policies to top-k expert selection (Montreuil et al., 2026a). The score-based methodology has also been applied across diverse classification domains (Keswani et al., 2021; Kerrigan et al., 2021; Hemmer et al., 2022; Benz and Rodriguez, 2022; Tailor et al., 2024; Liu et al., 2024; Montreuil et al., 2026c,b). Regression-based variants have been proposed by Mao et al. (2024c), who employ a dedicated allocation policy alongside a trainable predictor.

Robustness in L2D. Robustness has received comparatively little attention in the L2D literature. To the best of our knowledge, the only prior work addressing robustness is that of Montreuil et al. (2025a), who introduced an adversarially consistent formulation. Their analysis, however, is restricted to the two-stage setting, where experts are fixed and not jointly optimized with the allocation policy. This assumption simplifies the problem: robustness reduces to modifying the allocation strategy without accounting for the interaction between learning the predictor and adapting to the experts.

By contrast, the one-stage setting presents a qualitatively harder challenge as it requires jointly learning the predictor as well as the allocation policy. In this work, we extend the robustness framework of Montreuil et al. (2025a) to the technically more demanding one-stage L2D scenario, where adversarial perturbations may influence all components of the system simultaneously.

3 PRELIMINARIES

Task and data. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and let $\mathcal{Z}$ denote the target space, with

$$\mathcal{Z} = \begin{cases} \mathcal{Y} = \{1, \dots, K\}, & \text{classification,} \\ \mathcal{T} \subseteq \mathbb{R}^m, & \text{regression.} \end{cases}$$

Each training label satisfies $z \in \mathcal{Z}$. We assume access to i.i.d. samples $\mathcal{S}_n = \{(x_i, z_i, \mathbf{m}_i)\}_{i=1}^n \sim \big(\mathcal{D} \times \mathcal{M}\big)^n$, where $(x, z) \sim \mathcal{D}$ is drawn from the underlying data distribution and $\mathbf{m} \sim \mathcal{M} \mid (x, z)$ denotes the outputs of the J fixed experts conditioned on (x, z) (Madras et al., 2018; Mozannar and Sontag, 2020; Verma et al., 2022). Here $\mathbf{m} = (m_1, \dots, m_J)$, with each m_j belonging to the appropriate output space: in classification, $m_j \in \mathcal{Y}$ is a categorical prediction, while in regression, $m_j \in \mathcal{T}$ is a real-valued prediction. All expert predictions can thus be regarded as elements of the common target space $\mathcal{Z}$.

Classification L2D. We specialize to $\mathcal{Z} = \mathcal{Y} = \{1, \dots, K\}$. Each training label $y_i \in \mathcal{Y}$ and each expert output $m_{ij} \in \mathcal{Y}$ is categorical. Alongside the J experts $m_1, \dots, m_J$, we train a hypothesis $h \in \mathcal{H}$ with $h : \mathcal{X} \to \mathbb{R}^{K+J}$, where the first K coordinates correspond to class scores and the last J to expert scores (Mozannar and Sontag, 2020; Cao et al., 2024; Mao et al., 2024a; Montreuil et al., 2026a). Let $\mathcal{A}^c = \mathcal{Y} \cup \{K+1, \dots, K+J\}$ denote the augmented action space. For a vector $\xi(x)$, we write $\hat{\xi}(x)$ to denote its arg max index. The induced decision rule is then $\hat{h}(x) = \arg\max_{k \in \mathcal{A}^c} h(x)_k$, with ties broken uniformly at random. If $\hat{h}(x) \in \mathcal{Y}$, the system predicts the corresponding category; if $\hat{h}(x) = K + j$ for some $j \in \{1, \dots, J\}$, the system defers to expert m_j.

The predictor $h \in \mathcal{H}$ is trained to minimize the expected risk of the *true deferral loss* for classification

$$\ell^c_{\text{def}}(\hat{h}(x), y, \mathbf{m}) = \begin{cases} \mathbf{1}\{\hat{h}(x) \neq y\}, & \text{if } \hat{h}(x) \in \mathcal{Y}, \\ c^c_j(m_j, y), & \text{if } \hat{h}(x) = K + j. \end{cases} \tag{1}$$

where $c^c_j(m_j, y) \in [0, 1]$ is the cost of deferring to expert m_j. We set $c^c_j(m_j, y) = \alpha_j \, \mathbf{1}\{m_j \neq y\} + \beta_j$, with $\alpha_j \geq 0$ a scaling coefficient and β_j the fixed consultation cost for querying expert j.

A common way to minimize such a discontinuous loss is to replace it with a consistent surrogate (Steinwart, 2007; Zhang, 2002; Bartlett et al., 2006; Awasthi et al., 2022), ensuring that the learned predictor approximates the Bayes classifier $h^B \in \mathcal{H}$ that minimizes the expected true deferral loss. Given the augmented action space $\mathcal{A}^c$, we define the *surrogate deferral loss* for classification as

$$\begin{aligned} \Phi^{c,u}_{\text{def}}(h(x), y, \mathbf{m}) \;&=\; \Phi^u_{\text{cls}}(h(x), y) \\ &\quad + \sum_{j=1}^{J} (1 - c^c_j(m_j, y)) \, \Phi^u_{\text{cls}}(h(x), K + j), \end{aligned} \tag{2}$$

where Φ^u_{cls} is a classification surrogate from the cross-entropy family (Mao et al., 2023b), defined as $\Phi^u_{\text{cls}}(h(x), y) = \Psi^u\Big(\sum_{y' \in \mathcal{Y}} \exp(h(x)_{y'} - h(x)_y) - 1\Big)$ with outer transform

$$\Psi^u(v) = \begin{cases} \log(1 + v), & u = 1, \\ \frac{1}{1-u}\big[(1+v)^{1-u} - 1\big], & u \neq 1. \end{cases} \tag{3}$$

This family recovers well-known surrogates, including logistic loss (Ohn Aldrich, 1997), generalized cross-entropy (Zhang and Sabuncu, 2018), and mean absolute error (Ghosh et al., 2017). The surrogate risk is $\mathcal{E}_{\Phi^{c,u}_{\text{def}}}(h) = \mathbb{E}[\Phi^{c,u}_{\text{def}}(h(X), Y, M)]$ with optimal value $\mathcal{E}^*_{\Phi^{c,u}_{\text{def}}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{E}_{\Phi^{c,u}_{\text{def}}}(h)$. A surrogate is called *consistent* if minimizing its excess risk also minimizes the true excess risk (Zhang, 2002; Bartlett et al., 2006; Steinwart, 2007; Tewari and Bartlett, 2007).

Theorem 3.1 ($\mathcal{H}$-consistency bounds (Awasthi et al., 2022))**.** *The surrogate $\Phi^{c,u}_{def}$ is $\mathcal{H}$-consistent with respect to ℓ^c_{def} if there exists a non-decreasing function $\Gamma^u : \mathbb{R}^+ \to \mathbb{R}^+$ such that for every distribution $\mathcal{D}$,*

$$\begin{aligned} &\mathcal{E}_{\ell^c_{\text{def}}}(h) - \mathcal{E}^B_{\ell^c_{\text{def}}}(\mathcal{H}) + \mathcal{U}_{\ell^c_{\text{def}}}(\mathcal{H}) \leq \\ &\qquad \Gamma^u\Big(\mathcal{E}_{\Phi^{c,u}_{\text{def}}}(h) - \mathcal{E}^*_{\Phi^{c,u}_{\text{def}}}(\mathcal{H}) + \mathcal{U}_{\Phi^{c,u}_{\text{def}}}(\mathcal{H})\Big). \end{aligned}$$

Here $\mathcal{U}_{\Phi^{c,u}_{\text{def}}}(\mathcal{H})$ is the *minimizability gap*, quantifying the discrepancy between the best achievable excess risk within $\mathcal{H}$ and the expected pointwise minimum. This gap vanishes when $\mathcal{H} = \mathcal{H}_{\text{all}}$, recovering Bayes-consistency in the asymptotic limit (Zhang, 2002; Steinwart, 2007).

Regression L2D. Let $\mathcal{Z} = \mathcal{T} \subseteq \mathbb{R}^m$. Each training label $t_i \in \mathcal{T}$ and each expert output $m_{ij} \in \mathcal{T}$ is a real-valued vector. The system consists of a predictor $f \in \mathcal{F}$, with $f : \mathcal{X} \to \mathcal{T}$, and a rejector $r \in \mathcal{R}$, with $r : \mathcal{X} \to \mathbb{R}^{J+1}$. Let $\mathcal{A}^r = \{1, 2, \ldots, J+1\}$ denote this action space. The induced decision rule is $\hat{r}(x) = \arg\max_{k \in \mathcal{A}^r} r(x)_k$, so that if $\hat{r}(x) = 1$, the system outputs $f(x)$, while if $\hat{r}(x) = j$ with $j \geq 2$, the system defers to expert m_{j-1}. The *true deferral loss* for regression is

$$\ell^r_{\text{def}}(f(x), \hat{r}(x), t, \mathbf{m}) = \sum_{j=1}^{J+1} c^r_j(f(x), m_{j-1}, t) \mathbf{1}\{\hat{r}(x) = j\}, \tag{4}$$

where each $c^r_j(f(x), m_{j-1}, t) \geq 0$ is the cost of action j. For the predictor we define $c^r_1(f(x), t) = \alpha_1 \, L(f(x), t) + \beta_1$, and for expert $j \geq 2$, $c^r_j(m_{j-1}, t) = \alpha_j \, L(m_{j-1}, t) + \beta_j$, where $L : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ is a regression loss (e.g. squared error). Similarly to the classification case, we approximate the discontinuous loss with a consistent surrogate, which we refer to as the *surrogate deferral loss* for regression:

$$\begin{aligned} \Phi^{r,u}_{\text{def}}(f(x), r(x), t, \mathbf{m}) = &\sum_{j=1}^{J+1} \tau^r_j(f(x), \mathbf{m}, t) \, \Phi^u_{\text{cls}}(r(x), j) \\ &- (J-1) \, c^r_1(f(x), t), \end{aligned} \tag{5}$$

where Φ^u_{cls} is the same cross-entropy–type surrogate used in classification, and the weights $\tau^r_j(f(x), \mathbf{m}, t)$ are defined as $\tau^r_j(f(x), \mathbf{m}, t) = \sum_{i \neq j} c^r_i(f(x), m_{i-1}, t)$. This surrogate has been shown to be both Bayes consistent and $(\mathcal{R}, \mathcal{F})$-consistent (Mao et al., 2024c).

Robustness. Adversarially robust classification aims to train classifiers that remain reliable under small, often imperceptible, perturbations of the input (Goodfellow et al., 2014; Madry et al., 2017). The goal is to minimize the *true multiclass loss* ℓ_{01} evaluated on an adversarial input $x' = x + \delta$ (Gowal et al., 2020; Awasthi et al., 2022). A perturbation δ is constrained by its magnitude, with the adversarial region around x defined as $B_p(x, \gamma) = \{\, x' \in \mathbb{R}^d \mid \|x' - x\|_p \leq \gamma \,\}$, where $\|\cdot\|_p$ denotes the p-norm and $\gamma > 0$ bounds the perturbation size.

The *adversarial* 0–1 *loss* is $\widetilde{\ell}_{01}(h, x, y) = \sup_{x' \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x') \neq y\}$, and its *adversarial margin surrogate* is

$$\widetilde{\ell}_{01}(h, x, y) \leq \sup_{x' \in B_p(x,\gamma)} \Phi^{\rho,u}_{\text{cls}}(h(x'), y), \tag{6}$$

with $\Phi^{\rho,u}_{\text{cls}}(h(x'), y) = \Psi^u\Big(\sum_{k \neq y} \Psi_\rho\big(h(x')_k - h(x')_y\big)\Big)$. Here, Ψ^u and Ψ_ρ are transformations that characterize the surrogate family. In the analysis below, we

use the exponential ρ-soft margin transform $\Psi_\rho(v) = \exp(-v/\rho)$. Recent works show that algorithms based on smooth, regularized variants of these comp-sum ρ-margin losses achieve strong calibration and consistency guarantees (Awasthi et al., 2022, 2023; Mao et al., 2023b).

4 ATTACKING ONE-STAGE LEARNING-TO-DEFER APPROACHES

As discussed in Section 2, robustness has so far been investigated only in the two-stage setting of Learning-to-Defer (Montreuil et al., 2025a), while the one-stage formulation remains unexplored. Addressing robustness in this setting requires novel surrogate losses that simultaneously preserve consistency and provide robustness guarantees.

Learning-to-Defer seeks to route each query to the most reliable agent—either the predictor or one of the experts (Madras et al., 2018; Mozannar and Sontag, 2020). In classification, the augmented classifier jointly handles both prediction and deferral. In regression, allocation and prediction are decoupled. Such formulations perform well on clean inputs (Verma et al., 2023; Mozannar et al., 2023). The natural question is: do they remain effective under noisy or adversarial perturbations? We answer this in the negative. Extending the conclusions of Montreuil et al. (2025a), we show that adversarial attacks not only compromise the allocation policy but also predictive performance in both the regression and classification settings of one-stage Learning-to-Defer.

4.1 Introducing Attacks

Untargeted Attack. In the *untargeted* setting, the adversary seeks a perturbation $\delta \in B_p(0,\gamma)$ that maximally degrades the system's performance, without targeting a specific outcome.

Definition 4.1 (Untargeted Attack)**.** *An* untargeted adversarial attack *in L2D is the problem of finding a perturbed input $x' = x + \delta$, with $\delta \in B_p(0,\gamma)$, that maximizes a differentiable attack objective aligned with the deferral loss in either classification or regression. In our experiments, we instantiate this objective with the corresponding surrogate loss. Formally,*

$$x' = \underset{x' \in B_p(x,\gamma)}{\arg\max}\ \Phi^{\bullet,u}_{\text{def}}(x'),$$

where $\bullet \in \{c, r\}$ indicates whether the task is classification (c) or regression (r).

The attacker perturbs an input x so that, even if the clean input would be processed optimally, the perturbed version x' induces a worse outcome under the deferral policy. Such degradation can occur in several ways in the one-stage setting: (i) causing the predictor to output an incorrect class in classification or a high-error estimate in regression; (ii) obstructing deferral to a reliable expert when such deferral would reduce error; or (iii) inducing unnecessary deferral when the predictor is already sufficiently accurate. For instance, if the clean input x should be handled by the main predictor, an untargeted perturbation may instead redirect the query to an inappropriate expert, thereby increasing the overall system error.

Targeted Attack. In the *targeted* setting, the adversary specifies a desired outcome ν and perturbs the input so that the system's decision is steered toward this target. In classification, ν may correspond to a class label, $\nu \in \{1,\dots,K\}$, or to a deferral action, $\nu \in \{K+1,\dots,K+J\}$. In regression, the target space is defined analogously: $\nu = 1$ denotes trusting the predictor's output, while $\nu \in \{2,\dots,J+1\}$ corresponds to deferral to expert ν. Let $\boldsymbol{\delta} = (\delta_1,\dots,\delta_{|\mathcal{A}|})$ denote the family of outcome-specific perturbations, with $\delta_j \in \mathcal{X}$.

Definition 4.2 (Targeted Attack)**.** *A* targeted adversarial attack *in L2D is the problem of finding a perturbation $\delta_\nu \in \mathcal{X}$ from the family $\boldsymbol{\delta}$ such that the adversarial input $x'_\nu = x + \delta_\nu$ lies within the p-norm ball $B_p(x,\gamma)$ and drives the allocation policy toward the specified target $\nu \in \mathcal{A}$. Formally,*

$$x'_\nu = \underset{x'_\nu \in B_p(x,\gamma)}{\arg\min}\ \Phi^{u}_{\text{cls}}\big(\pi(x'_\nu), \nu\big),$$

where π is h in classification or r in regression.

Unlike untargeted attacks, which aim only to degrade performance, targeted attacks explicitly redirect the decision to a chosen outcome ν. The surrogate Φ^u_{cls} acts as a differentiable proxy for the indicator $\mathbf{1}\{\hat{\pi}(x'_\nu) \neq \nu\}$, thereby encouraging $\hat{\pi}(x'_\nu) = \nu$. Such attacks are particularly powerful: (i) in classification, an adversary may compel predictions into sensitive categories of $\mathcal{Y}$; and (ii) the adversary may force deferral to a specific expert—possibly one known to perform poorly or to exhibit exploitable dependencies. For example, even if the clean input x should defer to expert 1, a targeted perturbation may redirect the system to predict an unrelated class or to defer to a less reliable expert.

Remark 1. In the two-stage case (Montreuil et al., 2025a), both targeted and untargeted attacks focus on the allocation policy, whereas in our setting adversaries can attack both the predictive and allocation policies.

5 DEFENDING AGAINST ATTACKS WITH GUARANTEES

As previously discussed, current Learning-to-Defer approaches are highly sensitive to adversarial perturbations. This motivates the development of a formulation that explicitly incorporates adversarial robustness. Building on Montreuil et al. (2025a), we generalize and extend the analysis to the one-stage L2D setting and introduce *outcome-wise adversarial deferral losses* for both classification and regression.

5.1 Classification Setting

5.1.1 True Loss and Surrogates

We now move from the one-stage attack setting to optimization objectives that enable learning a robust hypothesis $h \in \mathcal{H}$. To this end, we define an outcome-wise adversarial deferral loss by distinguishing perturbations according to their decision outcomes. This quantity upper-bounds the conventional single-perturbation adversarial deferral loss.

Given an input x, let $x'_j \in B_p(x,\gamma)$ denote the adversarial example associated with outcome $j \in \mathcal{A}$, written as $x'_j = x + \delta_j$ where $\delta_j \in \mathcal{X}$ is the perturbation leading to outcome j. Collecting these perturbations, we write $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{K+J})$ for the family of outcome-specific perturbations. For example, x'_1 denotes the adversarial input aligned with outcome $1 \in \mathcal{A}$.

Lemma 5.1 (Outcome-wise Adversarial Deferral Loss for Classification)**.** *For any hypothesis $h \in \mathcal{H}$, input–label pair (x,y), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* outcome-wise adversarial deferral loss for classification *is*

$$\widetilde{\ell}^{c}_{\mathrm{def}}\big(h,x,y,\mathbf{m}\big) = \sum_{j=1}^{K+J} \mu_j(j,m_j,y) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) = j\}.$$

Here, μ_j denotes a shifted cost. For $j \in \{1,\ldots,K\}$, we set $\mu_j(j,y) = \alpha_j \mathbf{1}\{j \neq y\} + \beta_j$, which encodes the cost of predicting class j. For $j \in \{K+1,\ldots,K+J\}$, we define $\mu_j(m_j,y) = c_{j-K}(m_{j-K},y)$, capturing the cost of deferring to expert j.

We prove in Appendix A.3 that this quantity upper-bounds the conventional single-perturbation adversarial deferral loss, and therefore also upper-bounds the clean deferral loss in classification (Equation 1). Since this formulation is NP-hard, we next introduce a tractable surrogate.

Definition 5.2 (Adversarial Surrogate Deferral Loss for Classification)**.** *For any hypothesis $h \in \mathcal{H}$, input–label pair (x,y), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* adversarial surrogate deferral loss for classification *is defined as*

$$\widetilde{\Phi}^{c,u}_{\mathrm{def}}(h,x,y,\mathbf{m}) = \sum_{j=1}^{K+J} \sum_{i \neq j} \mu_i(i,m_i,y) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{\mathrm{cls}}(h(x'_j),j).$$

Here, $\Phi^{\rho,u}_{\mathrm{cls}}(h(x'_j),j)$ denotes the margin-based surrogate loss evaluated on the adversarial input x'_j associated with outcome $j \in \mathcal{A}$.

5.1.2 Theoretical Guarantees

$\mathcal{H}$-consistency. The surrogate in Definition 5.2 is a tractable relaxation of the outcome-wise adversarial loss in Lemma 5.1. A key requirement is that it be both Bayes- and $\mathcal{H}$-consistent with respect to the defined outcome-wise adversarial loss, ensuring that minimization recovers an optimal allocation policy for this loss. We establish these guarantees in the sense of Zhang (2002); Steinwart (2007), showing that optimization over the surrogate yields asymptotically optimal policies for the proposed adversarial loss.

Theorem 5.3 ($\mathcal{H}$-consistency bounds of $\widetilde{\Phi}^{c,u}_{\mathrm{def}}$)**.** *Let $\mathcal{H}$ be symmetric and locally ρ-consistent. Then, for the set $\mathcal{A}$, any hypothesis $h \in \mathcal{H}$, and any distribution $\mathcal{D}$, the following holds:*

$$\mathcal{E}_{\widetilde{\ell}^{c}_{def}}(h) - \mathcal{E}^{B}_{\widetilde{\ell}^{c}_{def}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\ell}^{c}_{def}}(\mathcal{H}) \leq \Psi^{u}(1)\Big(\mathcal{E}_{\widetilde{\Phi}^{c,u}_{def}}(h) - \mathcal{E}^{*}_{\widetilde{\Phi}^{c,u}_{def}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\Phi}^{c,u}_{def}}(\mathcal{H})\Big),$$

with $\Psi^{u}(1) = \begin{cases} \log(2), & u = 1, \\ \frac{1}{1-u}\big(2^{1-u}-1\big), & u \neq 1. \end{cases}$

We prove this theorem using novel proof techniques, provided in Appendix A.4. Intuitively, Theorem 5.3 shows that the surrogate loss $\widetilde{\Phi}^{c,u}_{\mathrm{def}}$ is $\mathcal{H}$-consistent up to the multiplicative constant $\Psi^{u}(1)$. If a sequence $(h_t) \subset \mathcal{H}$ satisfies

$$\mathcal{E}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(h_t) - \mathcal{E}^{*}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(\mathcal{H}) \to 0,$$

then by Theorem 5.3 we have $\mathcal{E}_{\widetilde{\ell}^{c}_{\mathrm{def}}}(h_t) - \mathcal{E}^{B}_{\widetilde{\ell}^{c}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\ell}^{c}_{\mathrm{def}}}(\mathcal{H}) \to 0$. Thus, minimizing the surrogate excess risk directly controls the excess risk of the defined outcome-wise adversarial deferral loss, ensuring that optimization with respect to $\widetilde{\Phi}^{c,u}_{\mathrm{def}}$ converges to the Bayes-optimal policy for this loss. Finally, note that the minimizability gap vanishes for realizable distributions or when $\mathcal{H}$ is the set of all measurable functions $\mathcal{H}_{\mathrm{all}}$ (Steinwart, 2007).

Relaxing non-convexity. While we established the consistency of the surrogate introduced in Definition 5.2, the formulation remains non-convex due

to its reliance on the margin-based surrogate $\widetilde{\Phi}^{\rho,u}_{\text{cls}}$. Non-convexity complicates optimization and limits the practical applicability of the guarantee. To address this, we construct a smooth upper bound by replacing the margin loss with a tractable relaxation, following the principled approaches of Awasthi et al. (2023); Mao et al. (2023b); Montreuil et al. (2025a). This relaxation preserves theoretical soundness while enabling tractable and stable optimization in practice.

Lemma 5.4 (Smooth Adversarial Surrogate Losses)**.** *Let $x \in \mathcal{X}$ be a clean input, and let $\rho > 0$ and $\kappa > 0$ be hyperparameters. The smooth adversarial surrogate losses are defined as*

$$\widetilde{\Phi}^{u}_{\text{cls,s}}(h,x,j) = \Phi^{u}_{\text{cls}}(h(x)/\rho, j) + \kappa \sup_{x'_j \in B_p(x,\gamma)} \left\| \overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x,j) \right\|_2,$$

We defer the proof in Appendix A.2. For any $x \in \mathcal{X}$, we define the pairwise margin differences as $\Delta_h(x,j,j') = h(x)_j - h(x)_{j'}$, and let $\overline{\Delta}_h(x,j) \in \mathbb{R}^{|\mathcal{A}|-1}$ be the vector of all pairwise differences, i.e., $\overline{\Delta}_h(x,j) = \big(\Delta_h(x,j,1), \ldots, \Delta_h(x,j,j-1), \Delta_h(x,j,j+1), \ldots, \Delta_h(x,j,K+J)\big)$. The first term, $\Phi^{u}_{\text{cls}}(h(x)/\rho, j)$, is the standard multiclass surrogate loss scaled by ρ, while the second term penalizes local instability by bounding the worst-case deviation of pairwise margins under adversarial perturbations. This yields a smooth tractable relaxation of the surrogate in Definition 5.2.

Definition 5.5 (Smooth Adversarial Surrogate Deferral Loss for Classification)**.** *For any hypothesis $h \in \mathcal{H}$, input–label pair (x,y), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* smooth adversarial surrogate deferral loss for classification *is defined as*

$$\widetilde{\Phi}^{c,u}_{\text{def,s}}(h,x,y,\mathbf{m}) = \sum_{j=1}^{K+J} \sum_{i \neq j} \mu_i(i, m_i, y) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{u}_{\text{cls,s}}(h(x'_j), j).$$

which yields the following pointwise upper bound.

Corollary 5.6 (Pointwise Upper Bound for $\widetilde{\Phi}^{c,u}_{\text{def,s}}$)**.** *For any hypothesis $h \in \mathcal{H}$, input–label pair (x,y), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the following holds:*

$$\widetilde{\Phi}^{c,u}_{\text{def}}(h,x,y,\mathbf{m}) \leq \widetilde{\Phi}^{c,u}_{\text{def,s}}(h,x,y,\mathbf{m}).$$

Corollary 5.6 shows that the smooth surrogate $\widetilde{\Phi}^{c,u}_{\text{def,s}}$ pointwise upper-bounds $\widetilde{\Phi}^{c,u}_{\text{def}}$. Combined with Theorem 5.3, this motivates optimizing the smooth relaxation as a tractable proxy for the nonsmooth adversarial surrogate.

Definition 5.7 (RERM-C: Regularized ERM for $\widetilde{\Phi}^{c,u}_{\text{def,s}}$)**.** *Assume $\mathcal{H}$ is symmetric and locally ρ-consistent. Let $\Omega : \mathcal{H} \to \mathbb{R}^+$ be a regularizer and let $\eta > 0$ be a hyperparameter. We define the regularized empirical risk minimization (ERM) objective as*

$$\min_{h \in \mathcal{H}} \left[\frac{1}{n} \sum_{k=1}^{n} \widetilde{\Phi}^{c,u}_{\text{def,s}}\big(h, x_k, y_k, \mathbf{m}_k\big) + \eta\, \Omega(h) \right],$$

5.2 Regression

5.2.1 True Loss and Surrogates

Unlike classification, regression requires jointly learning both the main predictor $f \in \mathcal{F}$ and the allocation policy defined by the rejector $r \in \mathcal{R}$ (see Equation 4). This interdependence introduces additional complexity, making a direct extension of the classification analysis infeasible. To address this, we formalize an outcome-wise adversarial deferral loss for regression, which characterizes a worst-case upper bound under perturbations. This formulation extends the deferral framework to settings where both prediction and deferral are adversarially sensitive. The result is summarized in Lemma 5.8, with the proof deferred to Appendix A.5.

Lemma 5.8 (Outcome-wise Adversarial Deferral Loss for Regression)**.** *For any hypothesis $r \in \mathcal{R}$, predictor $f \in \mathcal{F}$, input–label pair (x,t), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* outcome-wise adversarial deferral loss for regression *is defined as*

$$\widetilde{\ell}^{r}_{\text{def}}\big(r,f,t,x,\mathbf{m}\big) = \sum_{j=1}^{J+1} \tilde{c}^{r}_{j} \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{r}(x'_j) = j\},$$

Here, $\tilde{c}^{r}_{j} = \begin{cases} \alpha_1 \sup_{x'_1 \in B_p(x,\gamma)} L(f(x'_1), t) + \beta_1, & j = 1, \\ c^{r}_{j}(m_{j-1}, t), & j > 1. \end{cases}$

To make the regression setting amenable to optimization, we construct a surrogate that corrects for the adaptive penalization introduced by adversarial perturbations.

Definition 5.9 (Adversarial Surrogate Deferral Loss for Regression)**.** *For any hypothesis $r \in \mathcal{R}$, predictor $f \in \mathcal{F}$, input–label pair (x,t), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* adversarial surrogate deferral loss for regression *is defined as*

$$\widetilde{\Phi}^{r,u}_{\text{def}}(r,f,x,t,\mathbf{m}) = -(J-1)\tilde{c}^{r}_{1}(f,x,t) + \sum_{j=1}^{J+1} \sum_{i \neq j} \tilde{c}^{r}_{i}(f,x,m_{i-1},t) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{\text{cls}}(r(x'_j), j).$$

This surrogate loss provides a tractable relaxation of the outcome-wise adversarial deferral loss, while explicitly incorporating adaptive penalization through $\tilde{c}^{r}_{1}$.

Remark 2. This surrogate differs fundamentally from both its classification counterpart and the two-stage setting (Montreuil et al., 2025a), owing to the explicit involvement of the learnable predictor $f \in \mathcal{F}$.

5.2.2 Theoretical Guarantees

Similarly to the classification setting, we introduced both an adversarial surrogate deferral loss and its outcome-wise counterpart. It is therefore necessary to establish that the surrogate in Definition 5.9 is Bayes- and $(\mathcal{R},\mathcal{F})$-consistent. We prove the following theorem in Appendix A.6.

Theorem 5.10 ($(\mathcal{R},\mathcal{F})$-consistency bounds of $\widetilde{\Phi}^{r,u}_{\text{def}}$)**.** *Let $\mathcal{R}$ be symmetric and locally ρ-consistent. Then, for the set $\mathcal{A}^r$, any hypothesis $r \in \mathcal{R}$ and $f \in \mathcal{F}$, and any distribution $\mathcal{D}$, the following holds:*

$$\mathcal{E}_{\widetilde{\ell}^r_{def}}(r,f) - \mathcal{E}^B_{\widetilde{\ell}^r_{def}}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\ell}^r_{def}}(\mathcal{R},\mathcal{F}) \leq \overline{\Gamma}^u(1)\Big(\mathcal{E}_{\widetilde{\Phi}^{r,u}_{def}}(r,f) - \mathcal{E}^*_{\widetilde{\Phi}^{r,u}_{def}}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\Phi}^{r,u}_{def}}(\mathcal{R},\mathcal{F})\Big),$$

with $\overline{\Gamma}^u(1) = \max\{1, \Psi^u(1)\}$.

Theorem 5.10 shows that the adversarial surrogate loss for regression is $(\mathcal{R},\mathcal{F})$-consistent up to the factor $\max\{1,\Psi^u(1)\}$ with respect to the defined outcome-wise adversarial loss. Importantly, this factor differs from the classification case (see Theorem 5.3) precisely because regression involves the learnable predictor $f \in \mathcal{F}$ in addition to the rejector $r \in \mathcal{R}$. As a result, the analysis must bound a joint function of (r,f) rather than a single function, which makes the regression case fundamentally more challenging. Indeed, if two sequences $(f_t) \subset \mathcal{F}$ and $(r_t) \subset \mathcal{R}$ satisfy $\mathcal{E}_{\widetilde{\Phi}^{r,u}_{\text{def}}}(r_t,f_t) - \mathcal{E}^*_{\widetilde{\Phi}^{r,u}_{\text{def}}}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\Phi}^{r,u}_{\text{def}}}(\mathcal{R},\mathcal{F}) \to 0$, then by Theorem 5.10 we have $\mathcal{E}_{\widetilde{\ell}^r_{\text{def}}}(r_t,f_t) - \mathcal{E}^B_{\widetilde{\ell}^r_{\text{def}}}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\ell}^r_{\text{def}}}(\mathcal{R},\mathcal{F}) \to 0$. Therefore, minimizing the surrogate recovers an optimal predictor-rejector pair for the defined adversarial loss. In this case, the minimizability gap vanishes under realizability with respect to the pair (r,f), provided that $\mathcal{F} = \mathcal{F}_{\text{all}}$ and $\mathcal{R} = \mathcal{R}_{\text{all}}$.

Analogous to the classification case, we apply Lemma 5.4 to upper-bound the surrogate defined in Definition 5.9 by a smooth relaxation.

Corollary 5.11 (Smooth Adversarial Surrogate Deferral Loss for Regression)**.** *For any hypothesis $r \in \mathcal{R}$, predictor $f \in \mathcal{F}$, input–label pair (x,t), and expert outputs $\mathbf{m} = (m_1,\dots,m_J)$, the* smooth adversarial surrogate deferral loss for regression *is defined as*

$$\widetilde{\Phi}^{r,u}_{\text{def,s}}(r,f,x,t,\mathbf{m}) = -(J-1)\tilde{c}^r_1(f(x),t) + \sum_{j=1}^{J+1}\sum_{i\neq j}\tilde{c}^r_i(f,x,m_{i-1},t)\widetilde{\Phi}^u_{\text{cls,s}}(r(x),j).$$

This yields a novel surrogate with guarantees while ensuring tractable optimization.

Corollary 5.12 (Pointwise Upper Bound for $\widetilde{\Phi}^{r,u}_{\text{def, s}}$)**.** *For any hypothesis $r \in \mathcal{R}$, predictor $f \in \mathcal{F}$, input–label pair (x,t), and expert outputs $\mathbf{m} = (m_1,\dots,m_J)$, the following holds:*

$$\widetilde{\Phi}^{r,u}_{\text{def}}(r,f,x,t,\mathbf{m}) \leq \widetilde{\Phi}^{r,u}_{\text{def,s}}(r,f,x,t,\mathbf{m}).$$

Corollary 5.12 shows that the smooth surrogate $\widetilde{\Phi}^{r,u}_{\text{def,s}}$ pointwise upper-bounds $\widetilde{\Phi}^{r,u}_{\text{def}}$. This provides a tractable optimization objective aligned with the adversarial surrogate analyzed in Theorem 5.10.

Hence, it motivates the introduction of a regularized ERM algorithm for regression (RERM-R).

Definition 5.13 (RERM-R: Regularized ERM for $\widetilde{\Phi}^{r,u}_{\text{def,s}}$)**.** *Assume $\mathcal{R}$ is symmetric and locally ρ-consistent. Let $\Omega : \mathcal{R}\times\mathcal{F} \to \mathbb{R}_+$ be a convex regularizer, and let $\eta > 0$ be a hyperparameter. The regularized ERM objective is*

$$\min_{r\in\mathcal{R},\, f\in\mathcal{F}}\left[\frac{1}{n}\sum_{k=1}^{n}\widetilde{\Phi}^{r,u}_{\text{def,s}}\big(r,f,x_k,t_k,\mathbf{m}_k\big) + \eta\,\Omega(r,f)\right]$$

We will use this algorithm in Subsection 6.2. Its per-epoch computational cost in the h-score setting is summarized below.

Proposition 5.14 (Epoch cost of RERM-C in the h-score setting)**.** *Consider n training examples processed in mini-batches of size B, and let $\mathcal{A}$ denote the action space (for instance, $\mathcal{A} = \{1,\dots,K+J\}$). Suppose that each inner maximization is carried out by* PGD(T)*, that is, by T projected-gradient steps. Let C_{fwd} and C_{bwd} denote the costs of one forward and one backward pass through the score network h, respectively. Then one epoch of RERM-C has computational cost*

$$n\big(1 + |\mathcal{A}|T\big)\big(C_{\text{fwd}} + C_{\text{bwd}}\big).$$

Moreover, the peak memory requirement is the same as that of a standard forward-backward pass, up to the additional storage needed for one adversarial copy of each input currently being optimized.

6 EXPERIMENTS

We evaluate the robustness of one-stage L2D under adversarial perturbations, focusing on both classification and regression tasks. Our experiments show that standard one-stage L2D baselines can degrade sharply under adversarial perturbations, with attacks affecting both predictive accuracy and deferral decisions. By contrast, our algorithms (RERM-C and RERM-R) improve attacked performance while maintaining competitive clean performance. All results are averaged over three independent runs.

Comparison. Montreuil et al. (2025a) proposes a defense mechanism tailored to a more restrictive *two-stage* setting, in which only the router is learned while the experts are fixed. As a result, a direct comparison is not meaningful, since the learning setup and objectives differ fundamentally from ours. We therefore focus on standard one-stage L2D baselines; adapting generic adversarial training methods developed for pure classification problems (Awasthi et al., 2021) to the augmented L2D action space is beyond the scope of this work.

Metrics. We report the following evaluation metrics: *Clean Accuracy* (C.Acc, %) — the overall accuracy of the L2D system, accounting for both model predictions and expert deferrals on clean inputs; *Untargeted Accuracy* (U.Acc, %) — system accuracy under untargeted adversarial attack; *Targeted Accuracy* (T.Acc, %) — system accuracy under a targeted attack toward a randomly selected target action; *Adversarial Deferral Loss* (Def.Loss) — the empirical outcome-wise adversarial deferral loss.

6.1 Classification Setting

We evaluate on CIFAR-10 (Krizhevsky, 2009) and DermaMNIST from the MedMNIST benchmark suite (Yang et al., 2021, 2023; Tschandl et al., 2018; Codella et al., 2019). As a baseline, we compare against the consistent deferral framework of Mozannar and Sontag (2020); Mao et al. (2024a), which represent the standard approach for one-stage L2D. In particular, we instantiate our method with the logistic loss ($u = 1$) for RERM-C, as introduced in Definition 5.7. The main paper reports results on CIFAR-10, while additional results on DermaMNIST are deferred to the Appendix.

6.1.1 CIFAR10

Setting. The augmented classifier is implemented using ResNet-4 (He et al., 2016) and trained with the AdamW optimizer (Kingma and Ba, 2014) for 400 epochs. As experts, we employ three ResNet-16 models, each trained on a subset of the dataset; their accuracies are reported in Appendix A.8.2. The consultation costs are set as follows: $\beta_{j\leq K} = 0$ for predictions, and $\beta_{K+1} = 0.05$, $\beta_{K+2} = 0.075$, and $\beta_{K+3} = 0.1$ for the experts. The baseline method uses a learning rate of 0.005, while our approach employs a learning rate of 0.01. For the PGD attack, we set $\epsilon = 0.03137$, and in our method we additionally use the hyperparameters $\rho = 1.0$ and $\nu = 0.002$.

Results. Table 1 reports performance on the CIFAR-10 dataset, comparing our approach with Mao et al. (2024a). The baseline achieves slightly higher clean accuracy (82.67 vs. 75.60), but this advantage vanishes under adversarial perturbations: its accuracy collapses to 27.47 (untargeted) and 19.80 (targeted). In contrast, our method preserves robustness, improving untargeted accuracy by more than 24% and targeted accuracy by nearly 49%. Furthermore, our approach attains a substantially lower empirical adversarial deferral loss under the outcome-wise evaluation metric of Lemma 5.1. These results indicate that our surrogate-based methods improve attacked performance while retaining competitive clean performance.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024a)	82.67 ± 2.06	27.47 ± 0.63	19.80 ± 0.48	0.75 ± 0.02
Ours	75.60 ± 1.87	52.00 ± 1.31	68.67 ± 1.70	0.52 ± 0.01

Table 1: Performance under clean and adversarial inputs, compared against Mao et al. (2024a).

6.1.2 DermaMNIST

Setting. DermaMNIST is a subset of the MedMNIST dataset consisting of biomedical images for 7-class classification. The augmented classifier is implemented using ResNet-18 and trained for 100 epochs. As experts, we construct three specialized classifiers, each responsible for a randomly assigned subset of three classes (with overlap), predicting correctly with probability $p = 0.85$ on their assigned classes and uniformly at random otherwise; their accuracies are reported in the Appendix. The consultation costs are set as $\beta_{j\leq K} = 0$ for predictions, and $\beta_{K+1} = 0.05$, $\beta_{K+2} = 0.075$, and $\beta_{K+3} = 0.125$ for the experts. Both the baseline method and our approach use a learning rate of 0.005. For the PGD attack, we set $\epsilon = 0.03137$, while our approach additionally uses the hyperparameters $\rho = 1.75$ and $\nu = 0.001$.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024a)	83.39	30.82	27.08	69.60
Ours	81.80	71.12	80.65	31.66

Table 2: Performance under clean and adversarial inputs, compared against the approach of Mao et al. (2024a).

Results. On the DermaMNIST dataset, Table 2 shows that our approach remains close to the baseline under the clean setting. Under adversarial perturbations, our approach preserves substantially higher targeted and untargeted accuracy, while also achieving a lower empirical adversarial deferral loss.

6.2 Regression Task

We evaluate our approach on the Communities and Crime dataset (Redmond, 2002) and the Insurance Company Benchmark (COIL 2000) (Putten, 2000). As a baseline, we compare against Mao et al. (2024c), which represents the standard one-stage L2D approach. For our method, we instantiate the logistic loss ($u = 1$) within RERM-R, as introduced in Definition 5.13. The main paper reports results on Communities, while additional results on COIL 2000 are deferred to the Appendix.

6.2.1 Communities and Crime

Setting. The rejector is implemented as an MLP and trained using the AdamW optimizer (Kingma and Ba, 2014) for 500 epochs. The main predictor is a linear layer. As experts, we employ three MLPs specialized in different socio-economic factors (e.g., demographics, economics, housing), and report their accuracies in the Appendix. The consultation costs are set as $\beta_1 = 0$ for the main predictor, and $\beta_2 = 0.04$, $\beta_3 = 0.05$, and $\beta_4 = 0.07$ for experts m_1, m_2, m_3. Both the baseline and our method are trained with AdamW; the baseline uses a learning rate of 0.005, while our approach employs 0.01. For the PGD attack, we set $\epsilon = 0.5$; our method additionally uses the hyperparameters $\rho = 2.5$ and $\nu = 0.005$.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024c)	9.96 ± 0.21	18.08 ± 0.44	56.74 ± 1.57	18.93 ± 0.40
Ours	12.09 ± 0.39	12.13 ± 0.31	19.67 ± 0.12	12.94 ± 0.26

Table 3: Performance under clean and adversarial inputs, compared with Mao et al. (2024c). All accuracies are reported as RMSE, where lower values indicate better performance.

Results. Table 3 compares our method with Mao et al. (2024c) on the Communities & Crime dataset. All accuracies are reported as RMSE, where lower values indicate higher accuracy. On clean inputs, our approach performs comparably to the baseline. Under adversarial perturbations, however, our method provides substantial gains: in the targeted setting, RMSE is reduced by nearly 37%, and in the untargeted setting, it remains consistently lower than the baseline. Equally important, our approach achieves a markedly lower deferral loss. This gap suggests that our method attains a more favorable trade-off between predictive error, robustness, and consultation cost under the reported evaluation pipeline.

6.2.2 Insurance Company Benchmark (COIL 2000)

Setting. The rejector is implemented using an MLP and trained with the AdamW optimizer (Kingma and Ba, 2014) for 25 epochs. The main predictor is a linear layer. As experts, we employ four regression MLPs, each focusing on different customer segments (demographics, product ownership, high-value customers) and generating predictions using rules and noise; their accuracies are reported in the Appendix. The consultation costs are set as follows: $\beta_1 = 0$ for the main predictor, and $\beta_2 = 0.035$, $\beta_3 = 0.04$, $\beta_4 = 0.045$ and $\beta_5 = 0.05$ for the experts. The baseline method uses a learning rate of 0.005, while our approach employs a learning rate of 0.01. For the PGD attack, we set $\epsilon = 2$, and in our method we additionally use the hyperparameters $\rho = 2.75$ and $\nu = 0.01$.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024c)	7.02	11.61	8.31	11.98
Ours	7.39	7.41	7.40	7.81

Table 4: Performance under clean and adversarial inputs, compared against the approach of Mao et al. (2024c).

Results. On the COIL-2000 dataset, our approach remains close to the baseline under the clean setting while improving both attacked RMSE and adversarial deferral loss. This suggests that the approach transfers reasonably well across the reported regression benchmarks.

7 CONCLUSION

We presented a framework for adversarial robustness in one-stage Learning-to-Defer (L2D), addressing both classification and regression tasks. Our work makes three key advances: we formalized untargeted and targeted attacks that reveal how adversaries can jointly exploit prediction and deferral; we proposed outcome-wise adversarial surrogate losses with tractable relaxations and established consistency guarantees for the proposed adversarial losses; and we demonstrated empirically, across diverse benchmarks, that our methods improve attacked performance while maintaining competitive clean accuracy.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No. AISG2-PhD-2023-01-041-J) and by A*STAR, and forms part of the DesCartes

programme, which is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34:9804–9815, 2021.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class h-consistency bounds. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10077–10094. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/awasthi23c.html.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 02 2006. doi: 10.1198/016214505000000907.

Nina L Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, pages 453–463. PMLR, 2022.

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pages 2972–3005. PMLR, 2022.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1919–1925. AAAI Press, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *ArXiv*, abs/2010.03593, 2020. URL https://api.semanticscholar.org/CorpusID:222208628.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2478–2484. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/344. URL https://doi.org/10.24963/ijcai.2022/344. Main Track.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient LLM inference. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ka82fvJ5f1.

Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Learning-to-defer for sequential medical decision-making under uncertainty. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=0pn3KnbH5F.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4421–4434. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/234b941e88b755b7a72a1c1dd5022f30-Paper.pdf.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 154–165, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462516. URL https://doi.org/10.1145/3461702.3462516.

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

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

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pages 4816–4824. PMLR, 2024.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017. URL https://api.semanticscholar.org/CorpusID:3488815.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=GIlsH0T4b2.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Workshop on Combinatorial Image Analysis*, pages 107–135. Springer, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable h-consistent and bayes-consistent loss functions for learning to defer. *Advances in neural information processing systems*, 37:73638–73671, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024c.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2KlxjR6lsd.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=h3KHwZCnxH.

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=qmeNQpLiG5.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Why ask one when you can ask k? learning-to-defer to the top-k experts. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=mGbEv4kVoG.

Yannis Montreuil, Hoang Duy Dang, Maxime Meyer, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. Online learning-to-defer with varying experts. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026b. URL https://openreview.net/forum?id=1lix8ppUJ7.

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Optimal query allocation in extractive QA with LLMs: A learning-to-defer framework with theoretical guarantees. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026c. URL https://openreview.net/forum?id=kEVupwepTq.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International*

Conference on Artificial Intelligence and Statistics, 2023. URL https://api.semanticscholar.org/CorpusID:255941521.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29292–29304. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc8f76d9caadd48f77025b1c889d2e2d-Paper-Conference.pdf.

R A Ohn Aldrich. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–179, 1997.

Filippo Palomba, Andrea Pugnana, Jose Manuel Alvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=mkkFubLdNW.

Peter Putten. Insurance Company Benchmark (COIL 2000). UCI Machine Learning Repository, 2000. DOI: https://doi.org/10.24432/C5630S.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C53W3X.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007. URL https://api.semanticscholar.org/CorpusID:16660598.

Joshua Strong, Qianhui Men, and J Alison Noble. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28413–28421, 2025.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3475–3483. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007. URL http://jmlr.org/papers/v8/tewari07a.html.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, page 180161, 2018.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2022. URL https://api.semanticscholar.org/CorpusID:253237048.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11415–11434. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/verma23a.html.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32, 12 2002. doi: 10.1214/aos/1079120130.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. (Yes) see Appendix A.1
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. (Yes) see Appendix A.7
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. (Yes)

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. (Yes) see Appendix A.1
   (b) Complete proofs of all theoretical results. (Yes) see Appendix
   (c) Clear explanations of any assumptions. (Yes) see Appendix A.1

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

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

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

A Appendix

A.1 Important Definitions, Lemmas, and Theorems

Definition A.1 (Symmetric Hypothesis Class)**.** *Let $\mathcal{A}$ denote the set of possible actions (predictions and deferrals), and let Q be a class of score-valued hypotheses $q : \mathcal{X} \to \mathbb{R}^{|\mathcal{A}|}$. We say that Q is* symmetric *if it is closed under permutations of the coordinates indexed by $\mathcal{A}$, i.e., for any $q \in Q$ and any permutation $\Pi : \mathcal{A} \to \mathcal{A}$, the permuted hypothesis $q^{\Pi} : \mathcal{X} \to \mathbb{R}^{|\mathcal{A}|}$ defined by*

$$q^{\Pi}(x)_j = q(x)_{\Pi^{-1}(j)}, \qquad \forall x \in \mathcal{X},\ \forall j \in \mathcal{A},$$

also belongs to Q.

Definition A.2 (Locally ρ-consistent (Awasthi et al., 2023))**.** *A hypothesis set $\mathcal{Q}$ is locally ρ-consistent if, for any $x \in \mathcal{X}$, there exists a hypothesis $q \in \mathcal{Q}$ such that:*

$$\inf_{x' \in B_p(x,\gamma)} |q(x')_i - q(x')_j| \geq \rho,$$

where $\rho > 0$, $i \neq j \in \mathcal{A}$, and $x' \in B_p(x, \gamma)$. Moreover, the ordering of the values $\{q(x')_j\}$ is preserved with respect to $\{q(x)_j\}$ for all $x' \in B_p(x, \gamma)$.

Lemma A.3 ($\mathcal{Q}$-consistency bounds from Montreuil et al. (2025a))**.** *Assume $\mathcal{Q}$ is symmetric and locally ρ-consistent. Then, for the set $\mathcal{A}$, any hypothesis $q \in \mathcal{Q}$, and any distribution $\mathcal{P}$ with probabilities $p = (p_1, \ldots, p_{|\mathcal{A}|}) \in \Delta^{|\mathcal{A}|}$, the following inequality holds:*

$$\sum_{j \in \mathcal{A}} p_j \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{q}(x'_j) \neq j\} - \inf_{q \in \mathcal{Q}} \sum_{j \in \mathcal{A}} p_j \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{q}(x'_j) \neq j\} \leq$$
$$\Psi^u(1)\Big(\sum_{j \in \mathcal{A}} p_j \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{cls}(q(x'_j), j) - \inf_{q \in \mathcal{Q}} \sum_{j \in \mathcal{A}} p_j \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{cls}(q(x'_j), j) \Big).$$

with $\Psi^u(1) = \begin{cases} \log(2), & u = 1, \\ \frac{1}{1-u}\big(2^{1-u} - 1\big), & u \neq 1. \end{cases}$

A.2 Proof of Lemma 5.4

Lemma 5.4 (Smooth Adversarial Surrogate Losses)**.** *Let $x \in \mathcal{X}$ be a clean input, and let $\rho > 0$ and $\kappa > 0$ be hyperparameters. The smooth adversarial surrogate losses are defined as*

$$\widetilde{\Phi}^u_{\text{cls,s}}(h, x, j) = \Phi^u_{\text{cls}}(h(x)/\rho, j) + \kappa \sup_{x'_j \in B_p(x,\gamma)} \big\| \overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x, j) \big\|_2,$$

Proof. Fix a target class $j \in \mathcal{A}$. Define

$$\Phi^{\rho,u}_{\text{cls}}(h(x), j) = \Psi^u\left(\sum_{\substack{j' \in \mathcal{A} \\ j' \neq j}} \Psi_\rho\big(h(x)_{j'} - h(x)_j\big) \right), \qquad \widetilde{\Phi}^{\rho,u}_{\text{cls}}(h, x, j) = \sup_{x'_j \in B_p(x,\gamma)} \Psi^u\left(\sum_{\substack{j' \in \mathcal{A} \\ j' \neq j}} \Psi_\rho\big(h_{j'}(x'_j) - h_j(x'_j)\big) \right).$$

We take the exponential link $\Psi_{\text{e}}(v) = e^{-v}$ and the ρ-softening

$$\Psi_\rho(v) = \Psi_{\text{e}}\Big(\frac{v}{\rho}\Big) = \exp\Big(-\frac{v}{\rho}\Big), \qquad \rho > 0.$$

For $u > 0$ define

$$\Psi^u(v) = \begin{cases} \log(1 + v), & u = 1, \\ \dfrac{(1+v)^{1-u} - 1}{1-u}, & u \neq 1, \end{cases} \qquad v \geq 0.$$

Then Ψ^u is nondecreasing and 1-Lipschitz on $\mathbb{R}_+$ since $\left|\frac{\partial}{\partial v}\Psi^u(v)\right| = \frac{1}{(1+v)^u} \le 1$ for $v \ge 0$ and $u > 0$. Moreover, Ψ_ρ is $\frac{1}{\rho}$-Lipschitz on $\mathbb{R}$.

For $j' \neq j$ define the pairwise margin

$$\Delta_h(x, j, j') \;=\; h(x)_j - h(x)_{j'}.$$

Collect these $|\mathcal{A}| - 1$ margins into

$$\overline{\Delta}_h(x, j) = \big(\Delta_h(x, j, 1), \ldots, \Delta_h(x, j, j-1), \Delta_h(x, j, j+1), \ldots, \Delta_h(x, j, |\mathcal{A}|)\big) \in \mathbb{R}^{|\mathcal{A}|-1}.$$

Note that $h(x)_{j'} - h(x)_j = -\Delta_h(x, j, j')$, hence $\Psi_\rho\big(h_{j'}(\cdot) - h_j(\cdot)\big) = \Psi_\rho\big(-\Delta_h(\cdot, j, j')\big)$.

Using monotonicity and 1-Lipschitzness of Ψ^u on $\mathbb{R}_+$,

$$\begin{aligned}
\widetilde{\Phi}^{\rho,u}_{\mathrm{cls}}(h, x, j) &= \sup_{x'_j \in B_p(x,\gamma)} \Psi^u\left(\sum_{j' \neq j} \Psi_\rho\big(-\Delta_h(x'_j, j, j')\big)\right) \\
&\le \Psi^u\left(\sum_{j' \neq j} \Psi_\rho\big(-\Delta_h(x, j, j')\big)\right) + \sup_{x'_j \in B_p(x,\gamma)} \left|\sum_{j' \neq j} \Big(\Psi_\rho\big(-\Delta_h(x'_j, j, j')\big) - \Psi_\rho\big(-\Delta_h(x, j, j')\big)\Big)\right| \\
&= \Phi^{\rho,u}_{\mathrm{cls}}(h(x), j) \;+\; \sup_{x'_j \in B_p(x,\gamma)} \sum_{j' \neq j} \left|\Psi_\rho\big(-\Delta_h(x'_j, j, j')\big) - \Psi_\rho\big(-\Delta_h(x, j, j')\big)\right|. \qquad (7)
\end{aligned}$$

Since Ψ_ρ is $\frac{1}{\rho}$-Lipschitz and by Cauchy–Schwarz,

$$\begin{aligned}
\sum_{j' \neq j} \left|\Psi_\rho\big(-\Delta_h(x'_j, j, j')\big) - \Psi_\rho\big(-\Delta_h(x, j, j')\big)\right| &\le \frac{1}{\rho} \sum_{j' \neq j} \left|\Delta_h(x'_j, j, j') - \Delta_h(x, j, j')\right| \\
&\le \frac{\sqrt{|\mathcal{A}| - 1}}{\rho} \left\|\overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x, j)\right\|_2. \qquad (8)
\end{aligned}$$

Plugging (8) into (7) gives, for all j,

$$\widetilde{\Phi}^{\rho,u}_{\mathrm{cls}}(h, x, j) \;\le\; \Phi^{\rho,u}_{\mathrm{cls}}(h(x), j) \;+\; \kappa \sup_{x'_j \in B_p(x,\gamma)} \left\|\overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x, j)\right\|_2, \qquad \kappa \ge \frac{\sqrt{|\mathcal{A}| - 1}}{\rho}. \qquad (9)$$

Because $\Psi_\rho(v) = \exp(-v/\rho) = \Psi_{\mathrm{e}}(v/\rho)$ and Ψ^u is nondecreasing,

$$\Phi^{\rho,u}_{\mathrm{cls}}(h(x), j) = \Psi^u\left(\sum_{j' \neq j} \Psi_{\mathrm{e}}\left(\frac{h(x)_{j'} - h(x)_j}{\rho}\right)\right) = \Psi^u\left(\sum_{j' \neq j} \Psi_{\mathrm{e}}\left(\frac{h(x)_{j'}}{\rho} - \frac{h(x)_j}{\rho}\right)\right) = \;\Phi^u_{\mathrm{cls}}\left(\frac{h}{\rho}, x, j\right),$$

where $\frac{h}{\rho}$ denotes the score map $x \mapsto h(x)/\rho$ (componentwise). Hence (9) becomes

$$\widetilde{\Phi}^{\rho,u}_{\mathrm{cls}}(h, x, j) \;\le\; \Phi^u_{\mathrm{cls}}\left(\frac{h}{\rho}, x, j\right) \;+\; \kappa \sup_{x'_j \in B_p(x,\gamma)} \left\|\overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x, j)\right\|_2, \qquad (10)$$

Define the smooth upper-bounding adversarial surrogate

$$\widetilde{\Phi}^{\mathrm{smth},u}_{\mathrm{cls}}(h, x, j) = \;\Phi^u_{\mathrm{cls}}\left(\frac{h}{\rho}, x, j\right) \;+\; \kappa \sup_{x'_j \in B_p(x,\gamma)} \left\|\overline{\Delta}_h(x'_j, j) - \overline{\Delta}_h(x, j)\right\|_2,$$

Then (10) states, pointwise in (x, j),

$$\widetilde{\Phi}^{\rho,u}_{\mathrm{cls}}(h, x, j) \;\le\; \widetilde{\Phi}^{\mathrm{smth},u}_{\mathrm{cls}}(h, x, j).$$

□

A.3 Proof of Lemma 5.1

Lemma 5.1 (Outcome-wise Adversarial Deferral Loss for Classification)**.** *For any hypothesis $h \in \mathcal{H}$, input–label pair (x, y), and expert outputs $\mathbf{m} = (m_1, \ldots, m_J)$, the* outcome-wise adversarial deferral loss for classification *is*

$$\widetilde{\ell}^{c}_{\text{def}}(h, x, y, \mathbf{m}) = \sum_{j=1}^{K+J} \mu_j(j, m_j, y) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) = j\}.$$

Proof. We first show that we can rewrite the usual score-based loss (Mozannar and Sontag, 2020; Mao et al., 2024a) introduced in Equation 1 with a cost-sensitive formulation depending on a shifted cost μ_j. Let the following change of variable:

$$\mu_j(j, m_j, y) = \begin{cases} \alpha_j \mathbf{1}\{j \neq y\} + \beta_j & \text{if } j \leq K \\ \alpha_j \mathbf{1}\{m_{j-K} \neq y\} + \beta_j & \text{if } j > K, \end{cases} \tag{11}$$

leading to

$$\ell^{c}_{\text{def}}(\hat{h}(x), y, \mathbf{m}) = \sum_{j=1}^{K+J} \mu_j(j, m_j, y) \mathbf{1}\{\hat{h}(x) = j\}. \tag{12}$$

It is immediate that the formulation reduces to the standard score-based loss when setting $\alpha_j = 1$ and $\beta_j = 0$ for $j \leq K$. For instance, if $\mathcal{Y} = \{1, 2\}$ and the prediction is $\hat{h}(x) = 1$, then by Equation 11 we obtain $\ell^{c}_{\text{def}}(\hat{h}(x), y, \mathbf{m}) = \mathbf{1}\{1 \neq y\}$, which coincides with the usual score-based loss in Equation 1.

Now, we upper-bound this loss with the outcome-wise worst-case scenario under the newly cost-sensitive formulation:

$$\begin{aligned} \ell^{c}_{\text{def}}(\hat{h}(x), y, \mathbf{m}) &\leq \sup_{x' \in B_p(x,\gamma)} \sum_{j=1}^{K+J} \mu_j(j, m_j, y) \mathbf{1}\{\hat{h}(x') = j\} \\ &\leq \sum_{j=1}^{K+J} \mu_j(j, m_j, y) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) = j\} \\ &= \widetilde{\ell}^{c}_{\text{def}}(h, x, y, \mathbf{m}) \end{aligned} \tag{13}$$

□

A.4 Proof of Theorem 5.3

Theorem 5.3 ($\mathcal{H}$-consistency bounds of $\widetilde{\Phi}^{c,u}_{\text{def}}$)**.** *Let $\mathcal{H}$ be symmetric and locally ρ-consistent. Then, for the set $\mathcal{A}$, any hypothesis $h \in \mathcal{H}$, and any distribution $\mathcal{D}$, the following holds:*

$$\begin{aligned} &\mathcal{E}_{\widetilde{\ell}^{c}_{def}}(h) - \mathcal{E}^{B}_{\widetilde{\ell}^{c}_{def}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\ell}^{c}_{def}}(\mathcal{H}) \\ &\quad \leq \Psi^{u}(1)\Big(\mathcal{E}_{\widetilde{\Phi}^{c,u}_{def}}(h) - \mathcal{E}^{*}_{\widetilde{\Phi}^{c,u}_{def}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\Phi}^{c,u}_{def}}(\mathcal{H})\Big), \end{aligned}$$

with $\Psi^{u}(1) = \begin{cases} \log(2), & u = 1, \\ \frac{1}{1-u}\big(2^{1-u} - 1\big), & u \neq 1. \end{cases}$

Proof. Let $\mathcal{A} = \{1, \ldots, K+J\}$. For $x \in \mathcal{X}$ and radius $\gamma > 0$, write $B_p(x, \gamma) = \{x' \in \mathcal{X} : \|x' - x\|_p \leq \gamma\}$. Given a score vector $h(x) \in \mathbb{R}^{K+J}$, let the induced decision be $\hat{h}(x) \in \mathcal{A}$ (e.g., argmax with fixed tie-breaking). We begin by recalling the outcome-wise adversarial deferral loss

$$\widetilde{\ell}^{c}_{\text{def}}(h, x, y, \mathbf{m}) = \sum_{j=1}^{K+J} \mu_j(j, m_j, y) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) = j\}, \tag{14}$$

together with its surrogate counterpart introduced in Definition 5.2:

$$\widetilde{\Phi}^{c,u}_{\text{def}}(h,x,y,\mathbf{m}) = \sum_{j=1}^{K+J}\sum_{i\neq j}\mu_i(i,m_i,y)\sup_{x'_j\in B_p(x,\gamma)}\Phi^{\rho,u}_{\text{cls}}\big(h(x'_j),j\big). \tag{15}$$

True Loss Calibration Gap. Let define the conditional risk associated with the adversarial true loss:

$$\begin{aligned}\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) &= \mathbb{E}_{Y|X=x}\,\mathbb{E}_{M|X=x,Y}\big[\widetilde{\ell}^c_{\text{def}}(h,x,Y,M)\big]\\ &= \sum_{j=1}^{K+J}\mathbb{E}_{Y|X=x}\,\mathbb{E}_{M_j|X=x,Y}\Big[\mu_j(j,M_j,Y)\Big]\sup_{x'_j\in B_p(x,\gamma)}\mathbf{1}\{\hat{h}(x'_j)=j\}\\ &= \sum_{j=1}^{K+J}\overline{\mu}_j(x)\sup_{x'_j\in B_p(x,\gamma)}\mathbf{1}\{\hat{h}(x'_j)=j\}\end{aligned} \tag{16}$$

with $\overline{\mu}_j(x) = \mathbb{E}_{Y|X=x}\,\mathbb{E}_{M_j|X=x,Y}\Big[\mu_j(j,M_j,Y)\Big]$. Next, we assume there exist a function $h\in\mathcal{H}$ that is local-ρ-consistent (see Lemma A.2). For any h and x, define the reachability set

$$\mathcal{H}_\gamma(h,x) \;=\; \Big\{j\in\mathcal{A}:\ \exists\, x'_j\in B_p(x,\gamma)\text{ s.t. }\hat{h}(x'_j)=j\Big\}. \tag{17}$$

Then, for each $j\in\mathcal{A}$,

$$\sup_{x'_j\in B_p(x,\gamma)}\mathbf{1}\{\hat{h}(x'_j)=j\} = \mathbf{1}\{\,j\in\mathcal{H}_\gamma(h,x)\,\}. \tag{18}$$

By definition, the supremum over $x'_j\in B_p(x,\gamma)$ of the indicator equals 1 iff there exists at least one x'_j in the ball with $\hat{h}(x'_j)=j$, i.e., iff $j\in\mathcal{H}_\gamma(h,x)$; otherwise it is 0. Consequently,

$$\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) = \sum_{j=1}^{K+J}\overline{\mu}_j(x)\,\mathbf{1}\{\,j\in\mathcal{H}_\gamma(h,x)\,\}. \tag{19}$$

For any h, $\mathcal{H}_\gamma(h,x)\neq\varnothing$ because $x\in B_p(x,\gamma)$ and $\hat{h}(x)\in\mathcal{A}$, hence at least one label is realized in the ball.

$$\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) = \sum_{j\in\mathcal{A}}\overline{\mu}_j(x)\,\mathbf{1}\{j\in\mathcal{H}_\gamma(h,x)\} \;\geq\; \min_{j\in\mathcal{A}}\overline{\mu}_j(x)\mathbf{1}\{\mathcal{H}_\gamma(h,x)\neq\varnothing\} = \min_{j\in\mathcal{A}}\overline{\mu}_j(x). \tag{20}$$

Let $j^\star\in\arg\min_{j\in\mathcal{A}}\overline{\mu}_j(x)$. By local-$\rho$-consistency, there exists $h^{j^\star}$ with $\hat{h}^{j^\star}(x')=j^\star$ for all $x'\in B_p(x,\gamma)$. Therefore $\mathcal{H}_\gamma(h^{j^\star},x)=\{j^\star\}$.

$$\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h^{j^\star},x) = \sum_{j\in\mathcal{A}}\overline{\mu}_j(x)\,\mathbf{1}\{j=j^\star\} = \overline{\mu}_{j^\star}(x) = \min_{j\in\mathcal{A}}\overline{\mu}_j(x). \tag{21}$$

Combining the lower bound with achievability yields

$$\inf_{h\in\mathcal{H}}\;\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) \;=\mathcal{C}^B_{\widetilde{\ell}^c_{\text{def}}}(\mathcal{H},x) \;=\; \min_{j\in\mathcal{A}}\overline{\mu}_j(x). \tag{22}$$

Next, let's define the calibration gap:

$$\begin{aligned}\Delta\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) &= \mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) - \mathcal{C}^B_{\widetilde{\ell}^c_{\text{def}}}(\mathcal{H},x)\\ &= \mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) - \min_{j\in\mathcal{A}}\overline{\mu}_j(x)\\ &= \sum_{j=1}^{K+J}\Big(\overline{\mu}_j(x)\sup_{x'_j\in B_p(x,\gamma)}\mathbf{1}\{\hat{h}(x'_j)=j\}\Big) - \min_{j\in\mathcal{A}}\overline{\mu}_j(x)\end{aligned} \tag{23}$$

Surrogate Calibration Gap. Now, for the surrogate, we have:

$$\mathcal{C}_{\widetilde{\Phi}^{c,u}_{\text{def}}}(h,x) = \sum_{j=1}^{K+J} \sum_{i\neq j} \overline{\mu}_i(x) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{\text{cls}}\big(h(x'_j), j\big) \tag{24}$$

Let's define the probability distribution $p_j(x) = \sum_{i\neq j} \overline{\mu}_i(x)/S(x)$ with $S(x) = \sum_{j=1}^{K+J} \sum_{k\neq j} \overline{\mu}_k(x)$. It follows the calibration gap and by Lemma A.3:

$$\begin{aligned}
\Delta\mathcal{C}_{\widetilde{\Phi}^{c,u}_{\text{def}}}(h,x) &= S(x)\left(\sum_{j=1}^{K+J} p_j(x) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{\text{cls}}\big(h(x'_j), j\big) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{K+J} p_j(x) \sup_{x'_j \in B_p(x,\gamma)} \Phi^{\rho,u}_{\text{cls}}\big(h(x'_j), j\big) \right) \\
&\geq S(x)[\Psi^u(1)]^{-1}\left(\sum_{j=1}^{K+J} p_j(x) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) \neq j\}\big) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{K+J} p_j(x) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) \neq j\}\big) \right)
\end{aligned} \tag{25}$$

Relationship between Calibration gaps. We fix $h \in \mathcal{H}$ and analyze its excess risks at a given x. Define

$$\mathcal{E}_p(h,x) = \sum_{j=1}^{K+J} p_j(x) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j) \neq j\}.$$

Write for conciseness

$$\mu_*(x) = \min_{j\in\mathcal{A}} \overline{\mu}_j(x),$$

and use

$$M(x) = \sum_{i=1}^{K+J} \overline{\mu}_i(x), \qquad S(x) = \sum_{j=1}^{K+J} \sum_{k\neq j} \overline{\mu}_k(x) = (K+J-1)\, M(x),$$

so that $p_j(x) = \dfrac{M(x) - \overline{\mu}_j(x)}{S(x)}$. For a fixed h, $\sup_{x'_j} \mathbf{1}\{\hat{h}(x'_j) \neq j\} = 0$ iff $j \in \mathcal{H}_\gamma(h,x)$ and $\hat{h}(x') = j$ for all $x' \in B_p(x,\gamma)$, and equals 1 otherwise. Hence

$$\mathcal{E}_p(h,x) = 1 - \sum_{j\in\mathcal{A}} p_j(x)\, \mathbf{1}\big\{j \in \mathcal{H}_\gamma(h,x) \text{ and } \hat{h} \text{ is constant } j \text{ on } B_p(x,\gamma)\big\}.$$

Allowing a predictor that is constant on the ball with the most likely p-class yields the Bayes value

$$\mathcal{E}^\star_p(x) = 1 - \max_{j\in\mathcal{A}} p_j(x) = 1 - \frac{M(x) - \mu_*(x)}{S(x)}.$$

Therefore,

$$\mathcal{E}_p(h,x) - \mathcal{E}^\star_p(x) = \max_j p_j(x) \;-\; \sum_{j\in\mathcal{A}} p_j(x)\, \mathbf{1}\big\{j \in \mathcal{H}_\gamma(h,x) \text{ and } \hat{h} \text{ is constant } j \text{ on } B_p(x,\gamma)\big\}. \tag{26}$$

We therefore have two different cases to inspect.

Case (i): h is constant on $B_p(x,\gamma)$ with label j_0.

Then the RHS of (26) equals $\max_j p_j(x) - p_{j_0}(x)$. Using $p_j(x) = \frac{M(x)-\overline{\mu}_j(x)}{S(x)}$ and that $\max_j p_j$ occurs at any minimizer of $\overline{\mu}_j$,

$$\mathcal{E}_p(h,x) - \mathcal{E}^\star_p(x) = \frac{\overline{\mu}_{j_0}(x) - \mu_*(x)}{S(x)}.$$

Since here $\mathcal{H}_\gamma(h,x) = \{j_0\}$, the true excess is $\Delta\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x) = \overline{\mu}_{j_0}(x) - \mu_*(x)$, hence

$$\mathcal{E}_p(h,x) - \mathcal{E}^\star_p(x) = \frac{1}{S(x)}\, \Delta\mathcal{C}_{\widetilde{\ell}^c_{\text{def}}}(h,x). \tag{27}$$

Case (ii): h is not constant on $B_p(x,\gamma)$.

Then the second term in (26) vanishes and

$$\mathcal{E}_p(h,x) - \mathcal{E}_p^\star(x) = \max_j p_j(x) = \frac{M(x) - \mu_*(x)}{S(x)}.$$

Since $\sum_{j\in\mathcal{H}_\gamma(h,x)} \overline{\mu}_j(x) \le M(x)$, we obtain

$$\mathcal{E}_p(h,x) - \mathcal{E}_p^\star(x) \;\ge\; \frac{\sum_{j\in\mathcal{H}_\gamma(h,x)} \overline{\mu}_j(x) - \mu_*(x)}{S(x)} = \frac{1}{S(x)}\,\Delta\mathcal{C}_{\widetilde{\ell}^c_{\mathrm{def}}}(h,x). \tag{28}$$

Combining (27)–(28) yields, for all $h \in \mathcal{H}$,

$$\mathcal{E}_p(h,x) - \mathcal{E}_p^\star(x) \;\ge\; \frac{1}{S(x)}\,\Delta\mathcal{C}_{\widetilde{\ell}^c_{\mathrm{def}}}(h,x). \tag{29}$$

Lower-Bounding the Surrogate Calibration Gap. Using (29), we obtain:

$$\begin{aligned}
\Delta\mathcal{C}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(h,x) &= S(x)\left(\sum_{j=1}^{K+J} p_j(x) \sup_{x'_j\in B_p(x,\gamma)} \Phi^{\rho,u}_{\mathrm{cls}}\big(h(x'_j),j\big) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{K+J} p_j(x) \sup_{x'_j\in B_p(x,\gamma)} \Phi^{\rho,u}_{\mathrm{cls}}\big(h(x'_j),j\big)\right) \\
&\ge S(x)[\Psi^u(1)]^{-1}\left(\sum_{j=1}^{K+J} p_j(x) \sup_{x'_j\in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j)\neq j\}\big) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{K+J} p_j(x) \sup_{x'_j\in B_p(x,\gamma)} \mathbf{1}\{\hat{h}(x'_j)\neq j\}\big)\right) \\
&\ge S(x)[\Psi^u(1)]^{-1}\left(\frac{1}{S(x)}\,\Delta\mathcal{C}_{\widetilde{\ell}^c_{\mathrm{def}}}(h,x)\right) \\
&= [\Psi^u(1)]^{-1}\left(\Delta\mathcal{C}_{\widetilde{\ell}^c_{\mathrm{def}}}(h,x)\right)
\end{aligned} \tag{30}$$

Applying the expectation $\mathbb{E}_X[\Delta\mathcal{C}_\ell(h,X)] = \mathcal{E}_\ell(h) - \mathcal{E}_\ell^B(\mathcal{H}) + \mathcal{U}_\ell(\mathcal{H})$, leads to the desired inequality:

$$\mathcal{E}_{\widetilde{\ell}^c_{\mathrm{def}}}(h) - \mathcal{E}^B_{\widetilde{\ell}^c_{\mathrm{def}}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\ell}^c_{\mathrm{def}}}(\mathcal{H}) \le \Psi^u(1)\Big(\mathcal{E}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(h) - \mathcal{E}^*_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(\mathcal{H}) + \mathcal{U}_{\widetilde{\Phi}^{c,u}_{\mathrm{def}}}(\mathcal{H})\Big),$$

with $\Psi^u(1) = \begin{cases} \log(2), & u = 1, \\ \frac{1}{1-u}\big(2^{1-u}-1\big), & u \neq 1. \end{cases}$

□

A.5 Proof of Lemma 5.8

Lemma 5.8 (Outcome-wise Adversarial Deferral Loss for Regression)**.** *For any hypothesis $r \in \mathcal{R}$, predictor $f \in \mathcal{F}$, input–label pair (x,t), and expert outputs $\mathbf{m} = (m_1,\ldots,m_J)$, the* outcome-wise adversarial deferral loss for regression *is defined as*

$$\widetilde{\ell}^r_{\mathrm{def}}\big(r,f,t,x,\mathbf{m}\big) = \sum_{j=1}^{J+1} \tilde{c}^r_j \sup_{x'_j\in B_p(x,\gamma)} \mathbf{1}\{\hat{r}(x'_j)=j\},$$

Here, $\tilde{c}^r_j = \begin{cases} \alpha_1 \sup_{x'_1\in B_p(x,\gamma)} L(f(x'_1),t) + \beta_1, & j=1, \\ c^r_j(m_{j-1},t), & j>1. \end{cases}$

Proof. We begin by recalling the standard true deferral loss for regression from Mao et al. (2024c):

$$\ell^r_{\mathrm{def}}\big(\hat{r}(x),f(x),t,\mathbf{m}\big) = \sum_{j=1}^{J+1} c^r_j\big(f(x),m_{j-1},t\big)\,\mathbf{1}\{\hat{r}(x)=j\}. \tag{31}$$

To capture the worst-case scenario, we account for adversarial perturbations of the input while noting that the cost $c_1^r(f(x),t)$ depends explicitly on x and that the predictor $f \in \mathcal{F}$ is trainable:

$$\begin{aligned}\ell_{\text{def}}^r\big(\hat{r}(x), f(x), t, \mathbf{m}\big) &\leq \sup_{x' \in B_p(x,\gamma)} \sum_{j=1}^{J+1} c_j^r\big(f, x, m_{j-1}, t\big)\, \mathbf{1}\{\hat{r}(x') = j\} \\ &\leq \sum_{j=1}^{J+1} \tilde{c}_j^r\big(f, x, m_{j-1}, t\big) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{r}(x'_j) = j\} \\ &= \tilde{\ell}_{\text{def}}^r\big(r, f, x, t, \mathbf{m}\big).\end{aligned} \tag{32}$$

□

A.6 Proof of Theorem 5.10

Theorem 5.10 ($(\mathcal{R},\mathcal{F})$-consistency bounds of $\widetilde{\Phi}_{\text{def}}^{r,u}$)**.** *Let $\mathcal{R}$ be symmetric and locally ρ-consistent. Then, for the set $\mathcal{A}^r$, any hypothesis $r \in \mathcal{R}$ and $f \in \mathcal{F}$, and any distribution $\mathcal{D}$, the following holds:*

$$\begin{aligned}&\mathcal{E}_{\widetilde{\ell}_{def}^r}(r,f) - \mathcal{E}^B_{\widetilde{\ell}_{def}^r}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\ell}_{def}^r}(\mathcal{R},\mathcal{F}) \\ &\leq \overline{\Gamma}^u(1)\Big(\mathcal{E}_{\widetilde{\Phi}_{def}^{r,u}}(r,f) - \mathcal{E}^*_{\widetilde{\Phi}_{def}^{r,u}}(\mathcal{R},\mathcal{F}) + \mathcal{U}_{\widetilde{\Phi}_{def}^{r,u}}(\mathcal{R},\mathcal{F})\Big),\end{aligned}$$

with $\overline{\Gamma}^u(1) = \max\{1, \Psi^u(1)\}$.

Proof. The outcome-wise adversarial deferral loss for regression is defined as

$$\tilde{\ell}_{\text{def}}^r(f, r, x, t, \mathbf{m}) = \sum_{j=1}^{J+1} \tilde{c}_j^r(f, x, m_{j-1}, t) \sup_{x'_j \in B_p(x,\gamma)} \mathbf{1}\{\hat{r}(x'_j) = j\}, \tag{33}$$

□

A.7 Complexity

Proposition A.4 (Epoch cost of RERM-C in the h-score setting)**.** *Process n training examples in mini-batches of size B. Let $\mathcal{A}$ be action space (e.g. $\mathcal{A} = \{1,\dots,K{+}J\}$) and let $T \in \mathbb{N}$ be the number of projected-gradient steps used in each inner maximization (*PGD*(T)). Write C_{fwd} and C_{bwd} for the costs of one forward and one backward pass through the score network h. Then one epoch of RERM-C minimization incurs*

$$n\,\big(1 + |\mathcal{A}|\,T\big)\,\big(C_{\text{fwd}} + C_{\text{bwd}}\big) \tag{34}$$

network traversals, while the peak memory equals that of a single forward–backward pass plus the storage of one adversarial copy of each input currently being optimized.

Proof. Consider one mini-batch. RERM-C performs:

(i) one *clean* forward pass of h to compute scores $h(x)$ and the loss terms;

(ii) for each $j \in \mathcal{A}$ and each of the T PGD steps that update the adversarial proxy $x'_j \in B_p(x,\gamma)$, one forward *and* one backward pass of h (to obtain gradients w.r.t. the input);

(iii) one backward pass to update θ (the parameter gradient of the batch loss).

Thus, per mini-batch, the total number of network traversals equals

$$\underbrace{1}_{\text{clean forward}} + \underbrace{|\mathcal{A}|\,T}_{\text{PGD forwards}} + \underbrace{|\mathcal{A}|\,T}_{\text{PGD backwards}} + \underbrace{1}_{\text{parameter backward}} = 2\big(1 + |\mathcal{A}|\,T\big).$$

Multiplying by the number of mini-batches $\lceil n/B \rceil$ gives the total cost

$$\lceil n/B \rceil \cdot 2\big(1 + |\mathcal{A}|\, T\big) \;\approx\; \frac{n}{B} \cdot 2\big(1 + |\mathcal{A}|\, T\big),$$

which, when expressed in units of *per-example* forward/backward costs, yields (34): each example induces $\big(1 + |\mathcal{A}|\, T\big)$ forwards and the same number of backwards, for a total of $n\big(1 + |\mathcal{A}|\, T\big)\big(C_{\text{fwd}} + C_{\text{bwd}}\big)$.

□

A.8 Experiments

A.8.1 Resources

All experiments were conducted on an internal cluster using an NVIDIA A100 GPU with 80 GB of VRAM.

A.8.2 CIFAR10

Expert	1	2	3
Accuracy	37.55	35.92	38.54

Table A1: CIFAR10: Accuracy of Experts

Analysis in the main paper.

A.8.3 DermaMNIST

Expert	1	2	3
Accuracy	28.48	30.52	71.72

Table A2: DermaMNIST: Accuracy of Experts

Setting. DermaMNIST is a subset of the MedMNIST dataset consisting of biomedical images for 7-class classification. The augmented classifier is implemented using ResNet-18 and trained for 100 epochs. As experts, we construct three specialized classifiers, each responsible for a randomly assigned subset of three classes (with overlap), predicting correctly with probability $p = 0.85$ on their assigned classes and uniformly at random otherwise; their accuracies are reported above. The consultation costs are set as $\beta_{j \le K} = 0$ for predictions, and $\beta_{K+1} = 0.05$, $\beta_{K+2} = 0.075$, and $\beta_{K+3} = 0.125$ for the experts. Both the baseline method and our approach use a learning rate of 0.005. For the PGD attack, we set $\epsilon = 0.03137$, while our approach additionally uses the hyperparameters $\rho = 1.75$ and $\nu = 0.001$.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024a)	83.39	30.82	27.08	69.60
Ours	81.80	71.12	80.65	31.66

Table A3: Performance under clean and adversarial inputs, compared against the approach of Mao et al. (2024a).

Results. On the DermaMNIST dataset, Table A3 shows that our approach remains close to the baseline under the clean setting while substantially improving targeted and untargeted attacked accuracy, together with a lower empirical adversarial deferral loss.

A.8.4 Community and Crime

Expert	1	2	3
RMSE	0.5442	1.1373	1.5613

Table A4: community and crime: Accuracy of Experts (RMSE)

Analysis in the main paper.

A.8.5 Insurance Company Benchmark (COIL 2000)

Expert	1	2	3	4
RMSE	0.0744	0.0747	0.0741	0.831

Table A5: coil2000: Accuracy of Experts (RMSE)

Setting. The rejector is implemented using an MLP and trained with the AdamW optimizer (Kingma and Ba, 2014) for 25 epochs, while the main predictor is a linear layer. As experts, we employ four regression MLPs, each focusing on different customer segments (demographics, product ownership, high-value customers) and generating predictions using rules and noise; their accuracies are reported above. The consultation costs are set as follows: $\beta_1 = 0$ for the main predictor, and $\beta_2 = 0.035$, $\beta_3 = 0.04$, $\beta_4 = 0.045$ and $\beta_5 = 0.05$ for the experts. The baseline method uses a learning rate of 0.005, while our approach employs a learning rate of 0.01. For the PGD attack, we set $\epsilon = 2$, and in our method we additionally use the hyperparameters $\rho = 2.75$ and $\nu = 0.01$.

	C.Acc	U.Acc	T.Acc	Def.Loss
Mao et al. (2024c)	7.02	11.61	8.31	11.98
Ours	7.39	7.41	7.40	7.81

Table A6: Performance under clean and adversarial inputs, compared against the approach of Mao et al. (2024c).

Results. On the COIL-2000 dataset, our approach remains close to the baseline under the clean setting while improving both attacked RMSE and adversarial deferral loss, suggesting that the method transfers reasonably well across the reported regression benchmarks.