# Online Learning-to-Defer with Varying Experts

**Dang Hoang Duy**[*,1] **Yannis Montreuil**[*,1,4,5] **Maxime Meyer**[*,1,4]

**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) methods route each query either to a predictive model or to external experts. While existing work studies this problem in batch settings, real-world deployments require handling streaming data, changing expert availability, and shifting expert distribution. We introduce the first online L2D algorithm for multiclass classification with bandit feedback and a dynamically varying pool of experts. Our method achieves regret guarantees of $O((n+n_e)T^{2/3})$ in general and $O((n+n_e)\sqrt{T})$ under a low-noise condition, where $T$ is the time horizon, $n$ is the number of labels, and $n_e$ is the number of distinct experts observed across rounds. The analysis builds on novel $\mathcal{H}$-consistency bounds for the online framework, combined with first-order methods for online convex optimization. Experiments on synthetic and real-world datasets demonstrate that our approach effectively extends standard Learning-to-Defer to settings with varying expert availability and reliability.

## 1 INTRODUCTION

Learning-to-Defer (L2D) is a framework for integrating machine learning models with human experts—or more broadly, auxiliary decision-makers—to enhance the reliability of automated systems. The central idea is to let a predictive model handle routine or low-risk inputs autonomously, while deferring uncertain or high-stakes cases to more competent agents such as domain experts, larger models, or specialized subsystems (Madras et al., 2018; Mozannar and Sontag, 2020; Verma et al., 2022). This deference mechanism enables L2D systems to balance predictive accuracy against consultation cost, leveraging the scalability of machine learning while preserving safety through expert oversight. For instance, in medical diagnostics, a model may issue a confident prediction for straightforward cases but defer ambiguous ones to a clinician, thereby reducing risk without overburdening human resources (Johnson et al., 2016, 2023; Strong et al., 2024).

While L2D has shown promise, existing formulations rely on assumptions that rarely hold in practice. They posit a fixed pool of experts available throughout deployment and require predictions from every expert in training. In reality, expert availability and reliability are dynamic: clinicians may be available only during scheduled shifts, their accuracy may degrade under fatigue, and automated subsystems may be restricted by budget or resource constraints. Moreover, many applications generate data sequentially, demanding decisions in an online rather than batch setting. These conditions expose a fundamental gap between classical L2D assumptions and the requirements of real-world deployment.

To the best of our knowledge, we present the first online Learning-to-Defer framework, where both the data distribution and the pool of experts may evolve over time. The formulation further accounts for experts whose reliability may drift, and, unlike classical L2D methods, it only requires feedback from the experts that are actually queried, rather than predictions from all experts to compute the loss. To assess the theoret-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

ical performance of our approach, we analyze regret, which measures the gap between the learner's cumulative performance and that of the best strategy in hindsight over $T$ rounds. We prove that our algorithm achieves regret sublinear in $T$, ensuring that the average per-round regret vanishes asymptotically.

Our main contributions are as follows: (**i**) We introduce a general framework for online Learning-to-Defer under bandit feedback, where both the number of experts and the data distribution may vary over time. (**ii**) We extend the notion of $\mathcal{H}$-consistency bounds to the online setting, whereas previous analyses were limited to batch learning. (**iii**) We propose the first online algorithm for this setting, achieving regret $O((n+n_e)T^{2/3})$ in general and $O((n+n_e)\sqrt{T})$ under a low-noise assumption. (**iv**) We empirically validate our approach on both synthetic and real-world datasets, demonstrating its competitiveness and robustness across diverse scenarios.

## 2 RELATED WORK

**Learning-to-Defer.** Learning-to-Defer (L2D) generalizes selective prediction (Chow, 1970; Bartlett and Wegkamp, 2008; Cortes et al., 2016; Geifman and El-Yaniv, 2017) by allowing a model not only to abstain on uncertain inputs but also to defer them to external experts (Madras et al., 2018; Mozannar and Sontag, 2020; Verma et al., 2022). Prior work has focused exclusively on the batch-learning setting.

The score-based formulation of Mozannar and Sontag (2020) introduced the first Bayes-consistent surrogate, which has since become the standard for one-stage frameworks. Subsequent work has refined calibration (Verma et al., 2022; Cao et al., 2024), improved surrogate design (Charusaie et al., 2022; Mao et al., 2024a; Montreuil et al., 2026a), and established guarantees such as $\mathcal{H}$-consistency and realizability (Mozannar et al., 2023; Mao et al., 2024b, 2025). Applications span diverse classification tasks (Kerrigan et al., 2021; Keswani et al., 2021; Hemmer et al., 2022; Verma et al., 2022; Benz and Rodriguez, 2022; Cao et al., 2024; Tailor et al., 2024; Liu et al., 2024), while a special case of this formulation is the *two-stage* framework, in which the base predictor and experts are trained offline and only the allocation function is learned (Narasimhan et al., 2022; Mao et al., 2023a). Extensions of two-stage approaches further address regression (Mao et al., 2024c), multi-task learning (Montreuil et al., 2025b), adversarial robustness (Montreuil et al., 2025a, 2026c), and applied systems (Strong et al., 2024; Palomba et al., 2025; Montreuil et al., 2026b).

**Online-Learning.** The most relevant line of work to our paper is bandit multiclass classification. A common approach employs *proper learning algorithms* that fix a hypothesis before the next input (Kakade et al., 2008). Subsequent papers refine this idea (Hazan and Kale, 2011; Beygelzimer et al., 2017; van der Hoeven, 2020; van der Hoeven et al., 2021), with van der Hoeven (2020) exploiting the gap between surrogate and zero-one loss. We also use a proper learner, but we evaluate performance directly in terms of the zero-one loss rather than only through a surrogate. Our distributional assumptions are broader than those in Crammer and Gentile (2013), which treat a specific label distribution, while our results cover arbitrary label distributions.

Another relevant line of work is online logistic regression. Foster et al. (2018) give a statistically optimal strategy that is not computationally practical. Jézéquel et al. (2020) provide an efficient algorithm in the binary case. More recent work (Jézéquel et al., 2021; Agarwal et al., 2022) develops algorithms with near-optimal regret and efficient running time, and shows how to use them for bandit classification. Our analysis follows this template. We first study the underlying online convex optimization problem and then combine it with a separate argument to obtain guarantees for multiclass bandit classification.

With a fixed number of experts per round, our setting reduces to contextual multi-armed bandits (Auer et al., 2002; Li et al., 2010). In this regime it is known that $R_{\ell_{\text{def}}}(T) = \Omega(\sqrt{T})$ (Daniely and Helbertal, 2013). This lower bound motivates our focus on additional structure beyond a generic contextual-bandit reduction and explains why controlling the true zero-one loss within a proper-learner framework is natural for our setting.

## 3 PRELIMINARIES

**Multiclass Classification.** We consider the standard multiclass classification setting. The input space is $\mathcal{X} = \{x \in \mathbb{R}^d : 1 \leq \|x\|_2 \leq R\}$ for some radius $R > 0$, and the label space is $\mathcal{Y} = [n] := \{1, \ldots, n\}$ with $n \geq 2$ classes (Mohri et al., 2012). The predictions are obtained through a hypothesis $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, via the decision rule $h(x) = \arg\max_{j \in \mathcal{Y}} h(x, j)$. The performance of $h$ is measured by the 0–1 loss $\ell(h(x), y) = \mathbf{1}\{h(x) \neq y\}$, which directly captures misclassification error. Since the 0–1 loss is non-differentiable, learning algorithms typically minimize a surrogate that approximates it (Zhang, 2002; Steinwart, 2007). We denote by $\Phi_{01} : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a general surrogate loss. Two prominent families are of particular interest: (i) the cross-entropy family $\Phi_{01}^{\text{cse}}$,

which includes the exponential and logistic losses, and (ii) the constrained surrogate family $\Phi_{01}^{\text{cstnd}}$, which includes constrained hinge-type losses (Awasthi et al., 2021, 2022a; Mao et al., 2023c). A surrogate $\Phi_{01}$ is said to be *$\mathcal{H}$-consistent* if minimizing its excess surrogate risk ensures recovery of the Bayes-optimal predictor within the hypothesis class $\mathcal{H}$. Writing the risk of a loss $\ell$ as $\mathcal{E}_\ell(h) = \mathbb{E}[\ell(h(x), y)]$ in batch-learning, this property is formally characterized by consistency bounds (Awasthi et al., 2022a).

**Definition 3.1** ($\mathcal{H}$-Consistency Bounds)**.** *Suppose the surrogate $\Phi_{01}$ is $\mathcal{H}$-consistent w.r.t the 0-1 loss $\ell_{01}$ for any distribution $\mathcal{D}$. Then, there exists a non-decreasing function $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for all $h \in \mathcal{H}$,*

$$\mathcal{E}_{\ell_{01}}(h) - \mathcal{E}^*_{\ell_{01},\mathcal{H}} - \mathcal{U}_{\ell_{01},\mathcal{H}} \leq \Gamma\Big(\mathcal{E}_{\Phi_{01}}(h) - \mathcal{E}^*_{\Phi_{01},\mathcal{H}} - \mathcal{U}_{\Phi_{01},\mathcal{H}}\Big).$$

Here $\mathcal{U}_{\ell_{01},\mathcal{H}} = \mathcal{E}^*_{\ell_{01},\mathcal{H}} - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}\big[\ell_{01}(h(x), y)\big]\big]$ is the *minimizability gap*, quantifying the irreducible approximation error due to the expressive limitations of $\mathcal{H}$. When $\mathcal{H}$ is sufficiently rich—such as $\mathcal{H} = \mathcal{H}_{\text{all}}$—this gap vanishes, and the inequality recovers standard Bayes-consistency guarantees (Steinwart, 2007; Awasthi et al., 2022a).

**Batch Learning-to-Defer.** Standard batch L2D assumes access to a fixed pool of $n_e$ experts $\mathcal{G} = \{g_1, g_2, \ldots, g_{n_e}\}$, each being a mapping $g_i : \mathcal{X} \to \mathcal{Y}$ (Madras et al., 2018; Mozannar and Sontag, 2020). Training examples $(x, y)$ are drawn i.i.d. from an unknown distribution $\mathcal{D}$, and for every input $x$ we observe the tuple $(y, g_1(x), g_2(x), \ldots, g_{n_e}(x))$.

The learner then has to decide, for each input, whether to predict directly or defer to an expert. To formalize this, we define the augmented label space $\overline{\mathcal{Y}} = \mathcal{Y} \cup \{n + j : g_j \in \mathcal{G}\}$, where indices $1, \ldots, n$ correspond to class predictions and indices $n + j$ correspond to deferral to expert $g_j$.

The goal of the learner is to find a hypothesis $h : \mathcal{X} \times \overline{\mathcal{Y}} \to \mathbb{R}$ that minimizes the true score-based loss:

**Definition 3.2** (True Score-Based Loss)**.** *Let $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $h \in \mathcal{H}$. The true score-based loss is*

$$\ell^{\text{batch}}_{\text{def}}(h(x), y) = \begin{cases} \mathbf{1}\{h(x) \neq y\} & \text{if } h(x) \in \mathcal{Y}, \\ c_j(x, y) & \text{if } h(x) = n + j. \end{cases}$$

Here, $c_j(x, y)$ accounts for both the misclassification and the query cost of expert $g_j$. We denote by $\underline{c}_j \geq 0$ and $\overline{c}_j \leq 1$ the finite lower bound and upper bound of $c_j$, that is $c_j(x, y) \in [\underline{c}_j, \overline{c}_j]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Directly minimizing $\ell^{\text{batch}}_{\text{def}}$ is intractable due to its non-differentiability (Zhang, 2002; Steinwart, 2007). Following Mozannar and Sontag (2020); Mao et al. (2024a), we instead optimize a convex surrogate loss that is Bayes and $\mathcal{H}$-consistent for $\ell^{\text{batch}}_{\text{def}}$, and has since become the standard formulation.

**Definition 3.3** (Score-Based Surrogate)**.** *Let $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $h \in \mathcal{H}$. The score-based surrogate is*

$$\Phi^{\text{batch}}_{\text{def}}(h, x, y) = \Phi_{01}(h, x, y) + \sum_{j=1}^{n_e} (1 - c_j(x, y))\, \Phi_{01}(h, x, n + j),$$

# 4 EXTENDING LEARNING-TO-DEFER TO THE ONLINE-LEARNING SETTING

## 4.1 Problem Formulation

**Problem Setup.** Prior work assume access to a fixed pool of experts (Madras et al., 2018; Mozannar and Sontag, 2020; Verma et al., 2022), but this assumption rarely holds in practice. In realistic deployments, the set of available experts may evolve over time: human annotators can join or leave, domain specialists may be consulted only intermittently, and automated subsystems may become unavailable due to resource or reliability constraints. This motivates the introduction of the online framework, in which the learner receives information and makes decision sequentially. To capture this, we denote by $\mathcal{A} \subseteq \mathcal{P}(\mathcal{G})$ the collection of feasible expert subsets, and $\overline{\mathcal{X}} = \mathcal{X} \times \mathcal{A}$ the augmented input space. That is the augmented inputs $\overline{x} = (x, A)$ consist of both the classical input $x \in \mathcal{X}$ and the set of available experts $A \in \mathcal{A}$.

The interaction protocol is as follows: (**i**) At each round $t$, the adversary selects an augmented input $\overline{x}_t \in \overline{\mathcal{X}}$ adversarially (Hazan, 2016), which is then revealed to the learner. (**ii**) A label $y$ is drawn from a fixed but unknown distribution $p(\cdot \mid \overline{x}_t)$, depending only on $\overline{x}_t$. (**iii**) The learner commits to a prediction $y'_t$ from the augmented label set with experts available at round $t$, namely $\overline{\mathcal{Y}}_{A_t} = [n] \cup \{n + j : g_j \in A_t\}$. (**iv**) The learner then incurs a loss determined by its prediction.

We use the bold notation to denote a sequence of objects. Hence the sequence of inputs $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_T) \in \overline{\mathcal{X}}^T$ is denoted $\overline{\mathbf{x}}$.

**Bandit Feedback.** We work in the bandit feedback setting, where the learner only observes feedback about the chosen action. If it predicts a label $y'_t \in [n]$,

it does not receive the true label $y_t$ but only a binary signal indicating correctness, $\mathbf{1}\{y'_t = y_t\}$. If it defers to expert $g_j$ (i.e., $y'_t = n + j$), it observes whether the expert was correct, $\mathbf{1}\{g_j(x_t) = y_t\}$, together with the associated cost $c_j(x_t, y_t)$. In contrast, in the batch Learning-to-Defer framework the learner observes the true label and the correctness of *all* experts during training, regardless of whether they are queried. This restriction makes the bandit setting not only substantially more challenging but also more faithful to practical applications than standard supervised classification, where the learner always receives the full label information.

**Hypothesis Set.** As for standard batch L2D, we consider that the learner predicts through a hypothesis $h : \overline{\mathcal{X}} \times \overline{\mathcal{Y}} \to \mathbb{R}$. For every augmented input $\overline{x}_t = (x_t, A_t)$, the prediction of the learner is then $h(\overline{x}_t) = \arg\max_{y \in \overline{\mathcal{Y}}_{A_t}} h(\overline{x}_t, y)$. We only consider the linear hypothesis set, *id est* hypothesis in the form $h(\overline{x}, y) = \langle w_y, x\rangle + b_y$ where $w_y$ is a vector of some weight matrix $W \in \mathbb{R}^{N\times d}$, $b_y$ is the corresponding bias, and $N = n + n_e$ is the size of the augmented label set. We denote by $\mathcal{H}$ the set of such hypotheses. Note that in this work, we only consider proper learning algorithms. That is the hypothesis $h_t \in \mathcal{H}$ at each round $t$ is generated before seeing input $\overline{x}_t$. Finally, the goal of the learner is to minimize the gap between his sequence of hypotheses $\mathbf{h}$ and the best hypothesis in hindsight $h$.

**Semi-Adversarial Setting.** We deliberately adopt a semi-adversarial framework, where inputs $\overline{x}_t$ may be chosen adversarially while labels are drawn from a fixed distribution $p$. This choice reflects the core structure of Learning-to-Defer: in realistic deployments, the instances presented to the system—such as medical cases, flagged content, or decision-support queries—may arrive in arbitrary or even adversarial ways, yet both the learner and the experts must ultimately be evaluated against the same ground-truth distribution. Without this structure, the notion of deferring to experts would be ill-posed, since expert reliability can only be assessed relative to a consistent source of truth.

To denote the induced distributions, given any finite set $\mathcal{S}$, we thus define $\Delta_{\mathcal{S}} \subset \mathbb{R}^{|\mathcal{S}|}$ as the set of standard $|\mathcal{S}|$-dimensional simplexes representing distributions over $\mathcal{S}$. We denote by $\mathcal{D}$ a set of conditional distributions over $\mathcal{Y}$, indexed by $\overline{x}$, that is, $\mathcal{D} \subseteq \{p : \overline{\mathcal{X}} \to \Delta_{\mathcal{Y}}\}$. For ease of notation, write $p(\overline{x}, y) = \mathbb{P}(Y = y|\overline{X} = \overline{x})$.

This setting forces us to randomize the learner's predictions, as discussed in Appendix A. At each round $t$, we thus sample the predictions from a distribution $q_t \in \Delta_{\overline{\mathcal{Y}}_{A_t}}$:

$$q_t = (1 - \gamma_t)e_{h_t(\overline{x}_t)} + \frac{\gamma_t}{|\overline{\mathcal{Y}}_{A_t}|}\mathbf{1},$$

where $e_y$ is the unit vector in the direction of $y$ and $\gamma_t \in (0, 1)$ is the exploration rate.

### 4.2 Loss Formulation

As introduced in Definition 3.2, the classical Learning-to-Defer formulation (Mozannar and Sontag, 2020) is defined in a batch-learning setting, assuming access to a fixed dataset and a fixed pool of experts. To relax these assumptions, we introduce a *true deferral loss*, which naturally generalizes Definition 3.2 to accommodate scenarios where the set of available experts may be restricted.

**Definition 4.1** (True Deferral Loss)**.** *The true deferral loss is defined as follows for any $h \in \mathcal{H}$ and $(\overline{x}, y) = ((x, A), y) \in \overline{\mathcal{X}} \times \mathcal{Y}$*

$$\ell_{\text{def}}(h, \overline{x}, y) = \begin{cases} \mathbf{1}\{h(\overline{x} \neq y\} & \text{if } h(\overline{x}) \in \mathcal{Y}, \\ c_j(\overline{x}, y) & \text{if } h(\overline{x}) \in \{n + j : g_j \in A_t\}. \end{cases}$$

*Moreover, let the online learner predict according to a distribution $q \in \Delta_{\overline{\mathcal{Y}}_A}$. The corresponding true deferral loss with respect to the learner is defined by*

$$\ell_{\text{def}}(q, \overline{x}, y) = \mathbb{E}_{y' \sim q}\Big[\mathbf{1}\{y' \in \mathcal{Y}\}\mathbf{1}\{y' \neq y\} + \sum_{n+j \in \overline{\mathcal{Y}}_A} c_j(x, y)\mathbf{1}\{y' = n + j\}\Big].$$

Note that our true deferral loss, introduced in Definition 4.1, directly extends the formulations of Mozannar and Sontag (2020); Verma et al. (2022); Mao et al. (2024a).

Since $\ell_{\text{def}}$ is non-differentiable, the standard practice in online classification is to replace it with a convex surrogate $\Phi_{\text{def}}$ that upper bounds $\ell_{\text{def}}$, and then optimize $\Phi_{\text{def}}$ using Online Convex Optimization (OCO) (Hazan, 2016) methods such as Online Gradient Descent (OGD) (Zinkevich, 2003).

**Definition 4.2** (Varying Multiclass Surrogate Loss)**.** *For each expert set $A$, let $\Phi^A_{01} : \mathcal{H} \times \overline{\mathcal{X}} \times \overline{\mathcal{Y}}_A$ be the standard $\overline{\mathcal{Y}}_A$-class classification surrogate. The varying multiclass surrogate loss $\widetilde{\Phi}_{01} : \mathcal{H} \times \overline{\mathcal{X}} \times \overline{\mathcal{Y}}$ is defined as follows* $\widetilde{\Phi}_{01}(h, \overline{x}, j) = \begin{cases} \Phi^A_{01}(h, x, j) & \text{if } j \in \overline{\mathcal{Y}}_A, \\ 0 & \text{otherwise.} \end{cases}$

Intuitively, this definition guarantees that the surrogate $\widetilde{\Phi}_{01}$ is unaffected by the scores of experts that are not available. Building on this property, any multiclass surrogate loss $\widetilde{\Phi}_{01}$ can be naturally extended to obtain an online equivalent of Definition 3.3.

**Definition 4.3** (Surrogate Deferral Loss)**.** *, then for any $h \in \mathcal{H}$ and $(\overline{x}, y) = ((x, A), y) \in \overline{\mathcal{X}} \times \mathcal{Y}$,*

$$\Phi_{\text{def}}(h, \overline{x}, y) = \widetilde{\Phi}_{01}(h, \overline{x}, y) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y))\widetilde{\Phi}_{01}(h, \overline{x}, n + j).$$

### 4.3 Regret Formulation

We can now formalize the goal of the learner, which is to minimize the gap between the cumulative risk of his sequence of hypotheses, and the cumulative risk of the best hypothesis in hindsight. At every round $t$, the conditional $\ell_{\text{def}}$-risk is defined by $\mathcal{C}_{\ell_{\text{def}}}(h_t, \bar{x}_t) = \mathbb{E}_{y_t \sim p(\bar{x}_t)}[\ell_{\text{def}}(h_t, \bar{x}_t, y_t)]$. We can now analyze the cumulative $\ell_{\text{def}}$-risk $\mathcal{R}_{\ell_{\text{def}}}(\mathbf{h}, \bar{\mathbf{x}}) = \sum_{t \in [T]} \mathcal{C}_{\ell_{\text{def}}}(h_t, \bar{x}_t)$ and the optimal cumulative risk in hindsight $\mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\bar{\mathbf{x}}) = \inf_{h \in \mathcal{H}} \mathcal{R}_{\ell_{\text{def}}}(h, \bar{\mathbf{x}})$. The objective of the learner is to minimize the true deferral regret

$$R_{\ell_{\text{def}}}(T) = \mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \bar{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\bar{\mathbf{x}}). \tag{1}$$

In particular, the goal is to achieve *sublinear* regret scaling in $T$, that is $R_{\ell_{\text{def}}}(T) = o(T)$. This ensures that the average regret $\frac{R_{\ell_{\text{def}}}(T)}{T}$ converges to zero as $T \to \infty$, meaning that the learner's performance asymptotically approaches that of the best fixed strategy in hindsight.

The regret typically depends on the $(\ell_{\text{def}}, \mathcal{H})$-minimizability gap $\mathcal{M}_{\ell_{\text{def}}, \mathcal{H}}(\bar{\mathbf{x}}) = \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\bar{\mathbf{x}}) - \sum_{t \in [T]} \mathcal{C}^*_{\ell_{\text{def}}, \mathcal{H}}(\bar{x}_t)$. The minimizability gap measures the discrepancy between selecting the best hypothesis adaptively at each round and selecting the best single hypothesis in hindsight.

## 5 $\mathcal{H}$-CONSISTENCY BOUND FOR ONLINE LEARNING

A central property in batch learning is *consistency*, which ensures that minimizing a surrogate risk recovers the Bayes-optimal solution. This property has been extensively studied in the classical setting (Zhang, 2002; Steinwart, 2007; Bartlett et al., 2006), but remains largely unexplored in online learning, despite being crucial for understanding how well surrogate minimization minimizes the true loss. To address this gap, we introduce the notion of $\mathcal{H}$-consistency bounds in the online setting, drawing parallels to the bounds developed in Bartlett et al. (2006); Awasthi et al. (2022a); Mao et al. (2023c).

Formally, our goal is to relate the true deferral regret $R_{\ell_{\text{def}}}(T) = \mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})$ to the surrogate deferral regret $R_{\Phi_{\text{def}}}(T) = \mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})$. This reduction is valid whenever the underlying multiclass surrogate $\widetilde{\Phi}_{01}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$: that is, there exists a concave, non-decreasing calibration function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ and a tolerance $\epsilon \geq 0$ such that, for all $h \in \mathcal{H}$, $\overline{x} \in \overline{\mathcal{X}}$, and $p \in \mathcal{D}$,

$$\Delta\mathcal{C}_{\ell_{01}, \mathcal{H}}(h, \overline{x}) \mathbf{1}\{\Delta\mathcal{C}_{\ell_{01}, \mathcal{H}}(h, \overline{x}) > \epsilon\} \leq \Gamma\big(\Delta\mathcal{C}_{\widetilde{\Phi}_{01}, \mathcal{H}}(h, \overline{x})\big).$$

It follows that the surrogate provides a calibrated upper bound on the true deferral regret.

**Theorem 5.1.** *(Distribution-Dependent $\Gamma$-Bound) Assume that $\widetilde{\Phi}_{01}$ admits an $\mathcal{H}$-consistency bound w.r.t $\ell_{01}$ with a concave, non-decreasing calibration function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ with $\Gamma(0) = 0$. Then for any $\overline{\mathbf{x}} \in \overline{\mathcal{X}}^T$ and any distribution $p \in \mathcal{D}$, suppose the online learner generates corresponding hypotheses $\mathbf{h}$ and samples from $\mathbf{q}$ defined in Section 4.1, we have,*

$$\frac{R_{\ell_{\text{def}}}(T) + \mathcal{M}_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) - c\Big(\sum_{t \in [T]} \gamma_t\Big)}{T} - \epsilon \leq \max_{t \in [T]} S(\overline{x}_t)\, \Gamma\left(\frac{R_{\Phi_{\text{def}}}(T) + \mathcal{M}_{\Phi_{\text{def}}\mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t \in [T]} S(\overline{x}_t)}\right).$$

*Here, $S(\overline{x}_t) = \sum_{y \in \mathcal{Y}} p(\overline{x}_t, y) + \sum_{n+j \in \overline{\mathcal{Y}}_{A_t}} \Big(1 - \sum_{y \in \mathcal{Y}} p(\overline{x}_t, y) c_j(x_t, y)\Big)$ and $c$ is the upper bound on the expert cost $c = \max_j\{\overline{c}_j\}$.*

The proof is deferred to Appendix C. Note that the minimizability gap vanishes when the Bayes optimal cmulative risk coincides with the best-in-class cumulative risk, that is $\mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) = \mathcal{R}_{\ell_{\text{def}}, \mathcal{H}_{\text{all}}}(\overline{\mathbf{x}})$, this implies the following corollary.

**Corollary 5.2.** *Assume that $\widetilde{\Phi}_{01}$ admits an $\mathcal{H}$-consistency bound with respect to the multi-class zero-one classification loss $\ell_{01}$. Then, for all $h \in \mathcal{H}$, any $\overline{\mathbf{x}} \in \overline{\mathcal{X}}^T$ and any distribution such that $\mathcal{R}^*_{\Phi_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) = \mathcal{R}^*_{\Phi_{\text{def}}, \mathcal{H}_{\text{all}}}(\overline{\mathbf{x}})$, the following bound holds:*

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c\Big(\sum_{t \in [T]} \gamma_t\Big)}{T} - \epsilon \leq \max_{t \in [T]} S(\overline{x}_t)\, \Gamma\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t \in [T]} S(\overline{x}_t)}\right).$$

In the limit, this bound recovers an online analogue of the standard Bayes-consistency notion defined in Bartlett et al. (2006).

## 6 REGRET BOUNDS FOR ONLINE L2D

We can now derive explicit regret bounds from the H-Consistency Bound defined in the previous section.

Indeed, such bounds can be obtained directly by minimizing the right-hand side in Theorem 5.1.

Since the minimizability gap $\mathcal{M}_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})$ depends only on the hypothesis class $\mathcal{H}$ and the input sequence $\overline{\mathbf{x}}$, minimizing the right-hand side effectively becomes an online optimization problem of minimizing the surrogate deferral regret $R_{\Phi_{\text{def}}}(T)$ with bandit feedback. The resulting bound is especially tight when the best-in-class hypothesis coincides with the Bayes hypothesis, i.e., $\mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) = \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}_{\text{all}}}(\overline{\mathbf{x}})$, as Corollary 5.2 then implies a true deferral regret bound of $R_{\ell_{\text{def}}}(T) = O\Big(T\Gamma\Big(\frac{R_{\Phi_{\text{def}}}(T)}{T}\Big)\Big)$.

If we restrict the setting to bandit feedback with fixed number of experts, it becomes a special case of contextual multi-armed bandits (Auer et al., 2002; Li et al., 2010). A lower bound $R_{\ell_{\text{def}}}(T) = \Omega(\sqrt{T})$ is known (Daniely and Helbertal, 2013), but can only be reached by algorithms that are hard to implement in practice, such as EXP4 (Auer et al., 2002).

We show that even relaxing to a varying number of experts, we can still match this lower bound with an efficient implementation. The key idea is to select a surrogate loss $\Phi_{\text{def}}$ that admits a linear calibration function $\Gamma$ in Theorem 5.1, which yields the reduction $R_{\ell_{\text{def}}}(T) = O(R_{\Phi_{\text{def}}}(T))$. We then show that there exists an online algorithm with surrogate deferral regret $R_{\Phi_{\text{def}}}(T) = O(\sqrt{T})$, thereby achieving the desired bound $R_{\ell_{\text{def}}}(T) = O(\sqrt{T})$. In particular, we establish this result using the constrained hinge loss, and we present the corresponding algorithm and analysis in Section 6.1.

### 6.1 Constrained OCO with constrained hinge loss $\Phi^{\text{cstnd}}_{\text{hinge}}$

When the surrogate is chosen as the constrained hinge loss $\Phi^A_{01} = \Phi^{\text{cstnd}}_{\text{hinge}}$ (Lee et al., 2004), the induced surrogate deferral loss takes the form

$$\Phi_{\text{def}}(h,x,y) = \sum_{y'\in\overline{\mathcal{Y}}_A\setminus\{y\}} \widetilde{\Phi}_{\text{hinge}}(-h(x,y')) + \sum_{n+j\in\overline{\mathcal{Y}}_A}(1-c_j(x,y)) \sum_{y'\in\overline{\mathcal{Y}}_A\setminus\{n+j\}} \widetilde{\Phi}_{\text{hinge}}(-h(x,y')),$$

subject to the constraint $\sum_{y\in\overline{\mathcal{Y}}_A} h(\overline{x},y) = 0$. The resulting learning problem is therefore a constrained online convex optimization problem.

**Problem 1** (Constrained Online Convex Optimization)**.**

$$\begin{aligned} &\text{minimize} && R_{\Phi_{\text{def}}}(T) = \mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h},\bar{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\bar{\mathbf{x}}),\\ &\text{subject to} && \sum_{y\in\overline{\mathcal{Y}}_A} h_t(\bar{x}_t,y) = 0,\ \forall t\in[T]. \end{aligned}$$

Since $\mathcal{H}$ is the set of linear hypotheses, we can write $\mathcal{H} = \{h := (W,b)\in\mathbb{R}^{N\times(d+1)} | h(\overline{x},y) = \langle w_y, x\rangle + b_y\}$ where $w_y\in\mathbb{R}^d$ is the $y$-th row of $W$, and $b\in\mathbb{R}^N$ is the bias column. For ease of notation, denote $\widetilde{W} = (W,b)\in\mathbb{R}^{N\times(d+1)}$ and $\widetilde{w}_y$ the $y$-th row of $\widetilde{W}$. We use $h$ and $\widetilde{W}$ interchangeably when it is clear from the context. We place a bound on hypothesis's norm, that is $\mathcal{H} = \{\widetilde{W}\in\mathbb{R}^{N\times(d+1)} \,|\, ||\widetilde{W}||_F \le B\}$. We note that previous results for $\mathcal{H}$-consistency bound w.r.t constrained hinge loss still holds for our hypothesis set $\mathcal{H}$ with a suitably chosen bound $B$. The discussion is deferred to Appendix B.

For each expert set $A\in\mathcal{A}$, we define $\mathcal{K}_A = \{h\in\mathcal{H}_{all} | \sum_{y\in\overline{\mathcal{Y}}_A} h(\overline{x},y) = 0,\ \forall\overline{x}\in\overline{\mathcal{X}}\}$ as the unbounded minimum feasible hypothesis set w.r.t expert set $A$. That is $\mathcal{K}_A$ is the set of weight matrices $\widetilde{W}$ whose column sums to 0 over rows in $\overline{\mathcal{Y}}_A$. Denote $h|_{\mathcal{K}_A}$ the projection of a hypothesis $h\in\mathcal{H}$ onto $\mathcal{K}_A$.

Our setting differs from previous constrained OCO problems in that there potentially exists only the trivial hypothesis in the intersection of all feasible sets, that is, $\cap_{t\in[T]}\mathcal{K}_{A_t} = \{\mathbf{0}\}$ in the worst case. This renders previous approaches, which rely on minimizing regret against the optimal feasible hypothesis $\mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) = \inf_{h\in\cap_{t\in[T]}\mathcal{K}_{A_t}} \mathcal{R}_{\Phi_{\text{def}}}(h,\overline{\mathbf{x}})$, ineffective in our problem. To address this, we devise a variant of the OGD algorithm that allows us to minimize regret against the best hypothesis after projection in the original hypothesis set $\mathcal{H}$. In other words we define $\mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) = \inf_{h\in\mathcal{H}} \mathcal{R}_{\Phi_{\text{def}}}\Big(\big(h|_{\mathcal{K}_{A_t}}\big)_{t=1}^T, \overline{\mathbf{x}}\Big)$. The key idea is to utilize important properties of projections onto $\mathcal{K}_A$, which are described in the following proposition.

**Proposition 6.1.** *(Properties of $\Pi_{\mathcal{K}_A}$) For any linear hypothesis $h = \widetilde{W}\in\mathcal{H}$ and expert set $A\in\mathcal{A}$, we have the following properties:*

1. *$\widetilde{W}|_{\mathcal{K}_A} = \widetilde{W} - \mu^T\mathbf{1}_{\overline{\mathcal{Y}}_A} = (\widetilde{w}_i|_{\mathcal{K}_A})_{i=1}^N$, where $\mu = \frac{1}{|\overline{\mathcal{Y}}_A|}\sum_{i\in\overline{\mathcal{Y}}_A}\widetilde{w}_i$, $\widetilde{w}_i|_{\mathcal{K}_A} = \widetilde{w}_i - \mu$ if $i\in\mathcal{Y}_A$ and $\widetilde{w}_i|_{\mathcal{K}_A} = \widetilde{w}_i$ otherwise.*
2. *(Maxima preserving) For any $\overline{x} = (x,A)\in\overline{\mathcal{X}}$, then $h(\overline{x}) = h|_{\mathcal{K}_A}(\overline{x})$.*

Proposition 6.1 implies that projection onto $\mathcal{K}_A$ preserves the hypothesis's predictions on rounds whose label set is $\overline{\mathcal{Y}}_A$. As a result, if $\mathbf{h} = (h_1,h_2,\ldots,h_T)$ is any sequence of hypotheses generated, then $\overline{\mathbf{h}} = \big(\Pi_{\mathcal{K}_{A_t}}(h_t)\big)_{t=1}^T$. That is the sequences of projected hypotheses, satisfies $\mathcal{R}_{\ell_{\text{def}}}(\mathbf{h},\overline{\mathbf{x}}) = \mathcal{R}_{\ell_{\text{def}}}(\overline{\mathbf{h}},\overline{\mathbf{x}})$. In particular, when $\mathbf{h} = (h^*)_{t=1}^T$ is the fixed sequence of best-in-class hypothesis, this implies that the objective true deferral regret $R_{\ell_{\text{def}}}(T)$ is preserved, that is $\mathcal{R}_{\ell_{\text{def}}}(\mathbf{h},\overline{\mathbf{x}}) = \mathcal{R}_{\ell_{\text{def}}}(\overline{\mathbf{h}},\overline{\mathbf{x}}) = \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})$.

**Algorithm 1** OGD with Projected Gradients

**Input:** Learning rate $\eta_t > 0$, and exploration rate $\gamma_t$
**Initialize:** $\widetilde{W}_1 = \mathbf{0} \in \mathbb{R}^{N\times(d+1)}$
**for** $t = 1$ **to** $T$ **do**
  Obtain $\overline{x}_t = (x_t, A_t)$
  Project $\widetilde{W}_t$ to feasible hypothesis set: $\widetilde{W}_t|_{\mathcal{K}_{A_t}} = \Pi_{\mathcal{K}_{A_t}}(\widetilde{W}_t)$
  Let $y_t^* = \arg\max_y \langle w_{t,y}, x_t\rangle + b_t$
  Set $q_t = (1-\gamma_t)\mathbf{e}_{y_t^*} + \frac{\gamma_t}{|\overline{\mathcal{Y}}_{A_t}|}\mathbf{1}$
  Predict with label $y_t' \sim q_t$
  Obtain feedback and compute $\widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}})$ accordingly
  Compute gradient $\widehat{\nabla}_t = \nabla\widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}})$ and its projection onto $\mathcal{K}_{A_t}$: $\widehat{\nabla}_t|_{\mathcal{K}_{A_t}} = \Pi_{\mathcal{K}_{A_t}}(\widehat{\nabla}_t)$
  Update $\widetilde{W}_{t+1} = \widetilde{W}_t - \eta_t\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}$
**end for**

We now introduce our algorithm, with full details provided in Algorithm 1. At round $t$, let $w_{t,y}$ denote the $y$-th row of $W_t$, and let $\widehat{\Phi}_t$ denote the estimated loss function constructed from the available feedback. In the full-feedback setting, the estimate coincides with the true surrogate deferral loss, i.e., $\widehat{\Phi}_t\Big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}\Big) = \Phi_{\text{def}}\Big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t\Big)$. In the bandit-feedback setting, we have

$$\widehat{\Phi}_t(\widetilde{W}) = \sum_{n+j\in\overline{\mathcal{Y}}_{A_t}} v_{t,j}\widetilde{\Phi}_{01}(\widetilde{W}, \overline{x}_t, n+j)(1-c_j(x_t,y_t)) + v_{t,0}\widetilde{\Phi}_{01}(\widetilde{W}, \overline{x}_t, y_t'), \forall \widetilde{W} \in \mathcal{H}.$$

When $\Phi_{01}^A = \Phi_{\text{hinge}}^{\text{cstnd}}$, let $y_t'$ denote the learner's randomized prediction sampled from $q_t$, and define $v_{t,0} = \frac{\mathbf{1}\{y_t'\in[n]\}\mathbf{1}\{y_t'=y_t\}}{q_t(y_t')}$ and $v_{t,j} = \frac{\mathbf{1}\{y_t'=n+j\}}{q_t(n+j)}$. The learner's exploration rate $\gamma_t$, introduced in Section 4.1, is set to $\gamma_t = 0$ in the full-information case, and to $\gamma_t = \min\{\frac{1}{2}, \frac{\gamma}{\sqrt{t}}\}$ in the bandit case, for some constant $\gamma > 0$ specified later.

Note that $\widehat{\Phi}_t$ is always convex and $\mathbb{E}_{y_t'\sim q_t}\left[\widehat{\Phi}_t(\widetilde{W})\right] = \Phi_{\text{def}}\Big(\widetilde{W}, \overline{x}_t, y_t\Big)$ regardless of the feedback scenario. Suppose $h^* = \widetilde{W}^*$ is the best-in-class hypothesis w.r.t surrogate deferral loss $\Phi$, i.e. $h^* = \arg\min_{h\in\mathcal{H}} \mathcal{R}_{\Phi_{\text{def}}}\left(\left(h|_{\mathcal{K}_{A_t}}\right)_{t=1}^T, \overline{\mathbf{x}}\right)$. For any expert set $A$, denotes $\mu_A^* = \frac{1}{|\overline{\mathcal{Y}}_A|}\sum_{i\in\overline{\mathcal{Y}}_A}\widetilde{w}_i^*$, the following lemma gives an upper bound on the surrogate loss difference between learner's generated hypothesis and best-in-class hypothesis.

**Lemma 6.2.** *Let $\widehat{\nabla}_{t,i}$ be the gradient of the $i$-th row of $\widehat{\nabla}_t = \nabla\widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}})$ in round $t\in[T]$, and $\widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}$ the projected gradient onto $\mathcal{K}_{A_t}$, that is $\widehat{\nabla}_t|_{\mathcal{K}_{A_t}} = (\widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}})_{i=1}^N$. We have*

$$\widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}}) - \widehat{\Phi}_t(\widetilde{W}^*|_{\mathcal{K}_{A_t}}) \le \sum_{i\in\overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^*\rangle.$$

The proof is deferred to Appendix E. By applying Lemma 6.2 to every round and summing up, we obtain the following bound on the surrogate regret.

**Theorem 6.3.** *(Surrogate regret bound) Running Algorithm 1 with $\eta_t = \frac{B}{\sqrt{NR}\,t^{2/3}}$, $\gamma_t = \min\left\{\frac{1}{2}, \frac{1}{t^{1/3}}\right\}$ for $t = 1,\dots,T$, we have the following surrogate regret bound*

$$\sum_{t\in[T]}\left(\Phi_{def}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t) - \Phi_{def}(\widetilde{W}^*|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)\right) = O\big(N^{3/2}B\,R\,T^{2/3}\big).$$

The key idea is to bound $\langle\widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^*\rangle$ together on rounds $t\in[T]$ in which label $i$ is available and apply Lemma 6.2. The proof is analogous to Theorem 2.13 in Orabona (2019) and is deferred to Appendix E.1.

Note that by Proposition 6.1, we have

$$\sum_{t\in[T]}\left(\ell_{\text{def}}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t) - \ell_{\text{def}}(\widetilde{W}^*|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)\right) = \sum_{t\in[T]}\left(\ell_{\text{def}}(\widetilde{W}_t, \overline{x}_t, y_t) - \ell_{\text{def}}(\widetilde{W}^*, \overline{x}_t, y_t)\right)$$

Combining Corollary 5.2 and Theorem 6.3, we achieve the following true deferral regret bound.

**Theorem 6.4.** *(True Deferral Regret Bound) Under the same assumptions in Corollary 5.2 and Theorem 6.3. Running Algorithm 1 with $\eta_t = \frac{B}{\sqrt{N}Rt^{2/3}}$, $t = 1,\dots,T$, $\gamma_t = \min\left\{\frac{1}{2}, \frac{1}{t^{1/3}}\right\}$ gives us the following expected regret bound*

$$R_{\ell_{\text{def}}}(T) = \sum_{t\in[T]}\left(\ell_{def}(q_t, \overline{x}_t, y_t) - \ell_{def}(\widetilde{W}^*, \overline{x}_t, y_t)\right) = O\big(N^{3/2}BRT^{2/3}\big).$$

*Proof.* Note that when $\gamma_t = \min\left\{\frac{1}{2}, \frac{1}{t^{1/3}}\right\}$, $\sum_{t\in[T]}\frac{c\gamma_t}{T} = O(T^{-1/3})$. Hence, applying Corollary 5.2 and Theorem 6.3 directly gives us the desired bound.

**Near-Realizable Assumption.** We make use of the following definition for any $\overline{x} = (x, A) \in \overline{\mathcal{X}}$ and $y \in \overline{\mathcal{Y}}$

$$s(\overline{x}, y) = \begin{cases} p(\overline{x}, y) & \text{if } y \in \mathcal{Y}, \\ 1 - \sum_{y'\in\mathcal{Y}} p(\overline{x}, y')c_j(x, y') & \text{if } y = n+j. \end{cases}$$

To obtain the probability representation, we denote $\overline{s}(\overline{x}, y) = \frac{s(\overline{x},y)}{S(\overline{x})}$ the normalized counterparts of $s(\overline{x}, y)$,

where $S(\overline{x}) = \sum_{y \in \overline{\mathcal{Y}}_A} s(\overline{x}, y)$. Intuitively, $s(\overline{x}, y)$ indicates the expected "score" of label $y$ given input $\overline{x}$. By making further assumptions on the score distribution given an input $\overline{x}$, we can improve the $O(T^{2/3})$ regret bound to $O(\sqrt{T})$. Given input $\overline{x} \in \overline{\mathcal{X}}$, let $y_{\max}$ be the label with highest score, that is $y_{\max} = \arg\max_{y \in \mathcal{Y}} \overline{s}(\overline{x}, y)$. Under the assumption that $\overline{s}(\overline{x}_t, y_{\max})$ is close to 1, we can show the following regret result.

**Theorem 6.5.** *(Near-Realizable Assumption) Under the same assumptions in Corollary 5.2 and Theorem 6.3, suppose that for any $t \in [T]$ we have $s(\overline{x}_t, y_{\max}) \geq 1 - 1/\sqrt{T}$. Running Algorithm 1 with $\eta_t = \frac{B}{\sqrt{4RN^{1/2}(N+4)t}}$ and $\gamma = B\sqrt{RN^{5/2}(N+4)}$ gives us the following regret deferral bound*

$$R_{\ell_{\text{def}}}(T) = \sum_{t \in [T]} \left( \ell_{def}(q_t, \overline{x}_t, y_t) - \ell_{def}(\widetilde{W}^*, \overline{x}_t, y_t) \right) = O\left( B\sqrt{R^3 N^{7/2} T} \right)$$

The proof works by giving a tighter analysis of the gap between $\Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(q, \overline{x})$ and $\Delta\mathcal{C}_{\Phi_{\text{def}}, \mathcal{H}}(h, \overline{x})$. The details are deferred to Appendix E.2.

**Remark 6.6.** *Further analysis of the minimizability gap is conducted in Appendix E.3.*

# 7 EXPERIMENTS

Expert systems in deployment rarely operate under fixed conditions. Human and automated experts differ not only in accuracy but also in when they can be consulted, and both factors fluctuate over time due to workload, fatigue, or shifting domains. To examine how an online Learning-to-Defer algorithm adapts under such nonstationarity, we vary two fundamental axes of uncertainty: expertise and availability. This yields three regimes of increasing complexity—ranging from fully static to fully dynamic—allowing us to isolate the effect of each source of variability.

## 7.1 Overall Setting

Our experiments study the effectiveness of Algorithm 1 under a variety of expert settings described in Table 1, both for synthetic and real-world datasets. Each expert $g_j$ has expert cost $c_j(x, y) = \mathbf{1}\{g_j(x) \neq y\} + 0.1$.

Further details of each setting are deferred to Appendix F.1. Notice that the third setting is especially challenging, as a given expert is allowed to drift between 100% accuracy and totally random over time. We show that across all settings, our algorithm correctly adapts to varying expert expertise and availability. It consistently learns the right boundary between deferring and predicting, and preserves high accuracy with low deferral loss. In the main paper, we focus on the most challenging setting and report results on the Reuters4 dataset (Kakade et al., 2008). Additional experiments on Reuters4, synthetic datasets, and CIFAR10H are deferred to Appendix F.

Table 1: Characteristics of the experts in our different settings

| Varying Expertise | Varying Availability | Setting Description |
|---|---|---|
| ✗ | ✗ | The experts have constant expertise and are always available. |
| ✗ | ✓ | Expertise is fixed, availability varies. |
| ✓ | ✓ | Both expertise and availability vary (fully dynamic setting). |

## 7.2 Real-world Dataset: Reuters4

For real-world datasets, we make use of the Reuters4 dataset, commonly used for bandit classification experiments (Kakade et al., 2008; Crammer and Gentile, 2013; Beygelzimer et al., 2017). This dataset is generated from the RCV1 dataset (Lewis et al., 2004), extracting 685,071 examples that have exactly one label from the set {CCAT, ECAT, GCAT, MCAT}. It contains 47,236 features and 4 classes, denoted 1,2,3,4. For this experiment, we use the first 100,000 examples.

We generate 3 experts with overlapping knowledge regions. At initialization $t = 0$, expert $g_j$ is knowledgeable on labels $\{j, j+1\}$, $j = 1, 2, 3$. That is, expert $g_j$ predicts correctly on inputs from classes $\{j, j+1\}$ while predicting uniformly at random on other classes. These regions remain fixed for the first two settings, and evolve in the final setting.

We slightly modify the update step of Algorithm 1 by the one from the Adagrad algorithm (Duchi et al., 2011) with base learning rate 0.1 and exploration rate $\gamma_t = \min\{\frac{1}{2}, \frac{10}{\sqrt{t}}\}$.

**Results under Drifting Availability and Drifting Expertise.** Figure 2 summarizes our results. The varying expert expertise is shown in Figure 1. Even under challenging conditions with simultaneously varying availability and expertise, our approach maintains strong performance and adapts effectively to the evolving environment.

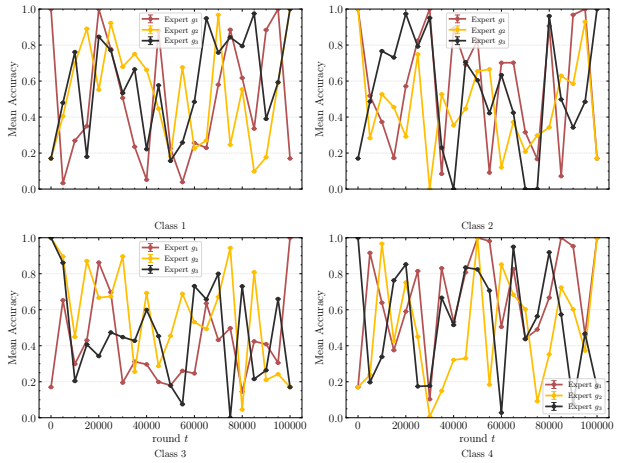
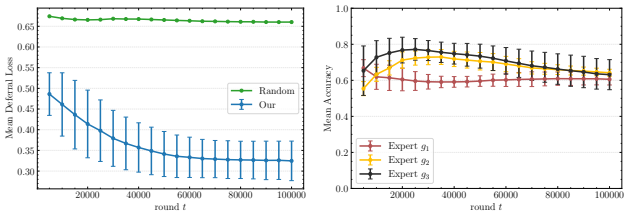



Figure 1: Expert Accuracies by Regions on Reuters4.

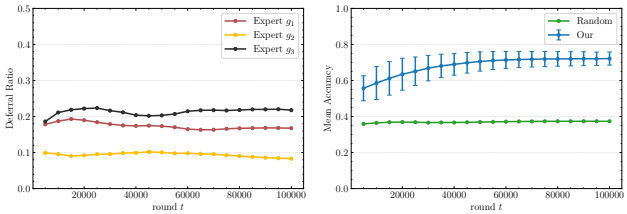
Figure 1 illustrates the evolution of expert expertise over time. In particular, expert $g_1$ gradually shifts from classes $\{1, 2\}$ to classes $\{3, 4\}$, expert $g_2$ shifts from classes $\{2, 3\}$ to classes $\{1, 4\}$, and expert $g_3$ shifts from classes $\{3, 4\}$ to classes $\{1, 2\}$. Figure 2 indicates that our method tracks these moving regions, defers with high frequency to the currently competent experts, and preserves stable calibration as expertise evolves. Therefore, overall task accuracy and deferral loss remain strong and comparable to the fixed-expertise baseline (reported in Appendix F.3) throughout the drift.

Remaining Reuters4 experiments are deferred to Appendix F.3. We also include experiments on synthetic datasets in Appendix F.2. These experiments are designed to closely mirror the Reuters4 setting, while providing a more controlled environment in which the optimal deferral behavior is explicit. As such, they provide additional evidence for the same conclusions suggested by the Reuters4 results. Finally, Appendix F.4 reports an image classification experiment on CIFAR10H, where the predictor is a WideResNet rather than a linear hypothesis and the experts are constructed from real human annotations. In this setting, we also compare against a confidence-based baseline that defers uniformly at random to an available expert whenever the classifier confidence falls below a fixed threshold. The results again show that our method achieves lower deferral loss and higher accuracy, indicating that the learned policy not only defers under uncertainty, but also makes better use of the available experts.



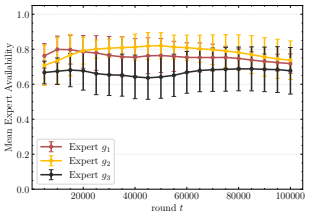

(a) Average true deferral loss.

(b) Accuracy of queried experts.




(c) Expert deferral ratio.

(d) Average accuracy.




(e) Expert availability.

Figure 2: Results of experiment on Reuters4 for **setting 3: drifting availability and drifting expertise**. Whiskers denote standard deviations computed over 5 independent runs.

## 8 CONCLUSION

We introduced the first framework for *Online Learning-to-Defer* (L2D) with bandit feedback and dynamically varying experts. Our analysis extends classical $\mathcal{H}$-consistency bounds to the online setting and establishes sublinear regret guarantees of $O((n + n_e)T^{2/3})$ in the general case and $O((n+n_e)\sqrt{T})$ under a near-realizable condition. These results demonstrate that consistent surrogate-based methods can be made robust to both adversarial environments and varying experts. We validated the practicality of our algorithm across synthetic and real-world datasets, showing that it maintains strong performance even under adversarial expert availability and drifting expertise.

## Acknowledgments

Yannis Montreuil and Maxime Meyer are supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2023-01-041-J and AISG Award No: AISG3-PhD-2026-01-068T). Yannis Montreuil is also supported by A*STAR, and is part of the programme DesCartes 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


Naman Agarwal, Satyen Kale, and Julian Zimmert. Efficient methods for online multiclass logistic regression. In *International Conference on Algorithmic Learning Theory*, pages 3–33. PMLR, 2022.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77, 2002.

Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses, 2021. URL https://arxiv.org/abs/2104.09658.

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, 2022a. Curran Associates Inc. ISBN 9781713871088.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class $h$-consistency bounds. *Advances in neural information processing systems*, 35: 782–795, 2022b.

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.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9: 1823–1840, June 2008.

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

Alina Beygelzimer, Francesco Orabona, and Chicheng Zhang. Efficient online bandit multiclass learning with $\tilde{O}(\sqrt{T})$ regret. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 488–497. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/beygelzimer17a.html.

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.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, January 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Koby Crammer and Claudio Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104. PMLR, 2013.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on learning theory*, pages 167–208. PMLR, 2018.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Elad Hazan. 2016. doi: 10.1561/2400000013.

Elad Hazan and Satyen Kale. Newtron: an efficient bandit algorithm for online multiclass prediction. *Advances in neural information processing systems*, 24, 2011.

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.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Conference on Learning Theory*, pages 2085–2108. PMLR, 2020.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Mixability made efficient: Fast online multiclass logistic regression. *Advances in Neural Information Processing Systems*, 34:23692–23702, 2021.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 05 2016. doi: 10.1038/sdata.2016.35.

Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:31, 01 2023. doi: 10.1038/s41597-023-01945-2.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447, 2008.

Varun Kanade and Thomas Steinke. Learning hurdles for sleeping experts. *ACM Transactions on Computation Theory (TOCT)*, 6(3):1–16, 2014.

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.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

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

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

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.

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. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023c.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *ISAIM*, 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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

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.

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.

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, 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*, 2026b. URL https://openreview.net/forum?id=kEVupwepTq.

Yannis Montreuil, Yu Letian, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in one-stage learning-to-defer. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026c. URL https://openreview.net/forum?id=gF8tv1SaR2.

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.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

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.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.

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 Alison Noble. Towards human-AI collaboration in healthcare: Guided deferral systems with large language models. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=4c5rg9y4me.

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.

Dirk van der Hoeven. Exploiting the surrogate gap in online multiclass classification. *Advances in Neural Information Processing Systems*, 33:9562–9572, 2020.

Dirk van der Hoeven, Federico Fusco, and Nicolò Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. *Advances in neural information processing systems*, 34:13280–13291, 2021.

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.

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.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

## Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

**In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.**

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] The mathematical setting and assumptions are described in Sections 3 and 4. Our main algorithm is clearly described in Algorithm 1.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] It is provided in Appendix.
   (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]
   (b) Complete proofs of all theoretical results. [Yes
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [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 Discussion of benchmarks

The objective in online learning is to minimize regret, the most general formulation of which is

$$R(T) = \mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}_{\ell_{\text{def}}}(\mathbf{h}^*, \overline{\mathbf{x}}).$$

Where $\mathbf{q} = (q_1, q_2, \ldots, q_T)$ is the learner's prediction distribution at round $t$ (defined in Section 4.1 and $\mathbf{h}^* = (h_1^*, h_2^*, \ldots, h_T^*) \in \mathcal{H}^T$ is the *best* sequence of hypotheses. How $\mathbf{h}^*$ is defined directly implies the difficulty of our problem. We present and discuss some natural definitions below

**No Restrictions**. Note that if no restriction is placed on $h_t^*$, then the regret is always $R(T) = \Omega(T)$ for any online algorithm. To see this, consider the setting where in each round $t \in [T]$, the adversary chooses $y_t$ uniformly at random. The optimal sequence of hypotheses is $h_t^*$ such that $h_t^*(\overline{x}_t) = y_t$. Assuming all experts are wrong, the learner incurs an expected regret of at least $(n-1)/n$ each round while the comparator $h_t^*$ incurs zero loss.

**Policy Regret**. We consider a more restrictive notion, called policy regret (adapted from Kanade and Steinke (2014), van der Hoeven et al. (2021)). We no longer allow $h_t^*$ to adapt to input-label $(x_t, y_t)$ in each round, but only to the expert set $A_t \in \mathcal{A}$. In other words, $h_t^*(\overline{x}_t) = h_{A_t}(\overline{x}_t)$, where $h_A \in \mathcal{H}$ is the best-in-class hypothesis over all rounds with expert set $A$. That is $h_A = \arg\min_{h \in \mathcal{H}} \sum_{t \in [T], \overline{x}_t = (x_t, A)} \ell_{\text{def}}(h, \overline{x}_t, y_t)$. Intuitively, $\mathbf{h}^*$ is determined by the best policy mapping from the set of expert sets $\mathcal{A}$ to the set of best-in-class hypotheses. Note that we assume $\overline{\mathbf{x}}$ can be chosen adversarially. Under such assumptions, we show that the regret bound for policy regret grows exponentially w.r.t the number of experts $n_e$. To see this, we first provide a lower bound for the standard online bandit classification problem, where input-label pairs are chosen adversarially.

**Theorem A.1.** *(Daniely and Helbertal (2013), Theorem 4) Consider the standard online bandit classification problem with $n \geq 2$ classes and adversarial assumptions. For any (possibly randomized) online algorithm and any integer $T \geq 1$, there exists an environment $E$ in which the algorithm must incur a true regret of at least $\Omega(\sqrt{nT})$.*

Note that while input-label pairs in our setting are chosen semi-adversarially, we make no assumptions about experts' performance. By applying Theorem A.1, we can show a lower bound on our problem when the number of experts is fixed, i.e. $A_t = \{g_1, g_2, \ldots, g_{n_e}\}$, $\forall t \in [T]$.

**Corollary A.2.** *Consider the online learning-to-defer problem with bandit feedback a fixed number of experts $n_e \geq 2$. For any (possibly randomized) online algorithm and any integer $T \geq 1$, there exists an environment $E$ in which the algorithm must incur an expected true regret of at least $\Omega(\sqrt{n_e T})$.*

*Proof.* Consider a uniform label distribution $p \in \mathcal{D}$, that is, $p(\overline{x}, y) = \frac{1}{n}$ $\forall \overline{x} \in \overline{\mathcal{X}}, y \in \mathcal{Y}$. Any algorithm that picks a true label incurs an expected cost of $\frac{n-1}{n}$. By Theorem A.1, there exists an environment $E$ (for the bandit classification problem) in which the algorithm must incur a true regret of at least $\Omega(\sqrt{n_e T})$. In other words, there exist a sequence $(x_t, y_t)_{t \in [T]}$ and a fixed comparator $h^* \in \mathcal{H}$ such that

$$\sum_{t \in [T]} (\mathbf{1}\{y_t' \neq y_t\} - \mathbf{1}\{h^*(x_t) \neq y_t\}) = \Omega(\sqrt{n_e T}),$$

where $y_t'$ is the algorithm's (possibly randomized) predictions. Now we consider an environment $E'$ (for the bandit learning-to-defer problem), with label distribution $p$ defined above and experts performance defined as follows: In round $t \in [T]$, expert cost of $g_{y_t}$ is 0, while the other experts suffer cost $\frac{n-1}{n}$. Define the new comparator $\overline{h^*}$ such that $\overline{h^*}(x_t) = h^*(x_t) + n$. Then over all rounds $t \in [T]$, the algorithm suffers a cumulative cost of at least

$$\sum_{t \in [T]} \min\{\frac{n-1}{n}, \mathbb{E}_{y \sim p_t}[c_{y_t'}(x_t, y)]\} \leq \sum_{t \in [T]} \mathbb{E}_{y \sim p_t}[c_{y_t'}(x_t, y)] \leq \mathbb{E}_{y \sim p_t}[c_{\overline{h^*}(x_t)}(x_t, y)] + \Omega(\sqrt{n_e T}),$$

which ends our proof. □

Now we are ready to show the following lower bound in the general varying number of expert case, which has an exponential dependence on $n_e$

**Theorem A.3.** *Consider the online learning-to-defer problem with bandit feedback and varying experts. Suppose the number of unique experts after $T$ rounds is $n_e \geq 2$ and let $\mathcal{A}' = \{A \in \mathcal{A} : |A| \geq 2\}$. For any (possibly randomized) online algorithm, there exists an environment $E$ in which the algorithm must incur an expected policy regret of at least $\Omega(\sqrt{T \min\{T, |\mathcal{A}'|\}})$.*

*Proof.* In the case of $|\mathcal{A}| = \min\{T, |\mathcal{A}|\}$, we divide $T$ into $|\mathcal{A}|$ segments $T_1, T_2, \ldots, T_{|\mathcal{A}|}$ (WLOG assume $|\mathcal{A}|$ divides $T$). Each segment $T_i$ is of length $m = T/|\mathcal{A}|$ and corresponds to a unique expert set $A_i$ where $\{i(g) : g \in A_i\} \in 2^{\{n+1,\ldots,n+n_e\}}$ and $|A_i| \geq 2$. For each segment $T_i$, based on Corollary A.2, there exists an environment $E_i$ and a comparator $h_i^*$ such that

$$\sum_{t \in T_i} (\ell_{\text{def}}(q_t, \overline{x}_t, y_t) - \ell_{\text{def}}(h_i^*, \overline{x}_t, y_t)) = \Omega(\sqrt{|T_i|}).$$

Choosing $h(x, A_i) = h_i^*$ and consider the environment $E$ as the combination of $E_i$, we have

$$R(T) \geq \sum_{i \in \mathcal{A}} \sum_{t \in T_i} (\ell_{\text{def}}(q_t, \overline{x}_t, y_t) - \ell_{\text{def}}(h_i^*, \overline{x}_t, y_t)) = \sum_{i \in \mathcal{A}} \Omega(\sqrt{|T_i|}) = \Omega(\sqrt{T|\mathcal{A}|}).$$

In the case $T = \min\{T, |\mathcal{A}|\}$, the adversary can choose a different expert set $A_t$ for each round. In round $t$, based on Corollary A.2, there exists an environment $E_t$ and a comparator $h_t^*$ such that the expected regret is at least some constant $c > 0$. Choosing $h^* = h_t^*$ and consider the environment $E$ as the combination of $E_t$, the expected regret is at least $cT \in \Omega(T)$, which ends our proof. □

As a result, to further relax this notion, we consider the overall best-in-class hypothesis as presented in our paper.

# B Discussion of bound B on the hypothesis set

In this section, we present a suitable choice for bound B. We make use of the characteristic of the minimizability gap for the constrained hinge loss. From Theorem E.5, we have shown that the minimal conditional surrogate $\Phi_{\text{def}}$-risk can be obtained by a hypothesis of the form

$$h(\overline{x}, y) = \begin{cases} |\overline{\mathcal{Y}}_A| - 1 & \text{if } y = y_{\max}, \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, it suffices to show that there exists a bound $B$ on the hypothesis set $\mathcal{H}$ such that for any $\overline{x}, y$, such hypothesis $h$ defined in (2) belongs in $\mathcal{H}$. Given any input $\overline{x} = (x, A) \in \overline{\mathcal{X}}$. Let $W|^{\overline{\mathcal{Y}}_A}$ be weight matrix $W$ restricted to $\overline{\mathcal{Y}}_A$. Then we can choose $\widetilde{W} = (W, b) \in \mathcal{H}$ such that $W|^{\overline{\mathcal{Y}}_A} = \frac{yx^\intercal}{||x||_2}, b = 0$. This gives us $W|^{\overline{\mathcal{Y}}_A} x + b = y$ and $(W|^{\overline{\mathcal{Y}}_A}, b|^{\overline{\mathcal{Y}}_A}) \in \mathcal{K}_A$. Furthermore, $||W|^{\overline{\mathcal{Y}}_A}||_F = \frac{||y||_2}{||x||_2} \leq |\mathcal{Y}_A| \leq N$. From this, a suitable bound $B$ is $B = N$.

**Projection onto $\mathcal{K}_A$ preserves bound**. Given $\widetilde{W} \in \mathcal{H}$, it is clear that $||\widetilde{W}|_{\mathcal{K}_A}||_F \leq ||\widetilde{W}||_F \leq B$, so $\widetilde{W}|_{\mathcal{K}_A} \in \mathcal{H}$.

# C Proof of $\mathcal{H}$-consistency bound for online learning

For any $\overline{x} = (x, A) \in \overline{\mathcal{X}}$, we denote by $H(\overline{x})$ the set of labels generated by hypotheses in $\mathcal{H}$: $H(\overline{x}) = \{h(\overline{x}) : h \in \mathcal{H}\} \cap \overline{\mathcal{Y}}_A$. With this, we can write the conditional risk and calibration gap of the deferral loss as follows.

## C.1 Calibration gap for true deferral loss

**Lemma C.1.** *For any $\overline{x} \in \overline{\mathcal{X}}$, the minimal conditional $\ell_{\text{def}}$-risk and the calibration gap for $\ell_{\text{def}}$ can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{x}) = 1 - \max_{y \in H(\overline{x})} s(\overline{x}, y),$$

$$\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h,\overline{x}) = \max_{y\in H(\overline{x})} s(\overline{x},y) - s(\overline{x},h(\overline{x})).$$

*Proof.* The conditional $\ell_{\text{def}}$-risk of $h$ can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\ell_{\text{def}}}(h,\overline{x}) &= \mathbb{E}_{y\sim p(\overline{x})}\big[\ell_{\text{def}}(h,\overline{x},y)\big] \\
&= \mathbb{E}_{y\sim p(\overline{x})}\left[\mathbf{1}\{h(\overline{x})\neq y\}\mathbf{1}\{h(\overline{x})\in[n]\} + \sum_{n+j\in\overline{\mathcal{Y}}_A} c_j(x,y)\mathbf{1}\{h(\overline{x})=n+j\}\right] \\
&= \sum_{y\in\mathcal{Y}} s(\overline{x},y)\mathbb{1}\{h(\overline{x})\neq y\}\mathbf{1}\{h(\overline{x})\in[n]\} + \sum_{n+j\in\overline{\mathcal{Y}}_A}\big(1-s(\overline{x},n+j)\big)\mathbf{1}\{h(\overline{x})=n+j\} \\
&= \big(1-s(\overline{x},h(\overline{x}))\big)\mathbf{1}\{h(\overline{x})\in[n]\} + \sum_{n+j\in\overline{\mathcal{Y}}_A}\big(1-s(\overline{x},h(\overline{x}))\big)\mathbf{1}\{h(\overline{x})=n+j\} \\
&= 1-s(\overline{x},h(\overline{x})).
\end{aligned}$$

Then, the minimal conditional $\ell_{\text{def}}$-risk is given by

$$\mathcal{C}^*_{\ell_{\text{def}}}(\mathcal{H},\overline{x}) = 1 - \max_{y\in H(\overline{x})} s(\overline{x},y),$$

and the calibration gap can be expressed as follows:

$$\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h,\overline{x}) = \mathcal{C}_{\ell_{\text{def}}}(h,\overline{x}) - \mathcal{C}^*_{\ell_{\text{def}}}(\mathcal{H},\overline{x}) = \max_{y\in H(\overline{x})} s(\overline{x},y) - s(\overline{x},h(\overline{x})),$$

which completes the proof. □

### C.2 Calibration gap for surrogate deferral risk

**Lemma C.2.** *For any $\overline{x}\in\overline{\mathcal{X}}$, the conditional surrogate $\Phi_{def}$-risk and calibration gap can be expressed as follows:*

$$\mathcal{C}_{\Phi_{def}}(h,\overline{x}) = \sum_{y\in\overline{\mathcal{Y}}_A} s(\overline{x},y)\ell(h,\overline{x},y),$$

$$\Delta\mathcal{C}_{\Phi_{def}}(h,\overline{x}) = \sum_{y\in\overline{\mathcal{Y}}_A} s(\overline{x},y)\Phi_{01}(h,\overline{x},y) - \inf_{h\in\mathcal{H}}\sum_{y\in\overline{\mathcal{Y}}_A} s(\overline{x},y)\Phi_{01}(h,\overline{x},y).$$

*Proof.* By definition, the conditional $\Phi_{\text{def}}$-risk $\mathcal{C}_{\Phi_{\text{def}}}(h,\overline{x})$ can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi_{\text{def}}}(h,\overline{x}) &= \mathbb{E}_{y\sim p(\overline{x})}[\Phi_{\text{def}}(h,\overline{x},y)] \\
&= \mathbb{E}_{y\sim p(\overline{x})}[\ell(h,\overline{x},y)] + \sum_{n+j\in\overline{\mathcal{Y}}_A}\mathbb{E}_{y\sim p(\overline{x})}\big[(1-c_j(\overline{x},y))\big]\Phi_{01}(h,\overline{x},n+j) \\
&= \sum_{y\in\mathcal{Y}} s(\overline{x},y)\Phi_{01}(h,\overline{x},y) + \sum_{n+j\in\overline{\mathcal{Y}}_A} s(\overline{x},n+j)\Phi_{01}(h,\overline{x},n+j) \\
&= \sum_{y\in\overline{\mathcal{Y}}_A} s(\overline{x},y)l(h,\overline{x},y),
\end{aligned}$$

which ends the proof.

□

We also make use of the following result for the zero-one loss $\ell_{01}(h, \overline{x}, y) = \mathbf{1}\{h(\overline{x}) \neq y\}$ with label space $\overline{\mathcal{Y}}_A$ and the conditional probability vector $\overline{s}(x, \cdot)$, which characterizes the minimal conditional $\ell_{01}$-risk and the corresponding calibration gap.

### C.3 Calibration gap for zero-one loss

**Lemma C.3.** *(Awasthi et al. (2022b) Lemma 3) For any $x \in \mathcal{X}$, the minimal conditional $\ell_{01}$-risk and the calibration gap for $\ell_{01}$ can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{01}}(x) = 1 - \max_{y \in H(x)} \overline{s}(x, y),$$

$$\Delta\mathcal{C}_{\ell_{01}}(h, x) = \max_{y \in H(x)} \overline{s}(x, y) - \overline{s}(x, h(x)).$$

Now we show that the surrogate upper bound on the calibration gap of the zero-one loss $\ell_{01}$ implies a surrogate upper bound on the calibration gap of the deferral loss $\ell_{\text{def}}$.

**Lemma C.4.** *Given $\Phi_{01}$ is a surrogate loss for the multi-class zero-one classification loss $\ell_{01}$ defined in Definition 4.2, $\Phi_{def}$ is the corresponding surrogate deferral loss. Assume that there exists a concave function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}, \overline{x} \in \overline{\mathcal{X}}$ and $p \in \mathcal{P}$ :*

$$[\Delta\mathcal{C}_{\ell_{01},\mathcal{H}}(h, \overline{x})]_\epsilon \leq \Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, \overline{x})).$$

*Then for any $\overline{\mathbf{x}} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_T) \in \overline{\mathcal{X}}^T$, any $\mathbf{h} = (h_1, h_2, \ldots, h_T) \in \mathcal{H}^T$ used by the randomized learner in Section 4.1, and any distribution $p \in \mathcal{P}$,*

$$[\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(q, \overline{x})]_\epsilon \leq S(\overline{x})\Gamma\left(\frac{1}{S(\overline{x})}\Delta\mathcal{C}_{\Phi_{def},\mathcal{H}}(h, \overline{x})\right) + c\gamma.$$

*Proof.* The calibration gap can be written as follows

$$\begin{aligned}
[\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(q, \overline{x})]_\epsilon &= (1-\gamma)[\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h, \overline{x})]_\epsilon + \frac{\gamma}{|\overline{\mathcal{Y}}_A|}\sum_{y \in \mathcal{Y}_A}[\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(y, \overline{x})]_\epsilon \\
&\leq [\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h, \overline{x})]_\epsilon + c\gamma \\
&= [\max_{y \in H(\overline{x})} s(\overline{x}, y) - s(\overline{x}, h(\overline{x}))]_\epsilon + c\gamma \quad \text{(Lemma C.1)} \\
&= S(\overline{x})[\max_{y \in H(\overline{x})} \overline{s}(\overline{x}, y) - \overline{s}(\overline{x}, h(\overline{x}))]_\epsilon + c\gamma \\
&= S(\overline{x})[\Delta\mathcal{C}_{\ell_{01},\mathcal{H}}(h, \overline{x})]_\epsilon + c\gamma \quad \text{(Lemma C.3)} \\
&\leq S(\overline{x})\Gamma(\Delta\mathcal{C}_{\Phi_{01},\mathcal{H}}(h, \overline{x})) + c\gamma \\
&= S(\overline{x})\Gamma\left(\sum_{y \in \overline{\mathcal{Y}}_A}\overline{s}(\overline{x}, y)\Phi_{01}(h, x, y) - \inf_{h \in \mathcal{H}}\sum_{y \in \overline{\mathcal{Y}}_A}\overline{s}(\overline{x}, y)\Phi_{01}(h, x, y)\right) + c\gamma \\
&= S(\overline{x})\Gamma\left(\frac{1}{S(\overline{x})}\left(\sum_{y \in \overline{\mathcal{Y}}_A} s(\overline{x}, y)\Phi_{01}(h, x, y) - \inf_{h \in \mathcal{H}}\sum_{y \in \overline{\mathcal{Y}}_A} s(\overline{x}, y)\Phi_{01}(h, x, y)\right)\right) + c\gamma \\
&= S(\overline{x})\Gamma\left(\frac{1}{S(\overline{x})}\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h, \overline{x})\right) + c\gamma \quad \text{(Lemma C.2)},
\end{aligned}$$

which completes the proof. □

### C.4 Proof of $\mathcal{H}$-consistency bounds for surrogate deferral losses (Theorem 5.1)

**Theorem 5.1.** *(Distribution-Dependent $\Gamma$-Bound) Assume that $\widetilde{\Phi}_{01}$ admits an $\mathcal{H}$-consistency bound w.r.t $\ell_{01}$ with a concave, non-decreasing calibration function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ with $\Gamma(0) = 0$. Then for any $\overline{\mathbf{x}} \in \overline{\mathcal{X}}^T$ and any distribution $p \in \mathcal{D}$, suppose the online learner generates corresponding hypotheses $\mathbf{h}$ and samples from $\mathbf{q}$ defined in Section 4.1, we have,*

$$\frac{R_{\ell_{\text{def}}}(T) + \mathcal{M}_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) - c\Big(\sum_{t\in[T]} \gamma_t\Big)}{T} - \epsilon \leq \max_{t\in[T]} S(\overline{x}_t)\, \Gamma\left(\frac{R_{\Phi_{\text{def}}}(T) + \mathcal{M}_{\Phi_{\text{def}}\mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t\in[T]} S(\overline{x}_t)}\right).$$

*Here, $S(\overline{x}_t) = \sum_{y\in\mathcal{Y}} p(\overline{x}_t, y) + \sum_{n+j\in\overline{\mathcal{Y}}_{A_t}} \Big(1 - \sum_{y\in\mathcal{Y}} p(\overline{x}_t, y) c_j(x_t, y)\Big)$ and $c$ is the upper bound on the expert cost $c = \max_j\{\overline{c}_j\}$.*

*Proof.* By definition, the LHS term can be written as follows:

$$\begin{aligned}
&\frac{1}{T}\big(\mathcal{R}_{\ell_{\text{def}}}(\mathbf{s}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) + \mathcal{M}_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})\big)\\
&= \frac{1}{T}\left(\sum_{t\in[T]} \Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(s_t, \overline{x}_t)\right)\\
&\leq \frac{1}{T}\left(\sum_{t\in[T]} [\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(s_t, \overline{x}_t)]_\epsilon\right) + \epsilon\\
&\leq \frac{1}{T}\sum_{t\in[T]} S(\overline{x}_t)\Gamma\left(\frac{1}{S(\overline{x}_t)}\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t)\right) + \frac{c(\sum_{t\in[T]}\gamma_T)}{T} + \epsilon \quad \text{(Lemma C.4)}\\
&\leq \max_{t\in[T]}\{S(\overline{x}_t)\}\frac{1}{T}\sum_{t\in[T]} \Gamma\left(\frac{1}{S(\overline{x}_t)}\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t)\right) + \frac{c(\sum_{t\in[T]}\gamma_T)}{T} + \epsilon\\
&\leq \max_{t\in[T]}\{S(\overline{x}_t)\}\Gamma\left(\sum_{t\in[T]} \frac{1}{TS(\overline{x}_t)}\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t)\right) + \frac{c(\sum_{t\in[T]}\gamma_T)}{T} + \epsilon\\
&= \max_{t\in[T]}\{S(\overline{x}_t)\}\Gamma\left(\frac{1}{T\min_{t\in[T]}\{S(\overline{x}_t)\}}\big(\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) + \mathcal{M}_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})\big)\right) + \frac{c(\sum_{t\in[T]}\gamma_T)}{T} + \epsilon.
\end{aligned}$$

when $\Gamma$ is linear we have

$$\begin{aligned}
&\frac{1}{T}\big(\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) + \mathcal{M}_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})\big)\\
&\leq \Gamma\left(\frac{1}{T}\big(\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) + \mathcal{M}_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})\big)\right) + \frac{c\Big(\sum_{t\in[T]}\gamma_t\Big)}{T} + \epsilon,
\end{aligned}$$

which completes the proof.

□

## D Examples of $\mathcal{H}$-consistency bound for common surrogate losses

### D.1 $\Phi_{01}$ being adopted as comp-sum losses

**Example:** $\Phi_{01} = \Phi_{\exp}$. Plug in $\Phi_{01} = \Phi_{\exp} = \sum_{y'\neq y} e^{h(\overline{x},y') - h(\overline{x},y)}$ in Definition 4.3, we obtain

$$\Phi_{\text{def}} = \sum_{y'\neq y} e^{h(\overline{x},y') - h(\overline{x},y)} + \sum_{n+j\in\overline{\mathcal{Y}}_A} \big(1 - c_j(x, y)\big) \sum_{y'\neq n+j} e^{h(\overline{x},y') - h(\overline{x},n+j)}.$$

By Mao et al. (2023b), Theorem 1, $\Phi_{\exp}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{2t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t\in[T]}\gamma_t)}{T} - \epsilon \leq \sqrt{2}\max_{t\in[T]} S(\overline{x}_t)\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T\min_{t\in[T]} S(\overline{x}_t)}\right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\log}$. Plug in $\Phi_{01} = \Phi_{\log} = -\log\Big(\frac{e^{h(\overline{x},y)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}\Big)$ in Definition 4.3, we obtain

$$\Phi_{\text{def}} = -\log\Big(\frac{e^{h(\overline{x},y)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}\Big) - \sum_{n+j\in\overline{\mathcal{Y}}_A}\big(1 - c_j(x,y)\big)\log\Big(\frac{e^{h(\overline{x},n+j)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}\Big).$$

By Mao et al. (2023b), Theorem 1, $\Phi_{\log}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{2t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t\in[T]}\gamma_t)}{T} - \epsilon \leq \sqrt{2}\max_{t\in[T]} S(\overline{x}_t)\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T\min_{t\in[T]} S(\overline{x}_t)}\right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\text{gce}}$. Plug in $\Phi_{01} = \Phi_{\text{gce}} = \frac{1}{\alpha}\big[1 - (\frac{e^{h(\overline{x},y)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}})^\alpha\big]$ in Definition 4.3, we obtain

$$\Phi_{\text{def}} = \frac{1}{\alpha}\Big[1 - \Big(\frac{e^{h(\overline{x},y)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}\Big)^\alpha\Big] + \frac{1}{\alpha}\sum_{n+j\in\overline{\mathcal{Y}}_A}(1 - c_j(x,y))\Big[1 - \Big(\frac{e^{h(\overline{x},n+j)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}\Big)^\alpha\Big].$$

By Mao et al. (2023b), Theorem 1, $\Phi_{\text{gce}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{2n^\alpha t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t\in[T]}\gamma_t)}{T} - \epsilon \leq \sqrt{2n^\alpha}\max_{t\in[T]} S(\overline{x}_t)\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T\min_{t\in[T]} S(\overline{x}_t)}\right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\text{mae}}$. Plug in $\Phi_{01} = \Phi_{\text{mae}} = 1 - \frac{e^{h(\overline{x},y)}}{\sum_{y'\in\overline{\mathcal{Y}}_A} e^{h(\overline{x},y')}}$ in Definition 4.3, we obtain

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

By Mao et al. (2023b), Theorem 1, $\Phi_{\text{mae}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = nt$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t\in[T]}\gamma_t)}{T} - \epsilon \leq n\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T}\right).$$

### D.2 $\Phi_{01}$ being adopted as sum losses

**Example:** $\Phi_{01} = \Phi^{\text{sum}}_{\text{sq}}$. Plug in $\Phi_{01} = \Phi^{\text{sum}}_{\text{sq}} = \sum_{y'\neq y}\Phi_{\text{sq}}\big(h(\overline{x},y) - h(\overline{x},y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\text{def}} = \sum_{y'\neq y}\Phi_{\text{sq}}\big(h(\overline{x},y) - h(\overline{x},y')\big) + \sum_{n+j\in\overline{\mathcal{Y}}_A}(1 - c_j(x,y))\sum_{y'\neq n+j}\Phi_{\text{sq}}\big(h(\overline{x},n+j) - h(\overline{x},y')\big),$$

where $\Phi_{\text{sq}}(t) = \max\{0, 1-t\}^2$.

By Awasthi et al. (2022b), Table 2, $\Phi^{\text{sum}}_{\text{sq}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t\in[T]}\gamma_t)}{T} - \epsilon \leq \max_{t\in[T]} S(\overline{x}_t)\left(\frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}})}{T\min_{t\in[T]} S(\overline{x}_t)}\right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\exp}^{\mathrm{sum}}$. Plug in $\Phi_{01} = \Phi_{\exp}^{\mathrm{sum}} = \sum_{y' \neq y} \Phi_{\exp}\big(h(\overline{x}, y) - h(\overline{x}, y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\mathrm{def}} = \sum_{y' \neq y} \Phi_{\exp}\big(h(\overline{x}, y) - h(\overline{x}, y')\big) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \sum_{y' \neq n+j} \Phi_{\exp}\big(h(\overline{x}, n+j) - h(\overline{x}, y')\big),$$

where $\Phi_{\exp}(t) = e^{-t}$.

By Awasthi et al. (2022b), Table 2, $\Phi_{\exp}^{\mathrm{sum}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{2t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\mathrm{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t \in [T]} \gamma_t)}{T} - \epsilon \leq \sqrt{2} \max_{t \in [T]} S(\overline{x}_t) \left( \frac{\mathcal{R}_{\Phi_{\mathrm{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t \in [T]} S(\overline{x}_t)} \right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\rho}^{\mathrm{sum}}$. Plug in $\Phi_{01} = \Phi_{\rho}^{\mathrm{sum}} = \sum_{y' \neq y} \Phi_{\rho}\big(h(\overline{x}, y) - h(\overline{x}, y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\mathrm{def}} = \sum_{y' \neq y} \Phi_{\rho}\big(h(\overline{x}, y) - h(\overline{x}, y')\big) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \sum_{y' \neq n+j} \Phi_{\rho}\big(h(\overline{x}, n+j) - h(\overline{x}, y')\big),$$

where $\Phi_{\rho}(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$.

By Awasthi et al. (2022b), Table 2, $\Phi_{\rho}^{\mathrm{sum}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = t$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\mathrm{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t \in [T]} \gamma_t)}{T} - \epsilon \leq \frac{\mathcal{R}_{\Phi_{\mathrm{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T}.$$

### D.3 $\Phi_{01}$ being adopted as constrained losses

**Example:** $\Phi_{01} = \Phi_{\mathrm{hinge}}^{\mathrm{cstnd}}$. Plug in $\Phi_{01} = \Phi_{\mathrm{hinge}}^{\mathrm{cstnd}} = \sum_{y' \neq y} \Phi_{\mathrm{hinge}}\big(-h(\overline{x}, y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\mathrm{def}} = \sum_{y' \neq y} \Phi_{\mathrm{hinge}}\big(-h(\overline{x}, y')\big) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \sum_{y' \neq n+j} \Phi_{\mathrm{hinge}}\big(-h(\overline{x}, y')\big),$$

where $\Phi_{\mathrm{hinge}}(t) = \max\{0, 1 - t\}$ with the constraint $\sum_{y \in \overline{\mathcal{Y}}_A} h(\overline{x}, y) = 0$.

By Awasthi et al. (2022b), Table 3, $\Phi_{\mathrm{hinge}}^{\mathrm{cstnd}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = t$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\mathrm{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t \in [T]} \gamma_t)}{T} - \epsilon \leq \frac{\mathcal{R}_{\Phi_{\mathrm{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T}.$$

**Example:** $\Phi_{01} = \Phi_{\mathrm{sq}}^{\mathrm{cstnd}}$. Plug in $\Phi_{01} = \Phi_{\mathrm{sq}}^{\mathrm{cstnd}} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}\big(-h(\overline{x}, y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\mathrm{def}} = \sum_{y' \neq y} \Phi_{\mathrm{sq}}\big(-h(\overline{x}, y')\big) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \sum_{y' \neq n+j} \Phi_{\mathrm{sq}}\big(-h(\overline{x}, y')\big).$$

By Awasthi et al. (2022b), Table 3, $\Phi_{\mathrm{sq}}^{\mathrm{cstnd}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = \sqrt{t}$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\mathrm{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t \in [T]} \gamma_t)}{T} - \epsilon \leq \max_{t \in [T]} S(\overline{x}_t) \left( \frac{\mathcal{R}_{\Phi_{\mathrm{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\mathrm{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t \in [T]} S(\overline{x}_t)} \right)^{1/2}.$$

**Example:** $\Phi_{01} = \Phi_{\rho}^{\mathrm{cstnd}}$. Plug in $\Phi_{01} = \Phi_{\rho}^{\mathrm{cstnd}} = \sum_{y' \neq y} \Phi_{\rho}\big(-h(\overline{x}, y')\big)$ in Definition 4.3, we obtain

$$\Phi_{\mathrm{def}} = \sum_{y' \neq y} \Phi_{\rho}\big(-h(\overline{x}, y')\big) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \sum_{y' \neq n+j} \Phi_{\rho}\big(-h(\overline{x}, y')\big),$$

where $\Phi_\rho(t) = \min\{\max\{0, 1 - t/\rho\}, 1\}$ with the constraint $\sum_{y \in \overline{\mathcal{Y}}_A} h(\overline{x}, y) = 0$.

By Awasthi et al. (2022b), Table 3, $\Phi_\rho^{\text{cstnd}}$ admits an $\mathcal{H}$-consistency bound with respect to $\ell_{01}$ with $\Gamma(t) = t$; using Corollary 5.2, we obtain

$$\frac{\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T} - \frac{c(\sum_{t \in [T]} \gamma_t)}{T} - \epsilon \leq \frac{\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T}.$$

# E Proof of constrained OCO with $\Phi_{\text{hinge}}^{\text{cstnd}}$

## E.1 Proof of constrained OCO with $\Phi_{\text{hinge}}^{\text{cstnd}}$ (Theorem 6.3)

We first give a proof for Lemma 6.2

**Lemma 6.2.** *Let $\widehat{\nabla}_{t,i}$ be the gradient of the $i$-th row of $\widehat{\nabla}_t = \nabla \widehat{\Phi}_t\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}\big)$ in round $t \in [T]$, and $\widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}$ the projected gradient onto $\mathcal{K}_{A_t}$, that is $\widehat{\nabla}_t|_{\mathcal{K}_{A_t}} = (\widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}})_{i=1}^N$. We have*

$$\widehat{\Phi}_t\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}\big) - \widehat{\Phi}_t\big(\widetilde{W}^*|_{\mathcal{K}_{A_t}}\big) \leq \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle.$$

*Proof.* By convexity we have

$$\begin{aligned}
\widehat{\Phi}_t\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}\big) - \widehat{\Phi}_t\big(\widetilde{W}^*|_{\mathcal{K}_{A_t}}\big) &\leq \langle \widehat{\nabla}_t, \widetilde{W}_t|_{\mathcal{K}_{A_t}} - \widetilde{W}^*|_{\mathcal{K}_{A_t}} \rangle_F \\
&= \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}, (\widetilde{w}_{t,i} - \mu_t) - (\widetilde{w}_i^* - \mu^*_{A_t}) \rangle \\
&= \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle + \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}, \mu^*_{A_t} - \mu_t \rangle \\
&= \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle + \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \frac{\sum_{i \in \overline{\mathcal{Y}}_{A_t}} \widehat{\nabla}_{t,i}}{|\overline{\mathcal{Y}}_{A_t}|}, \widetilde{w}_i^* - \widetilde{w}_{t,i} \rangle \\
&= \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i} - \frac{\sum_{i \in \overline{\mathcal{Y}}_{A_t}} \widehat{\nabla}_{t,i}}{|\overline{\mathcal{Y}}_{A_t}|}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle \\
&= \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \langle \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle,
\end{aligned} \tag{3}$$

where $\langle \rangle_F$ denotes the Frobenius inner product, $\mu_t = \frac{1}{|\overline{\mathcal{Y}}_{A_t}|} \sum_{i \in \overline{\mathcal{Y}}_{A_t}} \widetilde{w}_{t,i}$, $\mu^*_A = \frac{1}{|\overline{\mathcal{Y}}_A|} \sum_{i \in \overline{\mathcal{Y}}_A} \widetilde{w}_i^*$. The first equality is due to $\nabla_{t,i} = \mathbf{0}$, $\forall i \notin \overline{\mathcal{Y}}_{A_t}$, the third and fifth equalities are due to Proposition 6.1. □

To prove Theorem 6.3, we first show a bound on the sum of inner products $\langle \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle$ for each label $i$ over rounds where $i$ is available.

**Lemma E.1.** *Suppose that the row vector $\widetilde{w}$ is taken from a non-empty closed convex set with diameter $B$ and assume that $\eta_{t+1} \leq \eta_t, t = 1, \ldots, T$. For any label $y = i$, let $S_i$ be the set of rounds where $i$ is available, that is $S_i = \{t \in [T] : i \in \overline{\mathcal{Y}}_{A_t}\}$, we have*

$$2 \sum_{t \in S_i} \langle \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^* \rangle \leq \frac{B^2}{\eta_{t_k}} + \sum_{t \in S_i} \eta_t \big|\big| \widehat{\nabla}_{t,i}|_{\mathcal{K}_{A_t}} \big|\big|_2^2,$$

*where $t_k$ is the last round that $i$ is available.*

*Proof.* Assume the label is available in rounds $t_1, t_2, \ldots, t_k$ with $(t_1 < t_2 < \cdots < t_k)$, we have

$$\begin{aligned}
||\widetilde{w}_{t_{j+1},i} - \widetilde{w}_i^*||_2^2 &= ||\Pi(\widetilde{w}_{t_j,i} - \eta_{t_j}\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}) - \widetilde{w}_i^*||_2^2 \\
&\leq ||\widetilde{w}_{t_j,i} - \eta_{t_j}\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}} - \widetilde{w}_i^*||_2^2 \\
&= ||\widetilde{w}_{t_j,i} - \widetilde{w}_i^*||_2^2 + \eta_{t_j}^2 \left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2 - 2\eta_{t_j}\langle \widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}, \widetilde{w}_{t_j,i} - \widetilde{w}_i^*\rangle.
\end{aligned}$$

Hence

$$2\langle \widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}, \widetilde{w}_{t_j,i} - \widetilde{w}_i^*\rangle \leq \frac{||\widetilde{w}_{t_j,i} - \widetilde{w}_i^*||_2^2 - ||\widetilde{w}_{t_{j+1},i} - \widetilde{w}_i^*||_2^2}{\eta_{t_j}} + \eta_{t_j}\left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2.$$

Summing over $j$ we have

$$\begin{aligned}
\sum_{j\in[k]} 2\langle \widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}, \widetilde{w}_{t_j,i} - \widetilde{w}_i^*\rangle &\leq \frac{1}{\eta_{t_1}}||\widetilde{w}_{t_1,i} - \widetilde{w}_i^*||_2^2 - \frac{1}{\eta_{t_k}}||\widetilde{w}_{t_k,i} - \widetilde{w}_i^*||_2^2 \\
&\quad + \sum_{j\in[k-1]}\left(\frac{1}{\eta_{t_{j+1}}} - \frac{1}{\eta_{t_j}}\right)||\widetilde{w}_{t_{j+1},i} - \widetilde{w}_i^*||_2^2 + \sum_{j\in[k]}\eta_{t_j}\left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2 \\
&\leq \frac{1}{\eta_{t_1}}B^2 + \sum_{j\in[k-1]}\left(\frac{1}{\eta_{t_{j+1}}} - \frac{1}{\eta_{t_j}}\right)B^2 + \sum_{j\in[k]}\eta_{t_j}\left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2 \\
&= \frac{1}{\eta_{t_1}}B^2 + \left(\frac{1}{\eta_{t_k}} - \frac{1}{\eta_{t_1}}\right)B^2 + \sum_{j\in[k]}\eta_{t_j}\left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2 \\
&= \frac{B^2}{\eta_{t_k}} + \sum_{j\in[k]}\eta_{t_j}\left|\left|\widehat{\nabla}_{t_j,i}|_{\mathcal{K}_{A_{t_j}}}\right|\right|_2^2,
\end{aligned}$$

where the second inequality is due to $B$ being the diameter and $\eta_{t_j} \geq \eta_{t_{j+1}}$. This ends our proof. □

Next, we show that the projected gradient norm can be bounded.

**Lemma E.2.** *For any $W \in \mathcal{H}$ and any $((x, A), y) \in \overline{\mathcal{X}} \times \mathcal{Y}$, the following upper bound holds*

$$\left|\left|\nabla\Phi_{01}(W, \overline{x}, y)\right|\right|_F \leq \sqrt{N}R.$$

*Proof.* Due to the special properties of the projection, we can compute the exact form of $\left|\left|\nabla\Phi_{01}(W, \overline{x}, y)\right|\right|_F$. A subgradient of $\Phi_{01}(W, \overline{x}, y)$ w.r.t rows of $W$ is

$$\nabla_{w_i}\Phi_{01}(W, \overline{x}, y) = \alpha_i x, \tag{4}$$

where $\alpha_i = \mathbf{1}\{1 + \langle w_y', x\rangle + b_y > 0\}$ if $i \in \overline{\mathcal{Y}}_A \setminus \{y\}$ and $\alpha_i = 0$ otherwise. Let $\alpha = (\alpha_i)^\intercal \in \{0, 1\}^{|\mathcal{Y}_A|}$, then

$$\nabla\Phi_{01}(W, \overline{x}, y) = \alpha x^\intercal. \tag{5}$$

By Proposition 6.1, for any $1 \leq i \leq N$,

$$(\nabla\Phi_{01}(W, \overline{x}, y)|_{\mathcal{K}_A})_i = \left(\alpha_i - \frac{s}{|\overline{\mathcal{Y}}_A|}\right)x^\intercal, \tag{6}$$

where $s = \sum_{i\in[N]}\alpha_i$. Therefore,

$$\left|\left|\nabla\Phi_{01}(W, \overline{x}, y)\right|\right|_F^2 = \sum_{i\in[N]}\left(\alpha_i - \frac{s}{|\overline{\mathcal{Y}}_A|}\right)^2||x||_2^2 = \left(s - \frac{s^2}{|\overline{\mathcal{Y}}_A|}\right)||x||_2^2. \tag{7}$$

This implies

$$\left|\left|\nabla\Phi_{01}(W, \overline{x}, y)\right|\right|_F^2 = \sqrt{s\left(1 - \frac{s}{|\overline{\mathcal{Y}}_A|}\right)}||x||_2 \leq \sqrt{N}R, \tag{8}$$

which completes our proof. □

**Theorem 6.3.** *(Surrogate regret bound) Running Algorithm 1 with* $\eta_t = \frac{B}{\sqrt{N}Rt^{2/3}}$, $\gamma_t = \min\left\{\frac{1}{2}, \frac{1}{t^{1/3}}\right\}$ *for* $t = 1, \ldots, T$, *we have the following surrogate regret bound*

$$\sum_{t\in[T]} \left(\Phi_{def}\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t\big) - \Phi_{def}\big(\widetilde{W}^*|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t\big)\right) = O\big(N^{3/2}B\,R\,T^{2/3}\big).$$

*Proof.* For each label $i \in [N]$, let $S_i \subseteq [T]$ be the set of rounds in which $i$ is available, and let $\eta_{t_i} = \max S_i$ be the last available round of $i$. Combining Lemma 6.2 and Lemma E.1 we have

$$\begin{aligned}
\sum_{t\in[T]} 2\Big(\widehat{\Phi}_t\big(\widetilde{W}_t\big|_{\mathcal{K}_{A_t}}\big) - \widehat{\Phi}_t\big(\widetilde{W}^*\big|_{\mathcal{K}_{A_t}}\big)\Big) &\le 2\sum_{t\in[T]}\sum_{i\in\overline{\mathcal{Y}}_{A_t}} \Big\langle \widehat{\nabla}_{t,i}\big|_{\mathcal{K}_{A_t}}, \widetilde{w}_{t,i} - \widetilde{w}_i^*\Big\rangle \\
&\le \sum_{i\in[N]}\left(\frac{B^2}{\eta_{t_i}} + \sum_{t\in S_i}\eta_t\big\|\widehat{\nabla}_{t,i}\big|_{\mathcal{K}_{A_t}}\big\|_2^2\right) \\
&\le \frac{NB^2}{\eta_T} + \sum_{t\in[T]}\eta_t\big\|\widehat{\nabla}_t\big|_{\mathcal{K}_{A_t}}\big\|_F^2,
\end{aligned} \tag{9}$$

where the last inequality is due to $\eta_{t_i} \le \eta_T$, $\forall i \in [N]$.

Note that $\mathbb{E}_{y_t'\sim q_t}\big[\widehat{\Phi}_t(\widetilde{W})\big] = \Phi_{\text{def}}(\widetilde{W}, \overline{x}_t, y_t)$, $\forall \widetilde{W} \in \mathcal{H}$. Next, we bound the expected gradient's norm

$$\begin{aligned}
&\mathbb{E}_{y_t'\sim q_t}\big[\big|\big|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big|\big|_F^2\big] \\
&= \frac{1}{q_t(y_t)}\big|\big|\nabla\Phi_{01}(\widetilde{W}, \overline{x}_t, y_t)|_{\mathcal{K}_{A_t}}\big|\big|_F^2 + \sum_{n+j\in\overline{\mathcal{Y}}_{A_t}} \frac{(1-c_j(x_t,y_t))^2}{q_t(n+j)}\big|\big|\nabla\Phi_{01}(\widetilde{W}, \overline{x}_t, n+j)|_{\mathcal{K}_{A_t}}\big|\big|_F^2 \\
&\le \frac{N}{\gamma_t}NR^2 \le \max\{2N^2R^2, N^2R^2\sqrt{t}\},
\end{aligned}$$

whre the first inequality is due to Lemma E.2. Therefore, taking expectation on both sides in Inequality 9 gives

$$\begin{aligned}
&\sum_{t\in[T]} \left(\Phi_{\text{def}}\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t\big) - \Phi_{\text{def}}\big(\widetilde{W}^*|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t\big)\right) \\
&\qquad \le N^{3/2}BRT^{2/3} + \sum_{t\in[T]} \max\left\{\frac{2N^{3/2}BR}{t^{2/3}}, \frac{N^{3/2}BR}{t^{1/3}}\right\} = O(N^{3/2}BRT^{2/3}).
\end{aligned}$$

which ends our proof. □

### E.2 Proof of regret bound under near-realizable assumption (Theorem 6.5)

Consider an input $\overline{x}_t = (x_t, A_t)$. The difficulty of achieving a strong regret bound comes from the fact that we only observe the loss function partially in most situations, thus may lead to large variance in the gradient updates (as seen in the analysis of Theorem 6.3). One exception is when $|\mathcal{Y}_{A_t}| = 2$. This occurs when $n = 2$ and no experts are available in round $t$. In such rounds, bandit feedback is equivalent to full feedback. In other words, the learner always observes the full loss function, and the gradient updates are computed exactly. To address the case when $|\mathcal{Y}_{A_t}| \ge 3$, we give a tighter bound on the gap $\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h, \overline{x}) - \Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h, \overline{x})$. The analysis is inspired by van der Hoeven (2020). Given an input $\overline{x}$, $y_{\max}$ is the label with the highest score, the multi-class label margin given is defined as $\overline{s}(\overline{x}, y_{\max}) - \arg\max_{y\neq y_{\max}} \overline{s}(\overline{x}, y)$. Note that the near-realizable condition directly gives us a lower bound on the multi-class margin: for any $t \in [T]$ $\overline{s}(\overline{x}_t, y_{\max}) - \arg\max_{y\neq y_{\max}} \overline{s}(\overline{x}_t, y) \ge 1 - 2/\sqrt{T}$, which we will use extensively for the subsequent analysis.

**Lemma E.3.** *Under the near-realizable assumption, given any hypothesis $h \in \mathcal{H}$ and input $\overline{x} = (x, A) \in \overline{\mathcal{X}}$. If $|\mathcal{Y}_A| \geq 3$ and $h(\overline{x}) \neq y_{\max}$, we have the following bound*

$$\Delta\mathcal{C}_{\Phi_{def},\mathcal{H}}(h, \overline{x}) - \Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h, \overline{x}) \geq \frac{1 - 2/\sqrt{T}}{|\overline{\mathcal{Y}}_A| + 4}\mathcal{C}_{\Phi_{def}}(h, \overline{x}).$$

*Proof.* Let $c = 1 - 2/\sqrt{T}$. Similar to Theorem E.5, define $\mathcal{Y}_{\leq -1} = \{y \in \overline{\mathcal{Y}}_A : h(\overline{x}, y) \leq -1\}$, $\mathcal{Y}_{>1} = \overline{\mathcal{Y}}_A - \mathcal{Y}_{\leq -1}$, and $y_{\max} = \arg\max_{y \in \overline{\mathcal{Y}}_A} s(\overline{x}, y)$, we have $S(\overline{x}) - s(\overline{x}, y) \geq S(\overline{x}) - s(\overline{x}, y_{\max})$ and

$$\sum_{y \in \mathcal{Y}_{>-1}} \big(1 + h(\overline{x}, y)\big) = |\mathcal{Y}_{>-1}| - \big(\sum_{y \in \mathcal{Y}_{\leq -1}} h(\overline{x}, y)\big) \geq |\mathcal{Y}_{>-1}| + |\mathcal{Y}_{\leq -1}| = |\overline{\mathcal{Y}}_A|.$$

By Lemma C.2 and Theorem E.5, we can write

$$\begin{aligned}
\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) &= \sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(S(\overline{x}) - s(x, y)) \\
&= \sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(S(\overline{x}) - s(x, y_{\max})) + \sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(s(x, y_{\max}) - s(x, y)) \\
&\geq |\overline{\mathcal{Y}}_A|(S(\overline{x} - s(x, y_{\max})) + \sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(s(x, y_{\max}) - s(x, y)) \\
&\geq \mathcal{C}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{x}) + \sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(s(x, y_{\max}) - s(x, y)).
\end{aligned} \tag{10}$$

We consider two cases.

**Case 1.** $h(x, y_{\max}) \notin \mathcal{Y}_{>-1}$, we have

$$\begin{aligned}
&\sum_{y \in \mathcal{Y}_{>-1}} (1 + h(\overline{x}, y))(s(x, y_{\max}) - s(x, y)) \\
&= (s(x, y_{\max}) - s(x, h(\overline{x}))) + h(\overline{x}, h(\overline{x}))(s(x, y_{\max}) - s(x, h(\overline{x}))) + \sum_{y \in \mathcal{Y}_{>-1} \setminus \{h(\overline{x})\}} (1 + h(\overline{x}, y))(s(x, y_{\max}) - s(x, y)) \\
&\geq (s(x, y_{\max}) - s(x, h(\overline{x}))) + h(\overline{x}, h(\overline{x}))c + \sum_{y \in \mathcal{Y}_{>-1} \setminus \{h(\overline{x})} (1 + h(\overline{x}, y))c \\
&\geq (s(x, y_{\max}) - s(x, h(\overline{x}))) + (|\overline{\mathcal{Y}}_A| - 1)c.
\end{aligned} \tag{11}$$

Combining equations (10) and (11) we have

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h, \overline{x}) - \Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h, \overline{x}) - \frac{c}{|\overline{\mathcal{Y}}_A| + 4}\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) \\
&= \left(1 - \frac{c}{|\overline{\mathcal{Y}}_A| + 4}\right)\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) - \mathcal{C}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{x}) - \Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h, \overline{x}) \\
&\geq \left(1 - \frac{c}{|\overline{\mathcal{Y}}_A| + 4}\right)(|\overline{\mathcal{Y}}_A| - 1)c - \frac{c}{|\overline{\mathcal{Y}}_A| + 4}|\overline{\mathcal{Y}}_A|(S(\overline{x} - s(x, y_{\max})) - \frac{c}{|\overline{\mathcal{Y}}_A| + 4}(s(x, y_{\max}) - s(x, h(\overline{x}))) \\
&\geq 0,
\end{aligned} \tag{12}$$

where the first inequality is due to Theorem E.5 and the last inequality is due to $S(\overline{x}) - s(x, y_{\max}) \leq |\overline{\mathcal{Y}}_A| - 1$ and $|\overline{\mathcal{Y}}_A| \geq 3$.

**Case 2.** $h(x, y_{\max}) \in \mathcal{Y}_{>-1}$. We have

$$
\begin{aligned}
&\sum_{y\in\mathcal{Y}_{>-1}}(1+h(\overline{x},y))(s(x,y_{\max})-s(x,y))\\
&=(s(x,y_{\max})-s(x,h(\overline{x})))+h(\overline{x},h(\overline{x}))(s(x,y_{\max})-s(x,h(\overline{x})))\\
&+\sum_{y\in\mathcal{Y}_{>-1}\setminus\{h(\overline{x}),y_{\max}\}}(1+h(\overline{x},y))(s(x,y_{\max})-s(x,y))\\
&\geq(s(x,y_{\max})-s(x,h(\overline{x})))+h(\overline{x},h(\overline{x}))c+\sum_{y\in\mathcal{Y}_{>-1}\setminus\{h(\overline{x}),y_{\max}\}}(1+h(\overline{x},y))c\\
&\geq(s(x,y_{\max})-s(x,h(\overline{x})))+(\sum_{y\in\mathcal{Y}_{>-1}}(1+h(x,y))-h(x,y_{\max})-2)c\\
&\geq(s(x,y_{\max})-s(x,h(\overline{x})))+(\frac{1}{2}\sum_{y\in\mathcal{Y}_{>-1}}(1+h(x,y))-1)c\\
&\geq(s(x,y_{\max})-s(x,h(\overline{x})))+(\frac{|\overline{\mathcal{Y}}_A|}{2}-1)c,
\end{aligned}
\tag{13}
$$

where the second to last inequality is due to $1+h(x,y_{\max})\leq 1+h(x,h(x))$, so $1+h(x,y_{\max})\leq\frac{1}{2}\sum_{y\in\mathcal{Y}_{>-1}}(1+h(x,y))$. Combining equations (10) and (13) we have

$$
\begin{aligned}
&\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h,\overline{x})-\Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h,\overline{x})-\frac{c}{|\overline{\mathcal{Y}}_A|+4}\mathcal{C}_{\Phi_{\text{def}}}(h,\overline{x})\\
&=\left(1-\frac{c}{|\overline{\mathcal{Y}}_A|+4}\right)\mathcal{C}_{\Phi_{\text{def}}}(h,\overline{x})-\mathcal{C}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{x})-\Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h,\overline{x})\\
&\geq\left(1-\frac{c}{|\overline{\mathcal{Y}}_A|+4}\right)(\frac{|\overline{\mathcal{Y}}_A|}{2}-1)c-\frac{c}{|\overline{\mathcal{Y}}_A|+4}|\overline{\mathcal{Y}}_A|(S(\overline{x}-s(x,y_{\max}))-\frac{c}{|\overline{\mathcal{Y}}_A|+4}(s(x,y_{\max})-s(x,h(\overline{x})))\\
&\geq 0,
\end{aligned}
\tag{14}
$$

where the first inequality is due to Theorem E.5 and the last inequality is due to $S(\overline{x})-s(x,y_{\max})\leq|\overline{\mathcal{Y}}_A|-1$ and $|\overline{\mathcal{Y}}_A|\geq 3$. Combining the two cases ends our proof.

□

Define $\delta_t=\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h_t,\overline{x}_t)-\Delta\mathcal{C}_{\ell_{\text{def}}\ \mathcal{H}}(h_t,\overline{x}_t)$. Recall that the estimated loss at round $t$ is defined as

$$
\begin{aligned}
\widehat{\Phi}_t(\widetilde{W})=&\sum_{n+j\in\overline{\mathcal{Y}}_{A_t}}v_{t,j}\widetilde{\Phi}_{01}(\widetilde{W},\overline{x}_t,n+j)(1-c_j(x_t,y_t))\\
&+v_{t,0}\widetilde{\Phi}_{01}(\widetilde{W},\overline{x}_t,y_t'),\forall\widetilde{W}\in\mathcal{H}.
\end{aligned}
$$

The following lemma shows that $||\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}||^2$ is upper bounded by $\widehat{\Phi}_{\text{def}}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,y_t)$.

**Lemma E.4.** *Define $v_{\max}=\max_{t,j}v_{t,j}$, at any round $t\in[T]$, we have the following inequality*

$$
||\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}||^2\leq 2v_{\max}\sqrt{N}R\,\widehat{\Phi}_{def}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,y_t).
\tag{15}
$$

*Proof.* We can write the gradient's norm as follows

$$
\begin{aligned}
||\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}||^2&=v_{t,0}^2||\nabla\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,y_t)|_{\mathcal{K}_{A_t}}||^2+\sum_{n+j\in\overline{\mathcal{Y}}_{A_t}}v_{t,j}^2||\nabla\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,n+j)|_{\mathcal{K}_{A_t}}||^2\\
&\leq v_{\max}(v_{t,0}||\nabla\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,y_t)|_{\mathcal{K}_{A_t}}||^2+\sum_{n+j\in\overline{\mathcal{Y}}_{A_t}}v_{t,j}||\nabla\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}},\overline{x}_t,n+j)|_{\mathcal{K}_{A_t}}||^2).
\end{aligned}
\tag{16}
$$

We consider two cases.

**Case 1.** $h(\overline{x}_t) \notin \{y_t\} \cup \{n+j : n+j \in \overline{\mathcal{Y}}_{A_t}\}$. For $y \neq h(\overline{x}_t)$, we have

$$\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t) \geq 1 + h(x, h(x)) \geq 1,$$

therefore by Lemma E.2,

$$||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y)|_{\mathcal{K}_{A_t}}||^2 \leq \sqrt{N}R \leq \sqrt{N}R\, \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y).$$

This gives us

$$v_{t,0}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)|_{\mathcal{K}_{A_t}}||^2 + \sum_{n+j \in \overline{\mathcal{Y}}_{A_t}} v_{t,j}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, n+j)|_{\mathcal{K}_{A_t}}||^2 \leq \sqrt{N}R\, \widehat{\Phi}_{\text{def}}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t). \tag{17}$$

**Case 2.** $h(\overline{x}_t) \notin \{y_t\} \cup \{n+j : n+j \in \overline{\mathcal{Y}}_{A_t}\}$. Suppose $h(\overline{x}_t)$ corresponds to indicator $v_{t,j}$. Since $q_t(h(\overline{x}_t)) \geq 1 - \gamma_t \geq 1/2$, we have $v_{t,j} \leq 2$. Therefore

$$v_{t,j}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, h(\overline{x}_t))|_{\mathcal{K}_{A_t}}||^2 \leq 2\sqrt{N}R \leq \sqrt{N}R(\sum_{i \neq j} v_{t,j}\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)).$$

Combining with Equation (15) gives us

$$v_{t,0}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)|_{\mathcal{K}_{A_t}}||^2 + \sum_{n+j \in \overline{\mathcal{Y}}_{A_t}} v_{t,j}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, n+j)|_{\mathcal{K}_{A_t}}||^2 \leq 2\sqrt{N}R\, \widehat{\Phi}_{\text{def}}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t). \tag{18}$$

Now from Equation (16) we can write

$$\begin{aligned}
||\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}||^2 &\leq v_{\max}(v_{t,0}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t)|_{\mathcal{K}_{A_t}}||^2 + \sum_{n+j \in \overline{\mathcal{Y}}_{A_t}} v_{t,j}||\nabla \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, n+j)|_{\mathcal{K}_{A_t}}||^2) \\
&\leq 2v_{\max}\sqrt{N}R(v_{t,0}\Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t) + \sum_{n+j \in \overline{\mathcal{Y}}_{A_t}} \Phi_{01}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, n+j)) \\
&\leq 2v_{\max}\sqrt{N}R\, \widehat{\Phi}_{\text{def}}(\widetilde{W}_t|_{\mathcal{K}_{A_t}}, \overline{x}_t, y_t),
\end{aligned} \tag{19}$$

which ends our proof. □

We are now ready to prove Theorem 6.5.

**Theorem 6.5.** *(Near-Realizable Assumption) Under the same assumptions in Corollary 5.2 and Theorem 6.3, suppose that for any $t \in [T]$ we have $s(\overline{x}_t, y_{\max}) \geq 1 - 1/\sqrt{T}$. Running Algorithm 1 with $\eta_t = \frac{B}{\sqrt{4RN^{1/2}(N+4)t}}$ and $\gamma = B\sqrt{RN^{5/2}(N+4)}$ gives us the following regret deferral bound*

$$\begin{aligned}
R_{\ell_{\text{def}}}(T) &= \sum_{t \in [T]} \left(\ell_{def}(q_t, \overline{x}_t, y_t) - \ell_{def}(\widetilde{W}^*, \overline{x}_t, y_t)\right) \\
&= O\left(B\sqrt{R^3 N^{7/2} T}\right)
\end{aligned}$$

*Proof.* In Theorem 5.1, we upper bound $[\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(q_t, \overline{x}_t)]_\epsilon$ by $S(\overline{x}_t)\Gamma\left(\frac{1}{S(\overline{x}_t)}\Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h, \overline{x}_t)\right) + c\gamma_t$. With $\delta_t$ we can rewrite the bound as follow

$$\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(q_t, \overline{x}_t) \leq \Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t) + c\gamma_t = \Delta\mathcal{C}_{\Phi_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t) + c\gamma_t - \delta_t.$$

Moreover, when $h(\overline{x}_t) = y_{\max}$, we have $\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h_t, \overline{x}_t) = 0$ and

$$\Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(q_t,\overline{x}_t) \le \Delta\mathcal{C}_{\ell_{\text{def}},\mathcal{H}}(h_t,\overline{x}_t) + c\gamma_t = c\gamma_t.$$

Following the same arguments in Theorem 5.1 gives us

$$\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q},\overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) \le \mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h},\overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) - \sum_{t\in[T]}\delta_t + c\left(\sum_{t\in[T]}\gamma_t\right). \tag{20}$$

We want to want to show an upper bound on $\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h},\overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) - \sum_{t\in[T]}\delta_t$. In Theorem 6.3, we have shown that

$$\sum_{t\in[T]} 2\Big(\widehat{\Phi}_{\text{def},t}\big(\widetilde{W}_t|_{\mathcal{K}_{A_t}}\big) - \widehat{\Phi}_{\text{def},t}\big(\widetilde{W}^*|_{\mathcal{K}_{A_t}}\big)\Big) \;\le\; \frac{NB^2}{\eta_T} + \sum_{t\in[T]}\eta_t\big\|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big\|^2. \tag{21}$$

Therefore

$$\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h},\overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) - \sum_{t\in[T]}\delta_t = \frac{NB^2}{\eta_T} + \mathbb{E}_{y_t\sim p(\overline{x}_t),y'_t\sim q_t}\left[\sum_{t\in[T]}\left(\eta_t\big|\big|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big|\big|^2 - \delta_t\right)\right], \tag{22}$$

where the expectation is taken over the label distribution and the algorithm's sampling distribution. On rounds $t\in[T]$ in which $|\mathcal{Y}_{A_t}| = 2$, the exact gradient is computed, hence

$$\eta_t\big|\big|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big|\big|^2 = \eta_t\big|\big|\nabla_t|_{\mathcal{K}_{A_t}}\big|\big|^2 \le 4\eta_t NR^2. \tag{23}$$

On rounds $t\in[T]$ in which $h(\overline{x}_t) = y_{\max}$, making use of the condition that $\bar{s}(\overline{x}_t, y_{\max}) \ge 1 - 2/\sqrt{T}$, we can bound the gradient as follows.

$$\begin{aligned}
&\mathbb{E}_{y_t\sim p(\overline{x}_t),y'_t\sim q_t}\left[\big|\big|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big|\big|^2\right]\\
&= \mathbb{E}_{y_t\sim p(\overline{x}_t)}\left[\frac{1}{q_t(y_t)}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,y_t)|_{\mathcal{K}_{A_t}}\big|\big|^2 + \sum_{n+j\in\overline{\mathcal{Y}}_{A_t}}\frac{(1-c_j(x_t,y_t))^2}{q_t(n+j)}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,n+j)|_{\mathcal{K}_{A_t}}\big|\big|^2\right]\\
&\le \sum_{y\in\mathcal{Y}}\frac{s(\overline{x}_t,y)}{q_t(y)}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,y)|_{\mathcal{K}_{A_t}}\big|\big|^2 + \sum_{n+j\in\overline{\mathcal{Y}}_{A_t}}\frac{s(\overline{x}_t,n+j)}{q_t(n+j)}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,n+j)|_{\mathcal{K}_{A_t}}\big|\big|^2\\
&\le \sum_{y\in\overline{\mathcal{Y}}_{A_t}\setminus\{y_{\max}\}}\frac{s(\overline{x}_t,y)}{q_t(y)}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,y)|_{\mathcal{K}_{A_t}}\big|\big|^2 + \frac{s(\overline{x}_t,y_{\max})}{q_t(y_{\max})}\big|\big|\nabla\Phi_{01}(\widetilde{W},\overline{x}_t,y_{\max})|_{\mathcal{K}_{A_t}}\big|\big|^2\\
&\le \frac{(\sum_{y\in\overline{\mathcal{Y}}_{A_t}\setminus\{y_{\max}\}}s(\overline{x}_t,y))|\overline{\mathcal{Y}}_{A_t}|\sqrt{t}}{\gamma}NR^2 + \frac{s(\overline{x}_t,y_{\max})}{q_t(y_{\max})}NR^2\\
&\le O\left(\frac{|\mathcal{Y}_{A_t}|^2}{\gamma}\right)NR^2 = O(N^{5/4}R^2).
\end{aligned} \tag{24}$$

On rounds $t\in[T]$ in which $|\mathcal{Y}_{A_t}|\ge 3$ and $h(\overline{x}_t)\ne y_{\max}$, combining Lemma E.3 and Lemma E.4 we have

$$\mathbb{E}_{y_t\sim p(\overline{x}_t),y'_t\sim q_t}\left[\sum_{t\in[T]}\left(\eta_t\big|\big|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\big|\big|^2 - \delta_t\right)\right] \le 2\eta_t\, v_{\max}\sqrt{N}R\,\mathcal{C}_{\Phi_{\text{def}}}(h_t,\overline{x}_t) - \frac{1-2/\sqrt{T}}{|\mathcal{Y}_A|+4}\mathcal{C}_{\Phi_{\text{def}}}(h_t,\overline{x}_t). \tag{25}$$

Note that by our choice of $\gamma$ $v_{\max} = \max_{t,j} v_{t,j} \leq \frac{N\sqrt{T}}{\gamma}\}$ with large enough $T$. Moreover, our choice of $\eta_t$ guarantees that

$$2\eta_t v_{\max}\sqrt{N}R\mathcal{C}_{\Phi_{\text{def}}}(h_t, \overline{x}_t) - \frac{1-2/\sqrt{T}}{|\mathcal{Y}_A|+4}\mathcal{C}_{\Phi_{\text{def}}}(h_t, \overline{x}_t) \leq 0. \tag{26}$$

Plugging inequalities (23), (24), and (26) into equation (22) gives us

$$\begin{aligned}\mathcal{R}_{\Phi_{\text{def}}}(\mathbf{h}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\Phi_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) - \sum_{t\in[T]}\delta_t &= \frac{NB^2}{\eta_T} + \mathbb{E}_{y_t\sim p(\overline{x}_t), y'_t\sim q_t}\left[\sum_{t\in[T]}\left(\eta_t\left|\left|\widehat{\nabla}_t|_{\mathcal{K}_{A_t}}\right|\right|^2 - \delta_t\right)\right] \\ &\leq \frac{NB^2}{\eta_T} + \sum_{t\in[T]} O(\eta_t N^{5/4}R^2) = O\left(B\sqrt{R^3N^{7/2}T}\right).\end{aligned} \tag{27}$$

Combining with inequality (20), by our choice of $\eta_t$ and $\gamma_t$ we have

$$\mathcal{R}_{\ell_{\text{def}}}(\mathbf{q}, \overline{\mathbf{x}}) - \mathcal{R}^*_{\ell_{\text{def}},\mathcal{H}}(\overline{\mathbf{x}}) = O\left(B\sqrt{R^3N^{7/2}T}\right), \tag{28}$$

which ends our proof. □

### E.3 Analysis of the minimizability gap

**Theorem E.5.** *(Characterization of minimizability gaps) Assume that $\mathcal{H}$ is symmetric and complete, for any sequence of inputs $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_T)$ and any distribution, the minimizability gaps can be characterized as follows:*

$$\mathcal{M}_{\Phi_{def},\mathcal{H}}(\bar{\mathbf{x}}) = \mathcal{R}^*_{\Phi_{def},\mathcal{H}}(\bar{\mathbf{x}}) - \sum_{t\in[T]}|\overline{\mathcal{Y}}_{A_t}|\big(S(\bar{\mathbf{x}}) - s(\bar{x}_t, y_{t,\max})\big).$$

*Proof.* By Lemma C.2 we can write the conditional surrogate $\Phi_{\text{def}}$-risk as follows:

$$\begin{aligned}\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) &= \sum_{y\in\overline{\mathcal{Y}}_A} s(x,y)\Phi_{01}(h,x,y) \\ &= \sum_{y\in\overline{\mathcal{Y}}_A} s(x,y)\left(\sum_{y'\neq y, y'\in\overline{\mathcal{Y}}_A}\max\{0, 1+h(\overline{x}, y')\}\right) \\ &= \sum_{y\overline{\mathcal{Y}}_A}\max\{0, 1+h(\overline{x}, y)\}(S(\overline{x}) - s(\overline{x}, y)),\end{aligned}$$

where $\sum_{y\in\overline{\mathcal{Y}}_A} h(\overline{x}, y) = 0$. Define $\mathcal{Y}_{\leq -1} = \{y \in \overline{\mathcal{Y}}_A : h(\overline{x}, y) \leq -1\}$, $\mathcal{Y}_{>1} = \overline{\mathcal{Y}}_A - \mathcal{Y}_{\leq -1}$. We can rewrite the conditional surrogate risk as

$$\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) = \sum_{y\in\mathcal{Y}_{>-1}}\big(1 + h(\overline{x}, y)\big)\big(S(\overline{x}) - s(\overline{x}, y)\big).$$

Define $y_{\max} = \arg\max_{y\in\overline{\mathcal{Y}}_A} s(\overline{x}, y)$, we have $S(\overline{x}) - s(\overline{x}, y) \geq S(\overline{x}) - s(\overline{x}, y_{\max})$ and

$$\sum_{y\in\mathcal{Y}_{>-1}}\big(1 + h(\overline{x}, y)\big) = |\mathcal{Y}_{>-1}| - \big(\sum_{y\in\mathcal{Y}_{\leq -1}} h(\overline{x}, y)\big) \geq |\mathcal{Y}_{>-1}| + |\mathcal{Y}_{\leq -1}| = |\overline{\mathcal{Y}}_A|.$$

Combining the two aforementioned inequalities gives us

$$\mathcal{C}_{\Phi_{\text{def}}}(h, \overline{x}) \geq |\overline{\mathcal{Y}}_A|\big(S(\overline{x}) - s(\overline{x}, y_{\max})\big).$$

Equality occurs when $h(\overline{x}, y) = \begin{cases} |\overline{\mathcal{Y}}_A| - 1 & \text{if } y = y_{\max} \\ -1 & \text{otherwise.} \end{cases}$

Note that there exists such $h$ in $\mathcal{H}$, which ends our proof.

□

# F Experiments

## F.1 Details of Experiment Settings

### F.1.1 Setting 1: Fixed Availability and Expertise

Here, all experts are continuously available, and each expert's domain of competence remains stationary throughout the learning horizon. That is each expert achieves perfect accuracy on its region of expertise, but is uniformly random outside of it. This setting corresponds to the standard assumption in classical L2D frameworks (Madras et al., 2018; Mozannar and Sontag, 2020), but we study it here under the online learning protocol, where data arrives sequentially.

A representative example is a clinical decision-support system that integrates several automated diagnostic models, each specialized for a distinct imaging modality (e.g., chest X-rays, MRIs, or CT scans). These expert models remain permanently accessible and their competence boundaries do not change over time, while the system continuously processes incoming patient data. This setup reflects realistic deployments where the expert pool is fixed but the input stream evolves, such as hospital triage platforms or other continuous-monitoring systems.

### F.1.2 Setting 2: Drifting Availability and Fixed Expertise

Let $G_{j,t} \in \{0, 1\}$ denote whether expert $g_j$ is available at round $t$. Availability is sampled through $G_{j,t} \sim \text{Bern}(p_{j,t})$, with $p_{j,t}$ following a random walk initialized at 0.7 with standard deviation $\sigma = 2 \times 10^{-3}$ on $[0, 1]$.

A practical realization of this setting occurs in multi-model inference systems deployed under dynamic resource constraints. For example, a pool of pre-trained expert networks—each specialized for a distinct input modality or domain—may share a limited GPU or compute budget. At any given time, only a subset of models can be queried due to bandwidth or scheduling limitations, even though their predictive functions remain unchanged. Similar behavior appears in edge-device ensembles, where network connectivity or energy constraints temporarily restrict access to otherwise stable models.

### F.1.3 Setting 3: Drifting Availability and Expertise

Expert availability evolves as before, following the same random-walk process. Meanwhile, each expert's region of expertise drifts gradually over time. We fix both the initial and final regions, separated by $10^5$ rounds. With these boundaries fixed, the experts' regional accuracies evolve according to a Brownian bridge. This construction models environments where both participation and competence vary over time while preserving the overall structure of the expert population.

A concrete analogue arises in collaborative clinical decision-support systems. Clinicians differ in their availability due to shifts or workload, and their diagnostic expertise naturally evolves as they gain experience, adapt to new medical guidelines, or incorporate feedback from prior cases. The learning system must therefore adapt continuously to changing access to human experts and to gradual shifts in their decision boundaries. This dynamic also captures other human–AI ecosystems—such as annotation pipelines or hybrid moderation teams—where both availability and skill levels evolve throughout deployment.

## F.2 Details of Synthetic Datasets Experiments

We evaluate our approach on synthetic datasets with $n_e = 3$ experts and $n = 6$ classes. Following the SynSep dataset in Kakade et al. (2008); van der Hoeven et al. (2021), we generate inputs $x_t \in \mathbb{R}^d$ with $d = 120$. Each input–label pair $(x_t, y_t)_{t=1}^T$ is sampled from one of $n$ linearly separable clusters $\Omega_y = \{(x_t, y_t) : y_t = y\}$.

To introduce label noise, for each $x_t \in \Omega_y$ we flip the true label to a uniformly random alternative class in $[n] \setminus \{y\}$ with probability $p_{y,t}$. The noise vector at round $t$ is $\mathbf{p}_t = (p_{y,t})_{y \in [n]}$. We initialize $\mathbf{p}_0 = [0.3, 0.3, 0.3, 0.3, 0.0, 0.0]$ and evolve it via a random walk with Gaussian perturbations of mean 0 and standard deviation $\sigma = 2 \times 10^{-3}$.

At initialization $t = 0$, expert $g_1$ is knowledgeable on classes $\{1, 2\}$ and predicts the *post-noise* labels correctly on inputs from $\Omega_1$ and $\Omega_2$, while predicting uniformly at random on other clusters. Expert $g_2$ is knowledgeable on $\{3, 4\}$ with the same behavior. Expert $g_3$ is a weak expert that predicts uniformly at random on all inputs.

In all settings, we run our algorithm with learning rate $\eta_t = \frac{\eta_0}{\sqrt{t}}$ where $\eta_0$ is tuned to $5 \times 10^{-4}$. The exploration is fixed to $\gamma_t = \min\{\frac{1}{2}, \frac{10}{\sqrt{t}}\}$.

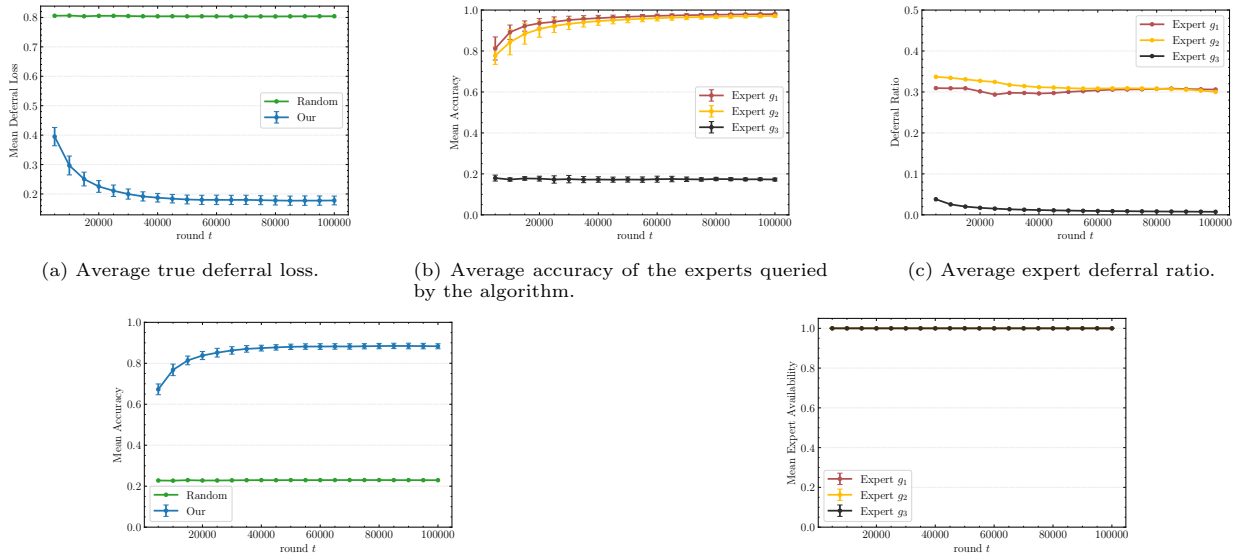
Note that in the first two settings where the expertise remains constant, our experiment design induces a clear optimal policy. A good learner must defer to experts when the input belongs to a region in $\{1, 2, 3, 4\}$ and make its own prediction on regions $\{5, 6\}$. In the third and final setting, the experts expertise regions gradually drift among $\{1, 2, 3, 4\}$.

**Setting 1: Fixed Availability and Expertise.** We report the results of this setting in Figure 3. Accuracies of experts are given in Table 2. Notice that those averages are significantly lower than the accuracy over queried rounds only (Figure 3b), which demonstrates our algorithm's ability to query experts adequately.

Table 2: Average Accuracies of Experts for setting 1.

| | **Expert** $g_1$ | **Expert** $g_2$ | **Expert** $g_3$ |
|---|---|---|---|
| Accuracy | 45.55 | 45.58 | 16.66 |

Our method rapidly converges to near-optimal predictive performance (Figure 3d) while maintaining a consistently low true deferral loss (Figure 3a), confirming an effective balance between autonomous prediction and expert deferral. Figures 3b and 3c show that experts $g_1$ and $g_2$ are queried most frequently and achieve near-perfect accuracy when selected, indicating that the algorithm learns to defer selectively to the most reliable experts. In contrast, the weaker expert $g_3$ exhibits both low accuracy (Figure 3b) and a negligible deferral ratio (Figure 3c), demonstrating that the model efficiently avoids uninformative experts. As expected in this fixed setting, expert availability remains constant across all rounds (Figure 3e).




(a) Average true deferral loss.

(b) Average accuracy of the experts queried by the algorithm.

(c) Average expert deferral ratio.

(d) Average accuracy.

(e) Expert availability.

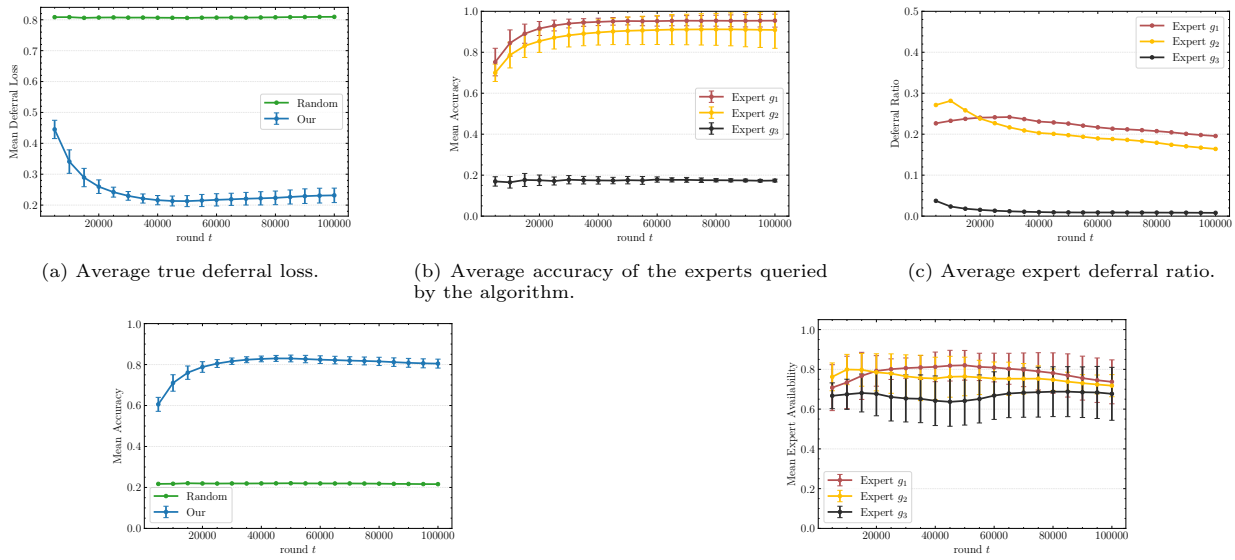
Figure 3: Results of the synthetic experiment for **setting 1: fixed availability and expertise**. Error bars represent standard deviations across five independent runs.

**Setting 2: Drifting Availability and Fixed Expertise.** We report results for this setting in Figure 4. Accuracies of experts are given in Table 3.

Table 3: Average Accuracies of Experts for setting 2.

| | **Expert** $g_1$ | **Expert** $g_2$ | **Expert** $g_3$ |
|---|---|---|---|
| Accuracy | 45.55 | 45.58 | 16.66 |

Compared to the previous experiment (Figure 3), overall performance declines slightly, as expected, due to the reduced and time-varying availability of experts. Nevertheless, once this availability constraint is taken into account, the observed deferral ratio remains effectively optimal (Figure 4c). Moreover, the near-perfect expert accuracy (Figure 4b) on deferred queries confirms that the algorithm maintains a well-calibrated decision boundary, consistently routing uncertain inputs to the most reliable experts.




(a) Average true deferral loss.

(b) Average accuracy of the experts queried by the algorithm.

(c) Average expert deferral ratio.

(d) Average accuracy.

(e) Expert availability.

Figure 4: Results of the synthetic experiment for **setting 2: fixed expert availability and expertise**. Error bars represent standard deviations across five independent runs.

**Setting 3: Drifting Availability and Expertise** We report results for this setting in Figure 6. Varying expertise of expert is shown in Figure 5.

Figure 5 shows that expert $g_1$ gradually shifts its expertise from region 2 to region 1, while expert $g_2$ moves in the opposite direction, from region 1 to region 2. Expert $g_3$ has no region of expertise, as it is uniformly random. The results show that our algorithm adapts effectively to these dynamic shifts. Even as expertise regions drift, the model continues to defer to the appropriate experts with high frequency, demonstrating stable calibration under evolving expertise (Figure 6b). Consequently, task accuracy and deferral loss remain strong and comparable to the fixed-expertise baseline (Figures 6a and 6d).

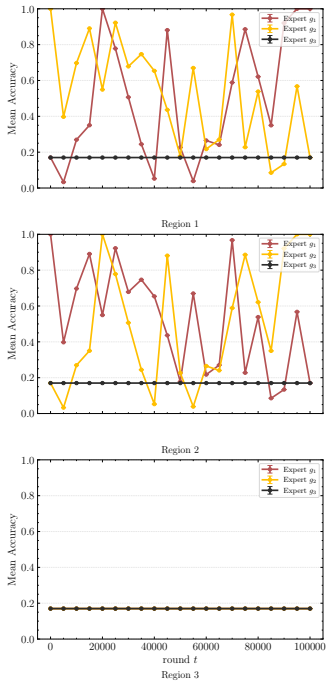
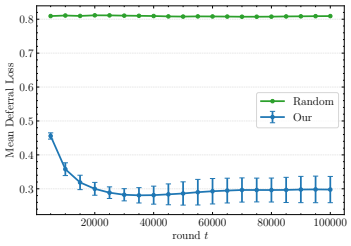
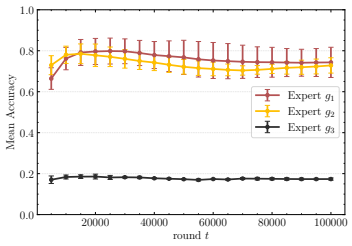


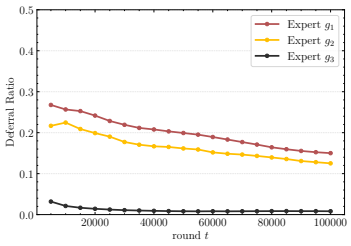

Figure 5: Evolution of the Expert Accuracies by Regions



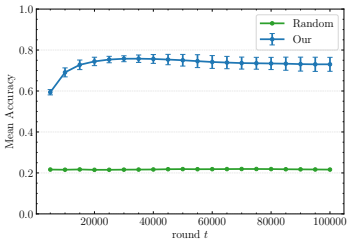

(a) Average true deferral loss.



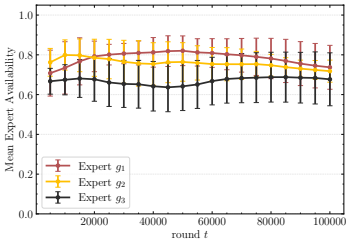

(b) Average accuracy of the experts queried by the algorithm.




(c) Average expert deferral ratio.




(d) Average accuracy.




(e) Expert availability.

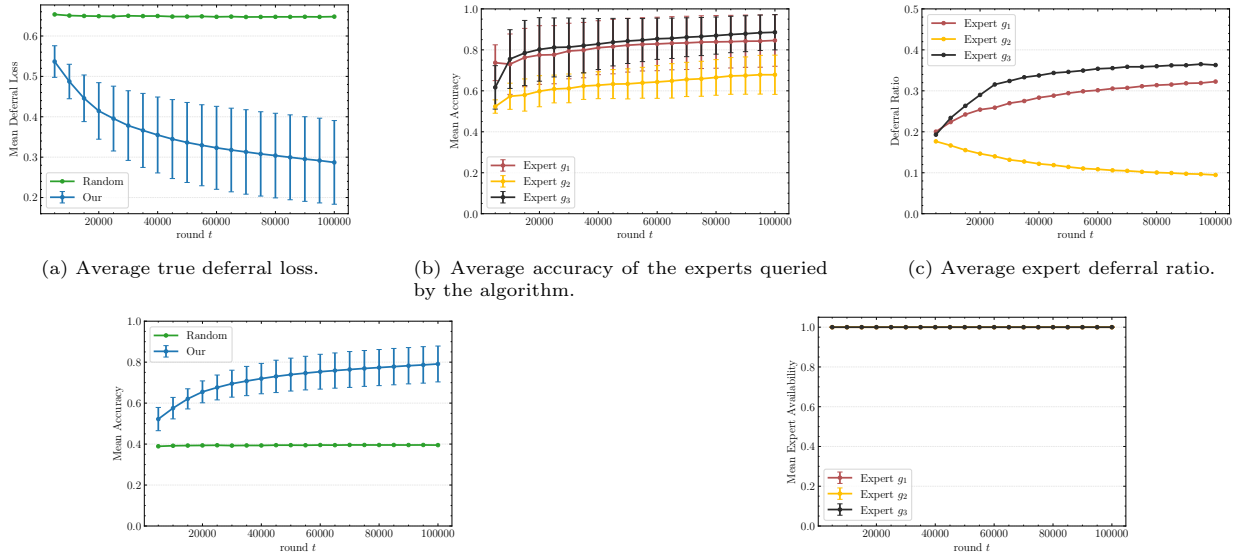
Figure 6: Results of the synthetic experiment for **setting 3: drifting availability and drifting expertise**. Whiskers denote standard deviations computed over 5 independent runs.

### F.3 Details of Reuters4 Dataset Experiments

**Setting 1: Fixed Availability and Expertise.** The results are summarized in Figure 7. Accuracies of experts are given in Table 4. The results show that the algorithm makes good use of given experts, shown by high expert accuracy in queried rounds and a high expert deferral ratio. This leads to a faster convergence rate for accuracy.

Table 4: Average Accuracies of Experts for setting 1 in Reuters4.

| | **Expert** $g_1$ | **Expert** $g_2$ | **Expert** $g_3$ |
|---|---|---|---|
| Accuracy | 62.21 | 61.09 | 50.01 |




(a) Average true deferral loss.

(b) Average accuracy of the experts queried by the algorithm.

(c) Average expert deferral ratio.

(d) Average accuracy.

(e) Expert availability.

Figure 7: Results of experiment on Reuters4 for **setting 1: fixed availability and expertise**. Whiskers denote standard deviations computed over 5 independent runs.

**Setting 2: Drifting Availability and Fixed Expertise.** Accuracies of experts are given in Table 5. Figure 8 show that the algorithm stays robust under varying expert availability. This is shown by the maintained high deferral ratio and high expert accuracies when queried.

Table 5: Average Accuracies of Experts for setting 2 in Reuters4.

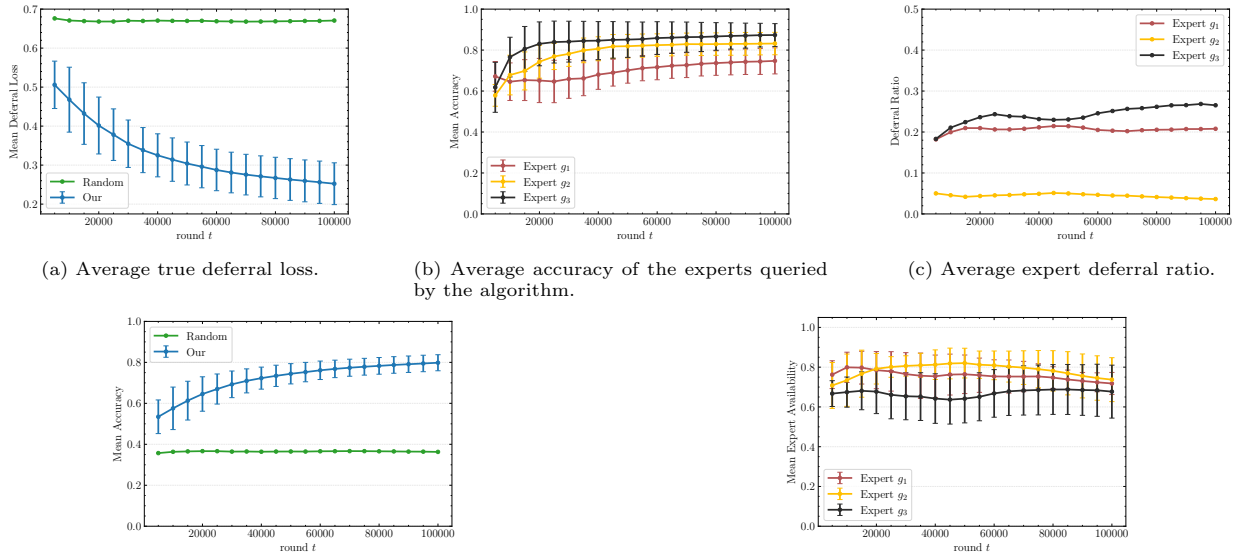
| | **Expert** $g_1$ | **Expert** $g_2$ | **Expert** $g_3$ |
|---|---|---|---|
| Accuracy | 62.21 | 61.09 | 50.01 |

**Setting 3: Drifting Availability and Drifting Expertise.** Figure 10 summarizes our results. Varying expertise of expert is shown in Figure 9. Even under challenging conditions with simultaneously varying availability and expertise, our approach maintains strong performance and adapts effectively to the evolving environment.

Figure 10 shows that expert $g_1$ gradually shifts its expertise from classes 1 and 2 to classes 3 and 4, expert $g_2$ shifts from classes 2 and 3 to classes 1 and 4, and expert $g_3$ shifts from classes 3 and 4 to classes 1 and 2. The results show that our algorithm adapts effectively to these dynamic shifts. Even as expertise regions drift, the model continues to defer to the appropriate experts with high frequency, demonstrating stable calibration under evolving expertise (Figure 10b). Consequently, task accuracy and deferral loss remain strong and comparable to the fixed-expertise baseline (Figures 10a and 10d).

### F.4 Details of CIFAR10H Dataset Experiments

In this section, we report additional results on the CIFAR10H dataset (Peterson et al., 2019), which augments CIFAR10 (Krizhevsky, 2009) with real human-annotated labels. We aggregate the human annotations into three




(a) Average true deferral loss.

(b) Average accuracy of the experts queried by the algorithm.

(c) Average expert deferral ratio.

(d) Average accuracy.

(e) Expert availability.

Figure 8: Results of experiment on Reuters4 for **setting 2: drifting availability and fixed expertise**. Whiskers denote standard deviations computed over 5 independent runs.

groups, each treated as a human expert $g_i$, $i = 1, 2, 3$. For an input image $x$, expert $g_i$ predicts by sampling uniformly at random from the labels in group $i$. We further introduce expert-specific noise regions to model heterogeneous expertise: $g_1$ is knowledgeable on labels $\{6, 7, 8, 9\}$, $g_2$ is knowledgeable on labels $\{0, 1, 2, 9\}$, $g_3$ is knowledgeable on labels $\{1, 2, 3, 4\}$. The mean accuracy of each expert after noise injection is reported in Table 6. Each expert $g_j$ has expert cost $c_j(x, y) = \mathbf{1}_{g_j(x) \neq y} + 0.05$.

Table 6: Average accuracies of experts in CIFAR10H post-noise.

| **Metric** | **Expert** $g_1$ | **Expert** $g_2$ | **Expert** $g_3$ |
|---|---|---|---|
| Mean Accuracy | 0.445 | 0.443 | 0.432 |

For this experiment, we use a Wide Residual Network (WideResNet) as the predictor $h \in \mathcal{H}$, instead of a linear hypothesis. To make expert consultation meaningful, we evaluate on corrupted inputs from CIFAR10C Hendrycks and Dietterich (2019) rather than on clean CIFAR10 images. In particular, we use 200,000 images corrupted with Gaussian noise at severity level 2. As a benchmark, we compare against a baseline that defers uniformly to one of the available experts whenever the classifier confidence falls below 0.5.

The results under fixed availability and expertise settings are summarized in Table 7, and the learning curves are shown in Figure 11. Our method achieves both lower mean deferral loss and higher mean accuracy than the baseline. This indicates that the learned policy attains a better trade-off between prediction performance and consultation cost, rather than relying on deferral in a naive manner. The queried expert accuracies are also substantially higher under our method, showing that the algorithm does not defer uniformly, but instead routes examples toward more suitable experts. The deferral ratios further support this behavior: the algorithm learns a non-uniform consultation pattern, which suggests that it is able to distinguish between examples that should be handled by the classifier and those for which expert advice is more beneficial.

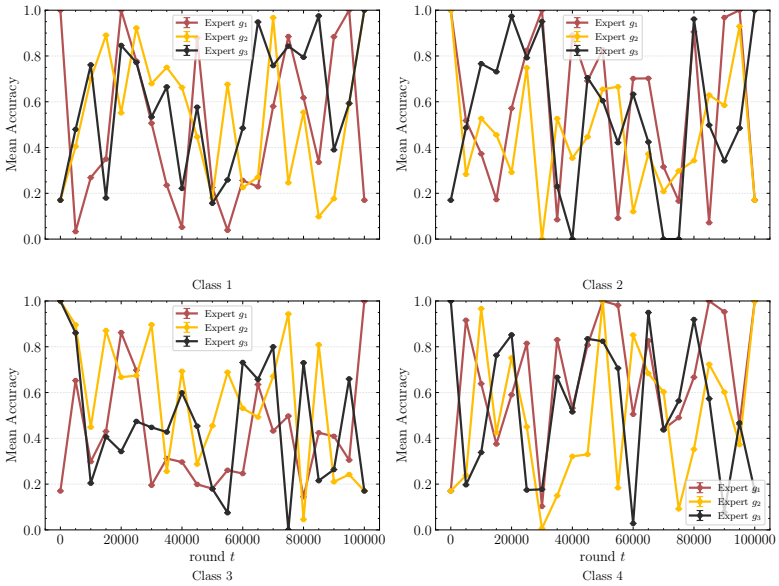
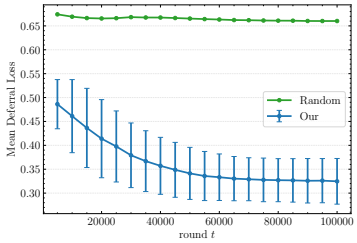


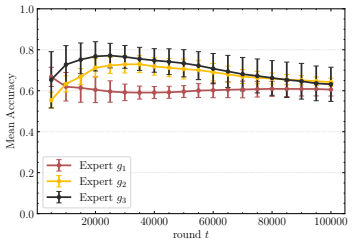

Figure 9: Expert Accuracies by Regions on Reuters4.

Table 7: Comparison between our method and the baseline, averaged over 5 runs. Better results are highlighted.

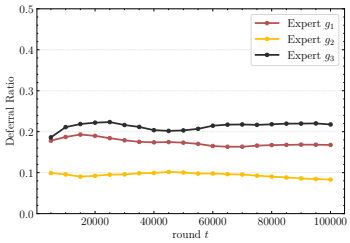
| Metric | Our Method | Baseline |
|---|---|---|
| Mean Deferral Loss | **0.3953** | 0.4133 |
| Mean Accuracy | **0.6519** | 0.6145 |
| Expert $g_1$ Queried Accuracy | **0.7253** | 0.4228 |
| Expert $g_2$ Queried Accuracy | **0.7967** | 0.3454 |
| Expert $g_3$ Queried Accuracy | **0.5764** | 0.4291 |
| Expert $g_1$ Deferral Ratio | **0.3233** | 0.1843 |
| Expert $g_2$ Deferral Ratio | **0.2266** | 0.1850 |
| Expert $g_3$ Deferral Ratio | **0.3981** | 0.1847 |



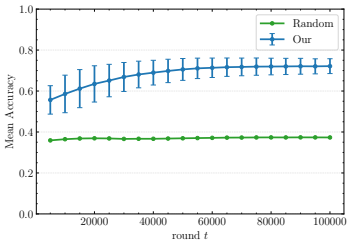

(a) Average true deferral loss.



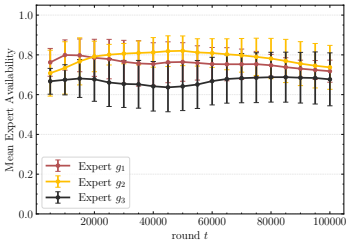

(b) Average accuracy of the experts queried by the algorithm.




(c) Average expert deferral ratio.

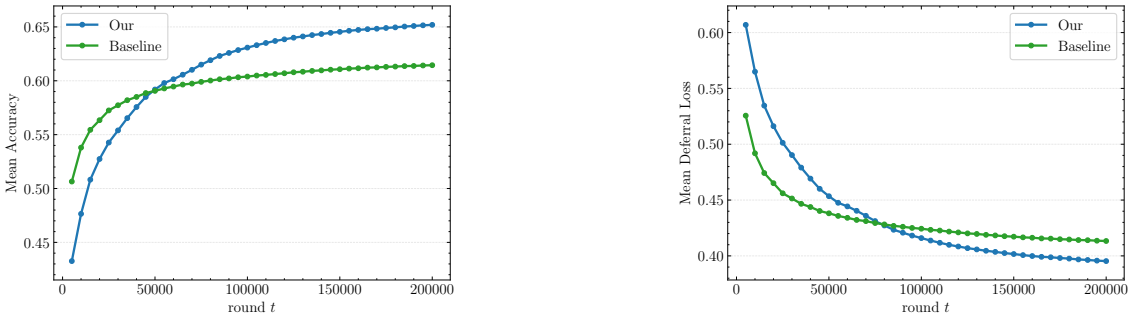



(d) Average accuracy.




(e) Expert availability.

Figure 10: Results of experiment on Reuters4 for **setting 3: drifting availability and drifting expertise**. Whiskers denote standard deviations computed over 5 independent runs.




(a) Average accuracy.

(b) Average true deferral loss.

Figure 11: Results of experiment on CIFAR10H with image corruption from CIFAR10C. Results are averaged over 5 independent runs.