

Budgeted Multiple-Expert Deferral

Giulia DeSalvo

Google DeepMind, Seattle

GIULIAD@GOOGLE.COM

Clara Mohri

Harvard University, Cambridge

CMOHRI@G.HARVARD.EDU

Mehryar Mohri

Google Research and Courant Institute of Mathematical Sciences, New York

MOHRI@GOOGLE.COM

Yutao Zhong

Google Research, New York

YUTAOZHONG@GOOGLE.COM

Abstract

Learning to defer uncertain predictions to costly experts offers a powerful strategy for improving the accuracy and efficiency of machine learning systems. However, standard training procedures for deferral algorithms typically require querying all experts for every training instance, an approach that becomes prohibitively expensive when expert queries incur significant computational or resource costs. This undermines the core goal of deferral: to limit unnecessary expert usage. To overcome this challenge, we introduce the *budgeted deferral* framework, which aims to train effective deferral algorithms while minimizing expert query costs during training. We propose new algorithms for both two-stage and single-stage multiple-expert deferral settings that selectively query only a subset of experts per training example. While inspired by active learning, our setting is fundamentally different: labels are already known, and the core challenge is to decide which experts to query in order to balance cost and predictive performance. We establish theoretical guarantees for both of our algorithms, including generalization bounds and label complexity analyses. Empirical results across several domains show that our algorithms substantially reduce training costs without sacrificing prediction accuracy, demonstrating the practical value of our budget-aware deferral algorithms.

1. Introduction

Learning algorithms can improve accuracy and efficiency by deferring uncertain predictions to experts, such as domain specialists or advanced pre-trained models. Examples include biomedical diagnosis with radiologists of different specialties and geological monitoring with sensor networks of varying capabilities. To do this effectively, it is essential to account for the cost associated with each expert, which may represent prediction quality, latency, or other computational aspects. These costs may vary with each input instance and depend on the possible output labels.

The problem of assigning each input to the most appropriate expert, balancing predictive accuracy with resource consumption is known as *learning to defer with multiple experts*. This task arises in many domains, including natural language generation with large language models (Wei et al., 2022; Bubeck et al., 2023), speech recognition, image classification, financial forecasting, and computer vision.

Recent work has extensively studied this problem in the context of classification (Hemmer et al., 2022; Keswani et al., 2021; Kerrigan et al., 2021; Straitouri et al., 2022; Benz and Rodriguez, 2022; Verma et al., 2023; Mao et al., 2023a, 2024a), and more recently in regression (Mao et al., 2024f) and multi-task learning settings (Montreuil et al., 2025d). Several algorithms with strong theoretical

results have been developed, including surrogate loss-based approaches for multiple-expert deferral with several consistency guarantees (Mao et al., 2025a).

These surrogate loss-based approaches can be broadly categorized into two settings (Mao, 2025). In the *single-stage* setting, the predictor and deferral function are learned jointly (Mozannar and Sontag, 2020; Verma and Nalisnick, 2022; Charusaie et al., 2022; Mozannar et al., 2023; Mao et al., 2024a,h). In contrast, the two-stage setting first trains a predictor, which is then held fixed and treated as an additional expert, while the deferral function is learned in a subsequent stage (Mao et al., 2023a). See Appendix A for further discussion.

However, despite extensive prior work, a central challenge remains: in many real-world applications, especially those involving costly experts such as LLMs, training a deferral algorithm can itself be computationally prohibitive. For each labeled example (x, y) , the outputs of all p experts must be computed, incurring the total cost of all experts, multiplied across the full training set. This is at odds with the core motivation of deferral: to reduce unnecessary expert usage.

Can we train effective deferral algorithms while reducing the computational burden? One natural idea is to selectively query only a subset of experts for each training instance, thereby lowering training cost. However, this introduces partial information and raises new questions: How should we choose which experts to query? Can we do so in a principled way that maintains performance guarantees for the deferral algorithm, while controlling computational cost?

This paper addresses these questions and introduces a formal study of this *budgeted deferral problem*, where the goal is to train deferral algorithms while minimizing the cost of querying experts. We propose new algorithmic solutions for both the two-stage multiple-expert deferral setting (Mao et al., 2023a) and the single-stage multiple-expert deferral setting (Verma et al., 2023; Mao et al., 2024a). While our methods draw inspiration from active learning (Beygelzimer et al., 2009; Cortes et al., 2019a,b, 2020), the setup is fundamentally different: unlike standard active learning, our training data consists of labeled examples, that is, pairs (x, y) are already known. The challenge is not whether to request a label, but rather which experts to query for each example, in order to balance cost and predictive performance.

Contributions. Our main contributions are as follows, and they also define the structure of the rest of this paper. In Section 3, we propose a budget-aware active querying algorithm for two-stage multiple-expert deferral. In Section 4, we introduce the *Sampling-Probs* subroutine used in our algorithm. This choice leads to favorable guarantees on the generalization bound of our algorithm. In Section 5, we establish a label complexity bound for our budgeted two-stage deferral algorithm. In the realizable case, the expected number of expert cost queries it issues scales as $\tilde{O}(\sqrt{T})$, a substantial improvement over the linear label complexity $n_e T$ incurred by standard two-stage methods that query all expert costs. For simplicity of exposition, the main body presents this square-root bound, while the stronger logarithmic bound (derived via Freedman’s inequality) is given in full detail in Appendix F. Even in the agnostic setting, the bound remains favorable when the optimal surrogate loss $\mathcal{E}^*$ is small. In Appendix E, we also present a novel algorithm for single-stage multiple-expert deferral with budgeted expert queries, supported by analogous theoretical guarantees. Finally, in Section 7, we report experimental results demonstrating the effectiveness of our methods, including comparisons with standard deferral baselines. Our budgeted deferral method matches the accuracy of the standard approach while reducing expert queries to below 40%, with even larger gains as the number of experts increases, demonstrating strong scalability to complex prediction tasks.

Novelty. This work introduces a novel two-stage deferral solution that significantly reduces expert cost queries in a cost-sensitive active learning framework. Unlike standard active learning,

our approach goes beyond deciding whether or not to query a label; instead, for each instance, we determine *which* expert cost to query and with what probability. We achieve this by extending existing active learning algorithms and conceptual tools to our setting. Our new technical solution offers strong generalization bounds and label complexity guarantees. In particular, in the realizable case the label complexity scales with the square-root of the time horizon T —and can be further sharpened to logarithmic dependence using Freedman’s inequality (see Appendix F). We also present a novel solution for the single-stage deferral setting, which incorporates an additional “no deferral” option. Our experiments across various binary and multi-class datasets demonstrate significant savings in expert queries (at most one per sample) while maintaining prediction accuracy comparable to full batch settings.

Related Work. The most relevant prior work to our study is by Reid et al. (2024), who model deferral to a single expert as a two-armed contextual bandit problem with budget constraints. This formulation enables the direct application of existing bandit algorithms with knapsack constraints (Agrawal and Devanur, 2016; Filippi et al., 2010; Li et al., 2017). However, extending these methods to multiple experts is non-trivial, and general bandit algorithms (Lattimore and Szepesvári, 2020) are not tailored to the deferral loss or its consistent surrogate losses, which are central to learning-to-defer approaches. Moreover, Reid et al. (2024) assume a generalized linear model for expert performance, an assumption that is often too restrictive in practice. While multi-armed bandit (MAB) algorithms are powerful tools, our problem does not naturally fit this framework. A direct mapping of experts to arms yields regret benchmarks (e.g., external or shifting regret) that are either uninformative or misaligned with the objectives of deferral. Even refined notions such as contextual or policy regret face challenges, as the reward definition must balance the immediate cost of querying an expert with the long-term benefit of acquiring labels for training. This trade-off, together with the dependence of generalization on the choice of policy class, makes it difficult to cast our setting as a standard bandit problem. For a detailed discussion, see Appendix B.

2. Two-stage expert deferral framework

We consider a standard multi-class classification setting with input space $\mathcal{X}$ and label space $\mathcal{Y} = [n] := \{1, \dots, n\}$ for $n \geq 2$ classes.

In the *two-stage multiple-expert deferral framework*, learning proceeds in two phases. In Stage 1, a multi-class classification predictor is trained and then fixed, becoming one of several available experts. In Stage 2, a *routing function* is learned, which selects, on a per-input basis, one of $n_e \geq 2$ predefined experts $g_1, \dots, g_{n_e}$ (including the trained predictor). Each expert is a scoring function $g_j: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, and its prediction on input x is $\mathbf{g}_j(x) = \operatorname{argmax}_{y \in \mathcal{Y}} g_j(x, y)$. The routing function $r \in \mathcal{R}$ selects an expert index according to:

$$r(x) = \operatorname{argmax}_{k \in [n_e]} r(x, k),$$

with ties broken deterministically. The set $\mathcal{R}$ is a finite hypothesis class of scoring functions $r: \mathcal{X} \times [n_e] \rightarrow \mathbb{R}$, and its cardinality is denoted by $|\mathcal{R}|$. Our results naturally extend to infinite $\mathcal{R}$ using covering numbers (see Appendix G).

The *two-stage deferral loss* is defined as:

$$\mathcal{L}_{\text{tdef}}(r, x, y, \mathbf{c}) = \sum_{k=1}^{n_e} c_k(x, y) \mathbb{1}_{r(x)=k},$$

where $c_k(x, y)$ is the cost incurred for selecting expert g_k and $\mathbf{c} = (c_1, \dots, c_{n_e})$ is the cost vector. A common choice is the 0-1 loss: $c_k(x, y) = \mathbb{1}_{g_k(x) \neq y}$, though c_k may also incorporate computation or fairness costs, for example $c_k(x, y) = \alpha_k \mathbb{1}_{g_k(x) \neq y} + \beta_k$, for some $\alpha_k, \beta_k > 0$. We adopt the 0-1 cost formulation for simplicity, following Mao et al. (2023a), but our results straightforwardly extend to other choices.

Since L_{def} is non-differentiable and its optimization is intractable due to the presence of indicator functions in its definition, we resort instead to a surrogate loss. The surrogate general loss function proposed by Mao et al. (2023a) is:

$$L(r, x, y, \mathbf{c}) = \sum_{k=1}^{n_e} (1 - c_k(x, y)) \ell(r, x, k),$$

where ℓ is a surrogate loss for multiclass classification. For example, if ℓ is the multiclass logistic loss (Verhulst, 1838, 1845; Berkson, 1944, 1951), then:

$$\ell(r, x, k) = \log \left(1 + \sum_{k' \neq k} e^{r(x, k') - r(x, k)} \right).$$

We assume (possibly after normalization) that L takes values in $[0, 1]$. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^{n_e}$. The *expected surrogate loss* or *generalization error* of a routing function r is defined as: $\mathcal{E}(r) = \mathbb{E}_{(x, y, \mathbf{c}) \sim \mathcal{D}} [L(r, x, y, \mathbf{c})]$, and the *best-in-class generalization error* over $\mathcal{R}$ is: $\mathcal{E}^*(\mathcal{R}) = \inf_{r \in \mathcal{R}} \mathcal{E}(r)$.

3. Budgeted deferral algorithm

How can we design a two-stage deferral solution that reduces expert queries? Our approach builds on an existing active learning algorithm (specifically, IWAL (Beygelzimer et al., 2009), but other similar algorithms could also be adapted). However, unlike standard active learning, our problem requires more than simply deciding whether or not to query a label. Instead, for each instance (x_t, y_t) , we must determine which expert costs $c_k(x_t, y_t)$ to query.

To address this, we decompose the decision into two parts: (1) selecting an expert k , and (2) determining the probability of querying $c_k(x_t, y_t)$ once k is chosen. As we show in our analysis, selecting experts uniformly gives the best worst-case bounds since all experts are treated equally. For the query probability $p_{t,k}$, we carefully design its expression based on the surrogate loss scores for expert k computed by each routing function in the current version space, along with the maximum cost in next section.

This formulation allows us to derive strong theoretical guarantees on the label complexity. Moreover, our experimental results empirically validate the effectiveness of this approach, demonstrating significant savings in expert queries. Remarkably, our algorithm queries at most one expert cost per sample.

Novelty. What sets our method apart is this novel adaptation of active learning tools to a deferral-based cost-sensitive setting. Unlike prior work that typically focuses on binary query decisions or uniform cost assumptions, we derive a principled two-stage mechanism that incorporates both expert selection and cost-sensitive querying into a single coherent framework.

Algorithm 1 outlines the core procedure of our budgeted two-stage deferral strategy in the multi-expert setting. For any $t \in [T]$ and $k \in [n_e]$, we denote by $q_{t,k}$ the probability of selecting expert k at

Algorithm 1 Budgeted Two-Stage Deferral with Multiple Experts (Subroutine SAMPLING-PROBS)

```

INITIALIZE  $S_0 \leftarrow \emptyset$ ;
for  $t = 1$  to  $T$  do
  RECEIVE( $x_t, y_t$ );
   $p_t \leftarrow$  SAMPLING-PROBS
  ( $x_t, y_t, \{x_s, y_s, \mathbf{c}_s, q_s, k_s, p_s, Q_s: 1 \leq s < t\}$ );
   $k_t \leftarrow$  SAMPLE( $n_e, q_t$ );
   $Q_{t,k_t} \leftarrow$  BERNOULLI( $p_{t,k_t}$ );
  if  $Q_{t,k_t} = 1$  then
     $c_{t,k_t} \leftarrow$  QUERY-COST( $k_t, (x_t, y_t)$ )
     $S_t \leftarrow S_{t-1} \cup \left\{ \left( x_t, y_t, c_{t,k_t}, \frac{1}{q_{t,k_t} p_{t,k_t}} \right) \right\}$ ;
  else
     $S_t \leftarrow S_{t-1}$ ;
  end if
   $r_t \leftarrow \operatorname{argmin}_{r \in \mathcal{R}} \sum_{(x,y,c,w) \in S_t} w(1-c)\ell(r, x, k_t)$ .
end for

```

time t . The optimal choice for the value of $q_t = (q_{t,1}, \dots, q_{t,n_e})$ is determined in Section 5, using our theoretical bounds.

At each round t , upon observing a labeled instance (x_t, y_t) , the learner calls a subroutine *Sampling-PROBS* (to be detailed later) that takes as input the current instance and historical data and returns the vector $p_t = (p_{t,1}, \dots, p_{t,n_e})$, where $p_{t,k}$ is the probability of querying the cost $c_{t,k}$ associated with expert k , conditioned on expert k being selected.

An expert k_t is then selected according to q_t , and its cost is queried with probability $Q_{t,k_t} \sim \text{BERNOULLI}(p_{t,k_t})$. The algorithm incrementally builds a labeled dataset, assigning an importance weight to each queried cost. Specifically, if expert k is selected and its cost $c_{t,k}$ is queried, the example is stored with a weight of $1/(q_{t,k} p_{t,k})$ to account for the sampling process.

Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^{n_e}$. Then, the expected surrogate loss of a hypothesis $r \in \mathcal{R}$ is:

$$\mathcal{E}(r) = \mathbb{E}_{(x,y,\mathbf{c}) \sim \mathcal{D}} [\mathcal{L}(r, x, y, \mathbf{c})] = \mathbb{E}_{(x,y,\mathbf{c}) \sim \mathcal{D}} \left[\sum_{k=1}^{n_e} (1 - c_k(x, y)) \ell(r, x, k) \right].$$

To estimate $\mathcal{E}(r)$ from data, the algorithm uses an importance weighted empirical estimate at time T :

$$\mathcal{E}_T(r) = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^{n_e} \frac{1_{k=k_t} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) \ell(r, x_t, k),$$

where $(k_t, Q_{t,k})$ are the random decisions used in sampling and querying at round t . It is straightforward to verify that this estimator is unbiased, i.e., $\mathbb{E}[\mathcal{E}_T(r)] = \mathcal{E}(r)$, where the expectation is over the internal randomness of the algorithm. Theorem 1 establishes high-probability concentration bounds for $\mathcal{E}_T(r)$, assuming the probabilities $p_{t,k}$ are chosen appropriately.

4. Sampling-Probs Strategy and Generalization Guarantees

A key component of the subroutine introduced in this section is the specific definition of the probabilities $p_{t,k}$, which depend on the selected expert k . We will demonstrate that this choice leads to favorable guarantees on the generalization bound and label complexity of our algorithm.

Algorithm 2 gives the pseudocode of the *Sampling-Probs* subroutine used within the budgeted two-stage multiple-expert deferral framework. This subroutine maintains a dynamically evolving subset of the hypothesis set (version space), denoted by $\mathcal{R}_t$, which is refined over time based on empirical performance.

Algorithm 2 Sampling-Probs Subroutine with Past History

```

INITIALIZE  $\mathcal{R}_0 \leftarrow \mathcal{R}$ ;
for  $t = 2$  to  $T$  do
     $\mathcal{E}_{t-1}(r) \leftarrow \frac{1}{t-1} \sum_{s=1}^{t-1} \sum_{k=1}^{n_e} \frac{1_{k=k_s} Q_{s,k}}{q_{s,k} p_{s,k}}$ 
     $\times (1 - c_{s,k}(x_s, y_s)) \ell(r, x_s, k)$ 
     $\mathcal{E}_{t-1}^* \leftarrow \min_{r \in \mathcal{R}_{t-1}} \mathcal{E}_{t-1}(r)$ ;
     $\mathcal{R}_t \leftarrow \{r \in \mathcal{R}_{t-1} : \mathcal{E}_{t-1}(r) \leq \mathcal{E}_{t-1}^* + \Delta_{t-1}\}$ ;
     $p_{t,k} \leftarrow \max_{r, r' \in \mathcal{R}_t} \{\ell(r, x_t, k) - \ell(r', x_t, k)\}$ .
end for
    
```

The version space is initialized with the full hypothesis class $\mathcal{R}$. After each round, it is pruned to retain only those hypotheses whose empirical error does not exceed that of the current best predictor (within $\mathcal{R}_t$) by more than a slack parameter Δ_t . Formally:

$$\mathcal{R}_{t+1} = \{r \in \mathcal{R}_t : \mathcal{E}_t(r) \leq \mathcal{E}_t^* + \Delta_t\}, \quad (1)$$

where $\mathcal{E}_t^* = \min_{r \in \mathcal{R}_t} \mathcal{E}_t(r)$ denotes the minimal empirical error at round t .

To control the size of Δ_t , we define $q_{\min} = \min_{k \in [n_e]} q_{t,k}$, which we assume is strictly positive without loss of generality, and let $\bar{q} = \frac{1}{q_{\min}} + 1$. Then, the threshold Δ_t is chosen as follows, based on standard sample complexity arguments:

$$\Delta_t = \sqrt{\bar{q}^2 \cdot 8/t \cdot \log(2t(t+1)|\mathcal{R}|^2/\delta)}. \quad (2)$$

This pruning strategy guarantees, with high probability, that the optimal predictor $r^* \in \mathcal{R}$ remains in the version space $\mathcal{R}_t$ at all times with high probability, while progressively eliminating suboptimal hypotheses (see Theorem 1).

For each instance (x_t, y_t) , the subroutine evaluates the informativeness of querying each expert's cost by examining the variability in the expert-specific component of the surrogate loss across hypotheses in $\mathcal{R}_t$, leveraging the decomposability of the loss. The sampling probability $p_{t,k}$ for expert k is then set to the maximum difference in this component over all pairs of hypotheses in $\mathcal{R}_t$:

$$\begin{aligned}
 p_{t,k} &= \max_{r, r' \in \mathcal{R}_t} \max_{c \in \{0,1\}} (1 - c) (\ell(r, x_t, k) - \ell(r', x_t, k)) \\
 &= \max_{r, r' \in \mathcal{R}_t} \{\ell(r, x_t, k) - \ell(r', x_t, k)\}.
 \end{aligned}$$

Since the surrogate loss is normalized within $[0, 1]$, the resulting sampling probabilities are also bounded in this range.

This design allocates the query budget adaptively, prioritizing experts and instances where the disagreement among remaining hypotheses is greatest, thus targeting high-uncertainty regions.

We now establish high-probability performance guarantees for the predictors output by the algorithm.

Theorem 1 (Two-Stage Generalization Bound)

Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^{n_e}$, and let $\mathcal{R}$ be a hypothesis class. Assume that $r^* \in \mathcal{R}$ minimizes the expected surrogate loss $\mathcal{E}(r)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $T \geq 1$:

- The optimal hypothesis r^* belongs to the retained set $\mathcal{R}_T$;
- For all $r, r' \in \mathcal{R}_T$, the generalization gap satisfies

$$\mathcal{E}(r) - \mathcal{E}(r') \leq 2\Delta_{T-1}.$$

In particular, the learned hypothesis r_T at time T satisfies

$$\mathcal{E}(r_T) - \mathcal{E}(r^*) \leq 2\Delta_{T-1}.$$

The proof is given in Appendix D.2. It relies on a concentration argument built around Lemma 10, which is stated and proved in Appendix D.1.

5. Label Complexity

In the previous section, we established that the generalization error of our budgeted two-stage deferral algorithm closely matches that of a standard deferral method with full access to all $n_e T$ expert costs. We now turn to analyzing the label complexity of our approach, that is, the expected number of expert cost queries it issues.

To derive label complexity guarantees for our algorithm, we must adapt existing tools and definition in active learning to our deferral setting. In particular, we will define a new notion of slope asymmetry, hypothesis distance metric, generalized disagreement coefficient, based on experts' costs c_k and tailored to our setting. These tools allow us to demonstrate that our budgeted deferral algorithm can achieve a favorable label complexity, in fact lower than its fully supervised counterpart when the learning problem is approximately realizable and the disagreement coefficient of the hypothesis set is not too large.

We focus on a family of multiclass surrogate losses ℓ that includes, among others, the multinomial logistic loss. We also assume a mild condition on the cost structure: for every input-label pair, there exists at least one expert that incurs zero cost. This assumption can be satisfied by expanding the expert pool to include a sufficiently diverse set of experts, such that at least one performs well on each instance. As in (Beygelzimer et al., 2009), a central property we require of the loss function ℓ is bounded slope asymmetry, which controls how differences in surrogate losses between hypotheses may be distorted by cost vectors. This condition is key to relating loss-based disagreement to label complexity.

Definition 2 (Slope Asymmetry for Two-Stage Deferral) The slope asymmetry of a multi-class loss function $\ell : \mathcal{R} \times \mathcal{X} \times [n_e] \rightarrow [0, \infty)$ is defined as: $K_\ell =$

$$\sup_{r, r', x, y} \frac{\max_{\mathbf{c} \in \{0, 1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(r, x, k) - \ell(r', x, k)|}{\min_{\mathbf{c} \in \{0, 1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(r, x, k) - \ell(r', x, k)|}.$$

This quantity is always well-defined and finite if, for every (x, y) , there exists at least one expert k^* with zero cost, $c_{k^*}(x, y) = 0$. In practice, K_ℓ is bounded for common convex surrogates such as the logistic loss, provided the range of score functions r is restricted to a compact interval (e.g., $[-B, B]$); see Appendix C for details and explicit bounds. Next, we define a distance measure over the hypothesis set that reflects variability in expert-specific loss components.

Definition 3 (Hypothesis Distance Metric) For any $r, r' \in \mathcal{R}$ and distribution $\mathcal{D}$, define $\rho(r, r') =$

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\mathbf{c} \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(r, x, k) - \ell(r', x, k)| \right].$$

Define ϵ -ball around r as $B(r, \epsilon) = \{r' \in \mathcal{R} : \rho(r, r') \leq \epsilon\}$.

Suppose $r^* \in \mathcal{R}$ minimizes the expected surrogate loss: $\mathcal{E}^* = \mathcal{E}(r^*) = \inf_{r \in \mathcal{R}} \mathcal{E}(r)$. At time t , the version space $\mathcal{R}_t$ contains only hypotheses with generalization error at most $\mathcal{E}^* + 2\Delta_{t-1}$. But how close are these hypotheses to r^* in ρ -distance? The following lemma provides an upper bound in terms of the slope asymmetry:

Lemma 4 For any distribution $\mathcal{D}$ and any multi-class loss function ℓ , we have $\rho(r, r^*) \leq K_\ell \cdot (\mathcal{E}(r) + \mathcal{E}^*)$ for all $r \in \mathcal{R}$.

The proof is provided in Appendix D.3. The following extends the notion of disagreement (Hanneke, 2007) to our setting:

Definition 5 The disagreement coefficient θ is the smallest value such that, for all $\epsilon > 0$,

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} \sup_{r \in B(r^*, \epsilon)} \sup_{k \in [n_e]} |\ell(r, x, k) - \ell(r^*, x, k)| \leq \theta \epsilon.$$

We now present an upper bound on the expected number of cost queries required by the algorithm. The proof is included in Appendix D.4.

Theorem 6 (Two-Stage Label Complexity Bound) Let $\mathcal{D}$ be a two-stage deferral distribution and $\mathcal{R}$ a hypothesis set. Suppose the loss function ℓ has slope asymmetry K_ℓ and the disagreement coefficient of the problem is θ . Then, with probability at least $1 - \delta$, the expected number of cost queries made by the budgeted two-stage deferral algorithm over T rounds is bounded by.

$$4\theta \cdot K_\ell \cdot \left(\mathcal{E}^* T + O\left((1/q_{\min} + 1)\sqrt{T \log(|\mathcal{R}|T/\delta)}\right) \right), \quad (3)$$

where the expectation is taken over the algorithm's randomness.

The theorem establishes a label complexity bound for our budgeted two-stage deferral algorithm. In the realizable case, the bound scales as $\tilde{O}(\sqrt{T})$, significantly improving over the linear label complexity $n_e T$ incurred by standard two-stage methods that query all expert costs. For simplicity of exposition, the main body presents this square-root bound, while the stronger logarithmic bound (derived via Freedman's inequality) is given in Theorem 22 in Appendix F. Even in the agnostic setting, the bound remains favorable when the optimal surrogate loss $\mathcal{E}^*$ is small. A high-probability version of this result is provided in Corollary 26 in Appendix H.

The dependence on the generalized disagreement coefficient is, in general, unavoidable, as shown by Hanneke (2014). However, this coefficient has been shown to be bounded for many common hypothesis classes, enabling meaningful guarantees in practice.

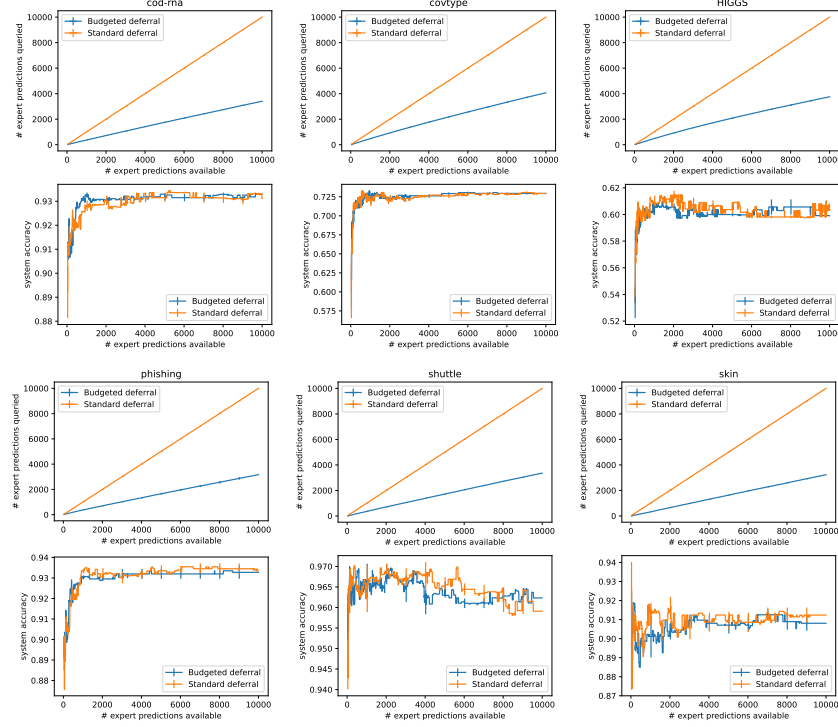


Figure 1: Standard vs. Budgeted Two-Stage Multiple-Expert Deferral on Binary Datasets.

Optimal q_t Finally, we note that both the generalization and label complexity bounds are minimized when the expert sampling probabilities are uniform: $q_{t,k} = 1/n_e$ for all t and k , yielding $q_{\min} = 1/n_e$ since all experts are treated symmetrically. Under this setting, the bounds simplify to:

$$\begin{aligned} \mathcal{E}(r_T) &\leq \mathcal{E}(r^*) + 2(n_e + 1)\sqrt{(8/(T-1))\log(2(T-1)T|\mathcal{R}|^2/\delta)} \\ \mathbb{E}\left[\sum_{t=1}^T \sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k}\right] &\leq 4\theta \cdot K_\ell \cdot \left(\mathcal{E}^*T + O\left((n_e + 1)\sqrt{T\log(|\mathcal{R}|T/\delta)}\right)\right). \end{aligned} \quad (4)$$

The linear dependence on n_e also appears in standard deferral methods. For example, [Mao et al. \(2024a, Theorem 3\)](#) establishes a generalization bound with linear dependence on the number of experts. By leveraging Freedman’s inequality ([Freedman, 1975](#)) in place of Azuma’s ([Mohri et al., 2018, Theorem D.7](#)) in our analysis, we can in fact derive learning and sample complexity bounds that depend only logarithmically on T in the realizable case (see Appendix F). This significantly strengthens our theoretical guarantees and further highlights the advantages of our approach.

6. Practical Implementation

In this section, we show that the budgeted two-stage deferral algorithm can be efficiently implemented for convex *comp-sum* losses ([Mao et al., 2023f](#)) and hypothesis classes $\mathcal{R}$ defined as convex sets of score functions $r: \mathcal{X} \times [n] \rightarrow \mathbb{R}$. The results apply not only to linear predictors but also to a broader range of models, including neural networks (where convexity is lost but the same optimization routines can be approximated via SGD).

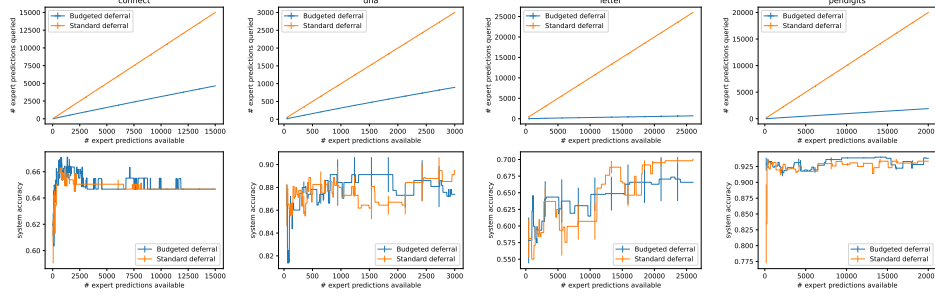


Figure 2: Standard vs. Budgeted Two-Stage Multiple-Expert Deferral on Multi-Class Datasets.

We consider *comp-sum losses* (Mao et al., 2023f) as the multi-class surrogate loss family ℓ , which includes many popular losses such as the multinomial logistic loss. A comp-sum loss is defined for any $(r, x, k) \in \mathcal{R} \times \mathcal{X} \times [n]$ as

$$\ell_{\text{comp}}(r, x, k) = \Psi\left(\frac{e^{r(x,k)}}{\sum_{k' \in [n]} e^{r(x,k')}}\right),$$

where $\Psi: [0, 1] \rightarrow \mathbb{R}_+ \cup \{+\infty\}$ is a non-increasing function. A notable instance is $\Psi(u) = -\log u$, which yields the *logistic loss* (Verhulst, 1838, 1845; Berkson, 1944, 1951). For suitable choices of Ψ , the loss ℓ_{comp} is convex in r .

Convex feasible region. At each round t , Algorithm 2 requires solving two optimization problems over the restricted hypothesis set $\mathcal{R}_t$, defined as the intersection of convex constraints accumulated up to round t : $\mathcal{R}_t = \bigcap_{t' < t} \left\{ r \in \mathcal{R}: \frac{1}{t'} \sum_{i=1}^{t'} \sum_{k=1}^{n_e} \frac{1_{k_i=k} Q_{i,k}}{q_{i,k} p_{i,k}} (1 - c_{i,k}(x_i, y_i)) \ell_{\text{comp}}(r, x_i, k) \leq \mathcal{E}_{t'}^* + \Delta_{t'} \right\}$. Since ℓ_{comp} is convex in r , each constraint defines a convex set, and thus $\mathcal{R}_t$ is convex.

First optimization. The first optimization problem at each round t computes the minimal empirical loss: $\mathcal{E}_t^* = \min_{r \in \mathcal{R}_t} \frac{1}{t} \sum_{i=1}^t \sum_{k=1}^{n_e} \frac{1_{k_i=k} Q_{i,k}}{q_{i,k} p_{i,k}} (1 - c_{i,k}(x_i, y_i)) \ell_{\text{comp}}(r, x_i, k)$. This is a convex program in r over the feasible region $\mathcal{R}_t$.

Second optimization. The second optimization determines the sampling probability $p_{t,k}$ by maximizing the difference in surrogate losses for each expert k : $\max_{r, r' \in \mathcal{R}_t} \left\{ \Psi\left(\frac{e^{r(x,k)}}{\sum_{k' \in [n]} e^{r(x,k')}}\right) - \Psi\left(\frac{e^{r'(x,k)}}{\sum_{k' \in [n]} e^{r'(x,k')}}\right) \right\}$. Since Ψ is non-increasing, this value is maximized when one term is minimized and the other maximized. Define $S_{\min}(x, k) \equiv \min_{r \in \mathcal{R}_t} \frac{e^{r(x,k)}}{\sum_{k' \in [n]} e^{r(x,k')}}$ and $S_{\max}(x, k) \equiv \max_{r \in \mathcal{R}_t} \frac{e^{r(x,k)}}{\sum_{k' \in [n]} e^{r(x,k')}}$. Then the optimal loss variation equals $\Psi(S_{\min}(x, k)) - \Psi(S_{\max}(x, k))$.

Interpretation. This shows that for convex comp-sum losses and convex hypothesis sets $\mathcal{R}$, both optimization problems required by Algorithm 2 are convex and can be solved efficiently. As an example, when $\mathcal{R}$ is the linear class $r(x, k) = \mathbf{w} \cdot \Phi(x, k)$ with $\|\mathbf{w}\| \leq B$, the optimizations reduce to convex problems in $\mathbf{w}$. For neural networks, the problems are no longer convex but can be handled in practice with standard stochastic gradient descent (SGD).

Heuristics. As in IWAL (Beygelzimer et al., 2009), practical heuristics can be used to further simplify the implementation. For the first optimization, one can minimize over $\mathcal{R}$ instead of $\mathcal{R}_t$, and for the second optimization, it suffices to impose only the most recent constraint (from round $t - 1$) instead of all past constraints. With these simplifications, the optimal solution remains within the feasible set, while computational efficiency is improved.

7. Experiments

This section evaluates the empirical performance of our proposed algorithm against an existing baseline in two-stage, multiple-expert learning-to-defer scenarios. We aim to confirm that our algorithm achieves two key outcomes predicted by our theoretical analysis: maintaining performance guarantees for deferral (i.e., minimizing deferral loss) and controlling computational cost, as measured by the number of queried expert predictions.

Experimental Setup For the budgeted two-stage multiple-expert deferral framework, we tested our approach, Algorithm 1, using ten publicly available benchmark datasets: shuttle, cod-rna, covtype, HIGGS, skin, phishing, dna, connect-4, letter, and pendigits¹. We compared our method to the state-of-the-art two-stage multiple-expert deferral algorithm (Mao et al., 2023a), which serves as the baseline. This algorithm imposes no budget restriction: it queries and uses the predictions of all n_e experts at every round, incurring a total cost of $t \times n_e$ expert predictions by time t . In the figures, the orange baseline curves correspond to this unconstrained state-of-the-art method, which incurs the full cost of $t \times n_e$ queries, while the blue curves correspond to our budgeted algorithm that makes at most one expert query per round.

We chose ℓ as the multinomial logistic loss for both our method and the baseline. For each dataset, the evaluation of the approaches was conducted over 5 trials, each with a random split of the unlabeled pool and test set. For each dataset, we generated a finite hypothesis set $\mathcal{H}$ of logistic regression models for deferral. For each dataset, features were normalized (zero mean, unit variance) and then scaled to ensure the maximum feature vector had unit norm.

Table 1 reports the number of features, the counts of test data points, the size of the hypothesis set, and the regularization parameter C . For all experiments, the finite hypothesis set consisted of logistic regression models from the scikit-learn library (Pedregosa et al., 2011). These models were trained using the `'liblinear'` solver for binary dataset and the `'lbfgs'` solver for multi-class datasets as well as an L_2 regularization parameter C , whose specific values are detailed in Table 1. Each hypothesis was trained on a random sample of data, with the sample size uniformly selected from the interval $[30, 500]$. We considered expert setups with clearly defined expertise across classes. Our approach assigns one expert per class. More precisely, for datasets with n classes, we define n experts, $g_1, \dots, g_n$, such that g_k is always correct on instances of the k -th class and predicts uniformly at random on instances from any other class.

The cost function is chosen as the 0–1 loss: $c_k(x, y) = \mathbb{1}_{g_k(x) \neq y}$. Following the IWAL-related literature (e.g., Cortes et al. (2019b,a); Amin et al. (2020)), we use publicly available datasets that are standard in active learning. We further adopt the zero-one cost from (Mao et al., 2023a) and expert setup from Mao et al. (2025a), where experts are accurate on specific classes. These choices enable controlled evaluation and ensure alignment with prior work.

1. <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

Table 1: Dataset Statistics: Number of Features (# F), Classes (# C), Test Data (# T), Model Class Size $|\mathcal{H}|$, and Regularization Parameter C .

DATASET	# F	# C	# T	$ \mathcal{H} $	$1/C$
Cod-rna	8	2	34535	4096	2^{-13}
Covtype	54	2	181012	4096	2^{-13}
Shuttle	9	2	14500	4096	2^{-13}
Skin	3	2	95057	2048	2^{-11}
Phishing	68	2	2055	2048	2^{-11}
Dna	180	3	1186	2048	2^{-13}
Letter	16	26	5000	2048	2^{-13}
Connect	126	3	66457	2048	2^{-13}
Pendigits	16	10	3498	2048	2^{-13}
HIGGS ²	28	2	10000	2048	2^{-13}

Results Figure 1 and Figure 2 demonstrate our method’s superior performance across ten datasets, achieving comparable system accuracy, defined as the average value of $[1 - L_{\text{tdef}}(h, x, y)]$ on the test data, to the standard deferral algorithm, while substantially reducing the number of expert queries.

The Expert Predictions Available (X-axis) represents the cumulative number of expert predictions that could have been made up to time step t , which equals $t \times n_e$. The Expert Predictions Queried (Y-axis of the top plots) is the cumulative number of times the algorithm actively queried an expert. Following Algorithm 1, this corresponds to $\sum_{s=1}^t Q_{s,k_s}$, while in the standard deferral setting we set this equal to $t \times n_e$.

In Figure 1 (binary datasets) and Figure 2 (multi-class datasets), our budgeted deferral method achieves the same accuracy while issuing far fewer expert queries than the standard method (which uses n_e expert predictions at every time step). The top plots show the exact reduction in queries, i.e., $\sum_{s=1}^t Q_{s,k_s} < t \times n_e$, while the bottom plots confirm that accuracy is preserved in both methods.

Because the standard method requires predictions from all n_e experts at every round, it accumulates a total of $t \times n_e$ expert predictions by time t . In contrast, our budgeted method makes at most one expert query per round. Consequently, the blue (budgeted) and orange (standard) curves diverge sharply in Figure 2, with the budgeted method achieving comparable accuracy while reducing expert queries by a large margin. This highlights the strong efficiency gains of our approach, particularly in settings with many experts.

While the exact percentage reduction varies by dataset, for binary datasets (Figure 1) it generally lies between 35–40%, already representing a substantial decrease. For multi-class datasets (Figure 2), the reduction is even larger, consistently below 30%. Moreover, as the number of experts increases, particularly in the `letter` dataset with 26 experts and the `pendigits` dataset with 10 experts, the performance improvements of our method become much more significant, underscoring its strong scalability to complex prediction tasks. Interestingly, in a previous version of this work, we stated that the multi-class results were less favorable than in the binary case; however, after correcting the analysis and properly accounting for the full cost in the multi-class setting of the non-budgeted systems, we find that the results are in fact more favorable.

Figure 1 further shows that system accuracy varied minimally across trials for all binary datasets, as indicated by their negligible error bars. In contrast, multi-class datasets in Figure 2 displayed

greater variation, particularly the `letter` dataset, which, with the largest number of classes among all tested datasets, exhibited wider error bars.

The fluctuations, or “jitters,” in the accuracy curves are not due to insufficient iterations but are inherent to the online, active learning process. They arise from stochastic components of the algorithm, including the random sampling of which expert to consider, the probabilistic decision of when to query, the dependency on individual data points, and the continual pruning of the hypothesis set.

Our empirical results suggest that some key components of the label complexity theoretical bound, the disagreement coefficient θ , the slope asymmetry K_ℓ , and the best-in-class loss $\mathcal{E}^*$, exert only a limited effect, thereby substantially reducing the impact of the higher-order term T .

8. Conclusion

Our results suggest that it is possible to train high-performing deferral algorithms while substantially reducing the cost of expert queries during training, thereby reconciling the practical constraints of real-world systems with the theoretical promise of learning to defer. This makes deferral strategies far more feasible in resource-constrained environments, especially in applications involving expensive experts such as large language models or human annotators. Our framework also opens the door to several possible extensions, including adaptive querying strategies that exploit structure across training examples, settings with dynamically available or context-dependent experts, as well as integration with reinforcement learning for sequential or interactive decision-making.

References

- Durmus Alp Emre Acar, Aditya Gangrade, and Venkatesh Saligrama. Budget learning via bracketing. In *International Conference on Artificial Intelligence and Statistics*, pages 4109–4119, 2020.
- Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In *Advances in neural information processing systems*, 2016.
- Jean Vieira Alves, Diogo Leitão, Sérgio Jesus, Marco OP Sampaio, Javier Liébana, Pedro Saleiro, Mario AT Figueiredo, and Pedro Bizarro. Cost-sensitive learning to defer to multiple experts with workload constraints. *Transactions on Machine Learning Research*, 2024.
- Kareem Amin, Corinna Cortes, Giulia DeSalvo, and Afshin Rostamizadeh. Understanding the effects of batching in online active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3482–3492, 2020.
- Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, pages 9804–9815, 2021a.
- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.
- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. $\mathcal{H}$ -consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, 2022a.

- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class $\mathcal{H}$ -consistency bounds. In *Advances in neural information processing systems*, 2022b.
- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094, 2023.
- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, 2024.
- Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.
- Nina L Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, pages 453–463, 2022.
- Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357–365, 1944.
- Joseph Berkson. Why I prefer logits to probits. *Biometrics*, 7(4):327–339, 1951.
- Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *International conference on machine learning*, pages 49–56, 2009.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie Gu, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *Advances in neural information processing systems*, 2022.
- Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *Advances in Neural Information Processing Systems*, 2023.
- Nicolo Cesa-Bianchi and Claudio Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.
- Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517, 2021.
- Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. In *Advances in Neural Information Processing Systems*, 2024.
- 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, 2022.

- Guanting Chen, Xiaocheng Li, Chunlin Sun, and Hanzhao Wang. Learning to make adherence-aware advice. In *International Conference on Learning Representations*, 2024.
- Xin Cheng, Yuzhou Cao, Haobo Wang, Hongxin Wei, Bo An, and Lei Feng. Regression with cost-based rejection. In *Advances in Neural Information Processing Systems*, 2023.
- C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- C.K. Chow. An optimum character recognition system using decision function. *IEEE Transactions on Computers*, 1957.
- Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82, 2016a.
- Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2016b.
- Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Active learning with disagreement graphs. In *International conference on machine learning*, 2019a.
- Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Region-based active learning. In *International Conference on Artificial Intelligence and Statistics*, 2019b.
- Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Adaptive region-based active learning. In *International Conference on Machine Learning*, pages 2144–2153, 2020.
- Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, pages 1–39, 2023.
- Corinna Cortes, Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Cardinality-aware set prediction and top- k classification. In *Advances in Neural Information Processing Systems*, 2024.
- Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *International Conference on Machine Learning*, 2025a.
- Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems*, 2025b.
- Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2611–2620, 2020.
- Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5905–5913, 2021.

- Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.
- Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.
- Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in neural information processing systems*, 2010.
- David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.
- Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2179–2187, 2021.
- Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. Human-ai collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614*, 2021.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.
- Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pages 2151–2159, 2019.
- Steve Hanneke. A bound on the label complexity of agnostic active learning. In *International Conference on Machine Learning*, pages 353–360, 2007.
- Steve Hanneke. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2-3): 131–309, 2014.
- 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. *arXiv preprint arXiv:2206.07948*, 2022.
- Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. *arXiv preprint arXiv:2304.07306*, 2023.
- Wenming Jiang, Ying Zhao, and Zehan Wang. Risk-controlled selective prediction for regression deep neural network models. In *International Joint Conference on Neural Networks*, pages 1–8, 2020.
- Wittawat Jitkrittum, Neha Gupta, Aditya K Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, 2023.
- Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*, 2021.

- Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in neural information processing systems*, 2008.
- Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In *Advances in Neural Information Processing Systems*, pages 4421–4434, 2021.
- Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080, 2017.
- Xiaocheng Li, Shang Liu, Chunlin Sun, and Hanzhao Wang. When no-rejection learning is optimal for regression with rejection. *arXiv preprint arXiv:2307.02932*, 2023.
- 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, 2024.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.
- Anqi Mao. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. PhD thesis, New York University, 2025.
- Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, 2023a.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds: Characterization and extensions. In *Advances in Neural Information Processing Systems*, 2023b.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, 2023c.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023d.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Structured prediction with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2023e.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023f.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024a.

- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pages 822–867, 2024b.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, pages 4753–4761, 2024c.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. H -consistency guarantees for regression. In *International Conference on Machine Learning*, pages 34712–34737, 2024d.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2024e.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *International Conference on Machine Learning*, 2024f.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems*, 2024g.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable H -consistent and Bayes-consistent loss functions for learning to defer. In *Advances in Neural Information Processing Systems*, 2024h.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable H -consistency and strong guarantees for learning to defer. In *International Conference on Machine Learning*, 2025a.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled algorithms for optimizing generalized metrics in binary classification. In *International Conference on Machine Learning*, 2025b.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Enhanced $\mathcal{H}$ -consistency bounds. In *International Conference on Algorithmic Learning Theory*, 2025c.
- Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.
- Mehryar Mohri and Yutao Zhong. Model margin noise and H -consistency bounds. In *International Symposium on Artificial Intelligence and Mathematics*, 2026.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.
- Yannis Montreuil, Shu Heng Yeo, 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. *arXiv preprint arXiv:2410.15761*, 2024.
- Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *International Conference on Machine Learning*, 2025a.

- Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Why ask one when you can ask k ? two-stage learning-to-defer to the top- k experts. *arXiv preprint arXiv:2504.12988*, 2025b.
- Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. One-stage top- k learning-to-defer: Score-based surrogates with theoretical guarantees. *arXiv preprint arXiv:2505.10160*, 2025c.
- Yannis Montreuil, Shu Heng Yeo, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. In *International Conference on Machine Learning*, 2025d.
- Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087, 2020.
- Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5323–5331, 2022.
- Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pages 10520–10545, 2023.
- Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, pages 29292–29304, 2022.
- Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to reject meets long-tail learning. In *International Conference on Learning Representations*, 2024.
- Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, pages 2582–2592, 2019.
- Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, pages 9140–9151, 2021.
- Filippo Palomba, Andrea Pugnana, José Manuel Alvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. *arXiv preprint arXiv:2405.18902*, 2024.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Melanie F Pradier, Javier Zazo, Sonali Parbhoo, Roy H Perlis, Maurizio Zazzi, and Finale Doshi-Velez. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Summits on Translational Science Proceedings*, 2021:525, 2021.

- Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.
- Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554, 2018.
- Mirabel Reid, Tom Sühr, Claire Vernade, and Samira Samadi. Online decision deferral under budget constraints. *arXiv preprint arXiv:2409.20489*, 2024.
- Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pages 19598–19615, 2022.
- Eleni Straitouri, Adish Singla, Vahid Balazadeh Meresht, and Manuel Gomez-Rodriguez. Reinforcement learning under algorithmic triage. *arXiv preprint arXiv:2109.11328*, 2021.
- Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Provably improving expert predictions with conformal prediction. *arXiv preprint arXiv:2201.12006*, 2022.
- Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *International Conference on Artificial Intelligence and Statistics*, pages 3475–3483, 2024.
- Pierre François Verhulst. Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique*, 10:113–121, 1838.
- Pierre François Verhulst. Recherches mathématiques sur la loi d’accroissement de la population. *Nouveaux Mémoires de l’Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1–42, 1845.
- Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202, 2022.
- Rajeev Verma, Daniel Barrejón, 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*, pages 11415–11434, 2023.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *CoRR*, abs/2206.07682, 2022.
- Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-ai dependence for learning to defer. In *International Conference on Machine Learning*, 2024.
- Yair Wiener and Ran El-Yaniv. Agnostic selective classification. In *Advances in Neural Information Processing Systems*, 2011.
- Yair Wiener and Ran El-Yaniv. Pointwise tracking the optimal regression function. In *Advances in Neural Information Processing Systems*, 2012.

- Yair Wiener and Ran El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.
- Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.
- Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(1), 2010.
- Ming Yuan and Marten Wegkamp. SVMs with a reject option. In *Bernoulli*, 2011.
- Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to knn. In *Advances in Neural Information Processing Systems*, pages 20073–20082, 2020.
- Zheng Zhang, Cuong Nguyen, Kevin Wells, Thanh-Toan Do, David Rosewarne, and Gustavo Carneiro. Coverage-constrained human-ai cooperation with multiple experts. *arXiv preprint arXiv:2411.11976*, 2024.
- Jason Zhao, Monica Agrawal, Pedram Razavi, and David Sontag. Directing human attention in event localization for clinical timeline creation. In *Machine Learning for Healthcare Conference*, pages 80–102, 2021.
- Yutao Zhong. *Fundamental Novel Consistency Theory: H-Consistency Bounds*. PhD thesis, New York University, 2025.

Contents of Appendix

A Related Work	23
B Relevance of Bandit Algorithms	25
B.1 Comparison with Contextual Bandit Frameworks	26
B.1.1 Mismatch 1: Primary Objective and Performance Benchmark	26
B.1.2 Mismatch 2: Label Complexity and the Nature of Queries	26
B.1.3 Mismatch 3: The Reward Signal	27
C Boundedness of Slope Asymmetry	27
D Proofs of Budgeted Two-Stage Deferral	29
D.1 Lemma 10 and Proof	29
D.2 Proof of Theorem 1	30
D.3 Proof of Lemma 4	30
D.4 Proof of Theorem 6	31
E Budgeted Single-Stage Multiple-Expert Deferral	32
E.1 Deferral and Surrogate Loss	32
E.2 Algorithm Overview	32
E.3 Sampling-Probs Strategy and Generalization Guarantees	33
E.4 Label Complexity	35
E.5 Practical Implementation	38
F Improved Sample Complexity and Label Complexity Bounds	39
G Budgeted Deferral with ϵ-Cover	43
H High-Probability Label Complexity Bounds	44

Appendix A. Related Work

The paradigm of single-stage learning to defer, where a predictor and a deferral function are jointly trained, was established by Cortes, DeSalvo, and Mohri (2016a,b, 2023) and has since spurred considerable research. This includes foundational studies on abstention with constant costs (Mao et al., 2024b,c; Mohri et al., 2024; Cao et al., 2022; Charoenphakdee et al., 2021; Cheng et al., 2023; Li et al., 2023; Narasimhan et al., 2024) and, more aligned with complex real-world scenarios, deferral involving instance- and label-dependent costs (Mao et al., 2024a; Cao et al., 2023; Mao et al., 2024h; Mozannar and Sontag, 2020; Mozannar et al., 2023; Verma and Nalisnick, 2022; Verma et al., 2023; Wei et al., 2024). In this established setup, the deferral function’s primary role is to optimally assign input instances to the most suitable expert. While this L2D paradigm offers distinct advantages over traditional *confidence-based* rejection methods (Chow, 1957; Ni et al., 2019; Yuan and Wegkamp, 2011; Bartlett and Wegkamp, 2008; Chow, 1970; Jitkrittum et al., 2023; Ramaswamy et al., 2018; Yuan and Wegkamp, 2010) and *selective classification* techniques that typically use fixed selection rates and cannot adapt to expert-modeled cost functions (Acar et al., 2020; El-Yaniv et al., 2010; El-Yaniv and Wiener, 2012; Gangrade et al., 2021; Geifman and El-Yaniv, 2017, 2019; Jiang et al., 2020; Shah et al., 2022; Wiener and El-Yaniv, 2011, 2012, 2015; Zaoui et al., 2020), these studies largely operate under the assumption that expert costs, even if variable, are specified or can be readily queried to train the deferral model. The distinct challenge of actively managing a budget for querying these expert costs during the learning phase of the deferral mechanism itself generally remains outside their primary scope, a research gap our work specifically targets.

The *learning to defer* (L2D) problem, which integrates human expert decisions into the cost function, was initially framed by Madras et al. (2018) and has been extensively studied (Mao et al., 2023a, 2024a; Charusaie et al., 2022; Mao et al., 2024h; Mozannar and Sontag, 2020; Mozannar et al., 2023; Pradier et al., 2021; Raghu et al., 2019; Verma and Nalisnick, 2022; Wilder et al., 2021). A core tenet is the formulation of a deferral loss function incorporating instance-specific costs for each expert. However, direct optimization of this loss is often intractable for commonly used hypothesis sets. Consequently, L2D algorithms typically resort to optimizing surrogate loss functions. This naturally raises the crucial question of the theoretical guarantees associated with such surrogate-based optimization.

This question, centered on the consistency guarantees of surrogate losses with respect to the target deferral loss, has been analyzed under two primary scenarios (Mao, 2025): the *single-stage* scenario, where predictor and deferral functions are learned jointly (Mao et al., 2024a; Charusaie et al., 2022; Mozannar and Sontag, 2020; Mozannar et al., 2023; Verma and Nalisnick, 2022), and a *two-stage* scenario, where a pre-trained predictor (acting as an expert) is fixed, and only the deferral function is subsequently learned (Mao et al., 2023a). While these consistency analyses are vital for understanding the reliability of L2D approaches, they generally presuppose that the necessary label and cost information for training and evaluating consistency is either given or can be acquired without explicit budgetary constraints that fundamentally shape the learning algorithm’s design. Our work diverges by concentrating on scenarios where the act of querying expert costs is itself a budgeted component integral to the training loop, necessitating a framework that explicitly manages this cost-acquisition process.

In particular, for the single-stage *single-expert* case, Mozannar and Sontag (2020), Verma and Nalisnick (2022), and Charusaie et al. (2022) proposed surrogate losses by generalizing standard classification losses. A key development by Mozannar et al. (2023) later revealed that these surrogates

did not satisfy *realizable $\mathcal{H}$ -consistency*, leading them to suggest an alternative. This was further refined by Mao et al. (2024h), who introduced a broader family of surrogates achieving Bayes-consistency, realizable $\mathcal{H}$ -consistency, and *$\mathcal{H}$ -consistency bounds* (Awasthi et al., 2022a,b; Mao et al., 2023f) (see also (Awasthi et al., 2021a,b, 2023, 2024; Mao et al., 2023b,e,c,d, 2024d,e,g; Cortes et al., 2024; Mao et al., 2025c,b; Zhong, 2025; Cortes et al., 2025a,b; Mohri and Zhong, 2026)). For the single-stage *multiple-expert* setting (Benz and Rodriguez, 2022; Hemmer et al., 2022; Kerrigan et al., 2021; Keswani et al., 2021; Straitouri et al., 2022), Verma et al. (2023) and Mao et al. (2024a) extended earlier surrogate losses. However, these extensions often inherit limitations regarding realizable $\mathcal{H}$ -consistency (Mao et al., 2025a). In the *two-stage* scenario, Mao et al. (2023a) introduced surrogate losses with strong consistency properties for constant costs, though their realizable $\mathcal{H}$ -consistency does not straightforwardly extend to practical classification error-based costs. These important theoretical investigations primarily concern the fidelity of surrogate losses, assuming the data (including expert costs) for loss computation is accessible for training. They do not typically address the strategic, budgeted acquisition of this costly expert information, which is a central focus of our research. Other extensions cover regression (Mao et al., 2024f), deferral to populations (Tailor et al., 2024), multi-task learning (Montreuil et al., 2025d), and aspects of adversarial robustness (Montreuil et al., 2025a), top- k learning (Montreuil et al., 2025b,c) or specialized query mechanisms (Montreuil et al., 2024), which, while related to querying, often adopt different problem formulations from our specific emphasis on minimizing expert query costs during the primary training of the deferral model under a budget.

The landscape of L2D extends beyond initial model training, with considerable work focusing on post-hoc refinements and deeper theoretical explorations. For instance, some studies propose alternative optimization schemes for the predictor and rejector components (Okati et al., 2021), or offer methods to ameliorate issues like underfitting in surrogate losses (Liu et al., 2024; Narasimhan et al., 2022). Frameworks for unified post-processing in multi-objective L2D have also been developed, drawing on principles like the generalized Neyman-Pearson Lemma (Charusaie and Samadi, 2024; Neyman and Pearson, 1933). Concurrently, theoretical advancements continue to shed light on L2D mechanisms; examples include the introduction of an asymmetric softmax function to derive more robust probability estimates (Cao et al., 2023), and investigations into dependent Bayes optimality to better understand the interdependencies in deferral decisions (Wei et al., 2024). The practical impact of these collective advancements is evident in the successful application of L2D and its variants across a range of domains, including regression tasks, human-in-the-loop systems, and reinforcement learning scenarios (Chen et al., 2024; De et al., 2020, 2021; Gao et al., 2021; Hemmer et al., 2023; Joshi et al., 2021; Mozannar et al., 2022; Palomba et al., 2024; Straitouri et al., 2021; Zhao et al., 2021). These contributions, while advancing the field, typically engage with models or data where expert interactions are already defined or have occurred, rather than focusing on the strategic acquisition of costly expert knowledge during the primary training phase.

Separately, another avenue of studies address deferral decisions under operational constraints or workload considerations, often with a focus on the inference phase (Charusaie and Samadi, 2024; Alves et al., 2024; Zhang et al., 2024). While this line of research is crucial for using L2D systems effectively, its foundational assumptions differ markedly from our work. Such studies generally proceed with the assumption that expert query costs are predetermined and fully known during the training stage. In direct contrast, our research is grounded in a budgeted deferral setting. The central objective here is to train deferral algorithms that not only learn the deferral policy but do so while

actively minimizing the expert querying expenses that are incurred and accounted for as an integral component of the learning process itself, without prior knowledge of these costs.

The most relevant prior work to our study is by [Reid et al. \(2024\)](#), who model deferral to a single expert as a two-armed contextual bandit problem with budget constraints. This formulation is interesting and allows for the direct application of existing bandit algorithms with knapsack constraints ([Agrawal and Devanur, 2016](#); [Filippi et al., 2010](#); [Li et al., 2017](#)). However, extending these methods to multiple experts is non-trivial. Moreover, such general bandit algorithms ([Lattimore and Szepesvári, 2020](#)) are not tailored to the deferral loss or its consistent surrogate losses, which are central to learning-to-defer approaches. Importantly, [Reid et al. \(2024\)](#) assume a generalized linear model for the expert and model performance ([Li et al., 2017](#)), an assumption that is often too restrictive for practical deferral scenarios.

Appendix B. Relevance of Bandit Algorithms

While multi-armed bandit (MAB) algorithms are powerful tools, our problem does not naturally fit into this framework. This section clarifies why both standard and contextual bandit formulations are misaligned with our setting.

Reward definition mismatch. Our objective depends jointly on the computational cost of querying an expert and the benefit of using that information to improve the generalization of the deferral algorithm. It is unclear how such a two-fold objective could be meaningfully expressed as a single reward for the “action” (choice of expert k) in a bandit framework.

Ambiguous benchmark. The standard contextual bandit benchmark, the best policy in hindsight is ill-defined here, since our guarantees concern both generalization and label complexity. A single “best policy” does not capture both objectives simultaneously.

Mismatch of guarantees. Even if one could design a meaningful bandit formulation, the resulting algorithms would not directly yield the dual guarantees we seek: (i) strong generalization for the deferral predictor, and (ii) favorable label complexity. These guarantees are central to our analysis and theoretical contributions.

Adversarial bandits. In adversarial settings (e.g., EXP3), treating each expert as an arm leads to regret benchmarks such as external or shifting regret, which are either uninformative or inadequate for deferral. More refined notions such as policy regret or contextual regret also face challenges, since the reward definition remains ambiguous, and the quality of the benchmark depends heavily on the choice of policy class. Moreover, the trade-off between the short-term cost of querying and the long-term benefit of acquired labels is not captured in these formulations.

Prior work. [Reid et al. \(2024\)](#) formulate deferral to a single expert as a two-armed contextual bandit problem with budget constraints, enabling the use of general bandit algorithms with knapsack constraints ([Agrawal and Devanur, 2016](#); [Filippi et al., 2010](#); [Li et al., 2017](#)). However, extending such methods to multiple experts is non-trivial and remains unexplored in that framework. More importantly, these general-purpose algorithms are not designed to minimize deferral loss or to leverage consistent surrogate losses, both of which are central to our framework. Furthermore, the generalized linear model assumption made by [Reid et al. \(2024\)](#) can be overly restrictive in practical scenarios involving black-box experts (e.g., large language models).

Summary. For these reasons, it is difficult to cast our problem as a standard bandit problem while still achieving the type of dual guarantees, generalization and label complexity bounds, that our analysis provides.

B.1. Comparison with Contextual Bandit Frameworks

While the problem of budgeted deferral involves sequential, context-dependent decisions, framing it within a standard contextual bandit setting is problematic. The core reason is a fundamental mismatch between the objectives and theoretical guarantees of each framework. Our approach is designed to solve a pool-based active learning problem, whereas contextual bandits are designed for online learning and performance optimization. We detail the key distinctions below.

B.1.1. MISMATCH 1: PRIMARY OBJECTIVE AND PERFORMANCE BENCHMARK

The primary objective of our framework is to train a final, static routing function r_T that exhibits low *generalization error*, $\mathcal{E}(r_T)$, on unseen data drawn from the true distribution $\mathcal{D}$. The querying strategy is a mechanism to gather the most *informative* labels to train this high-quality global model efficiently. Our performance is benchmarked against the best possible router in the hypothesis class, $r^* = \operatorname{argmin}_{r \in \mathcal{R}} \mathcal{E}(r)$, and our guarantees bound the excess error $\mathcal{E}(r_T) - \mathcal{E}(r^*)$.

In contrast, the objective of a contextual bandit algorithm is to maximize the *cumulative immediate reward* over a sequence of T rounds. Its performance is measured by *regret*, which compares its total reward to that of the best policy π^* from a fixed policy class Π in hindsight:

$$\operatorname{Reg}_T = \sum_{t=1}^T R(x_t, \pi^*(x_t)) - \sum_{t=1}^T R(x_t, a_t),$$

where a_t is the action (expert) chosen at round t . A low-regret algorithm is effective at online decision-making but provides no direct guarantee on the generalization performance of a model trained from its experiences. An algorithm could achieve low regret by exploiting the single best expert for a given context, thereby failing to gather the diverse data needed to learn the more nuanced decision boundaries of the globally optimal router r^* .

B.1.2. MISMATCH 2: LABEL COMPLEXITY AND THE NATURE OF QUERIES

Our framework explicitly addresses the dual goal of achieving low generalization error while minimizing the cost of supervision. This is reflected in our *label complexity* bounds.

- *Active Querying Framework (Ours):* The decision to query an expert is an integral part of the algorithm. The total number of queries, $\sum_{t=1}^T Q_{t,k_t}$, is a random variable that our analysis proves to be small, for instance, sublinear in T ($O(\log T)$ in the realizable case). The algorithm learns effectively even from examples where it makes no query, by using them to prune the version space.
- *Contextual Bandit Framework:* In a standard bandit setting, the agent *must* pull an arm (i.e., query an expert) at every timestep $t = 1, \dots, T$. The total number of queries is therefore always T . There is no built-in mechanism for cost-free learning. While one could introduce a “do not query” arm, defining its reward is non-trivial, as it involves an unknown *opportunity cost* related to the information lost for training the final model, a concept that standard regret analysis does not capture.

Our method is fundamentally a cost-sensitive active learning algorithm designed to build a minimally small, high-quality training set, which is a different goal from the online performance optimization of bandits.

B.1.3. MISMATCH 3: THE REWARD SIGNAL

A crucial difficulty in applying a bandit model is defining an immediate reward signal R_t that aligns with the long-term goal of learning a good router. If we choose an expert k for an example (x_t, y_t) , a natural reward might be its accuracy, $R_t = 1 - c_k(x_t, y_t)$. However, this myopic reward would merely encourage the bandit to identify the most accurate expert for each context.

The true “value” of a query in our setting is the *information gain* it provides for reducing the size of the version space $\mathcal{R}_t$. A query is valuable if it reveals high disagreement among the surviving hypotheses, even if the queried expert is not the most accurate one for that instance. This notion of value is internal to the state of the learning algorithm and cannot be expressed as a simple, external reward signal, making the problem fall outside the standard contextual bandit model.

In summary, our framework is tailored to the specific problem of training a deferral model under a budget, providing dual guarantees on generalization and label complexity that a contextual bandit formulation cannot naturally replicate.

Appendix C. Boundedness of Slope Asymmetry

This section shows that the slope asymmetry constant K_ℓ (Definition 2) is finite for the multinomial logistic (softmax) loss on compact logit domains, in the same spirit as the IWAL analysis (Beygelzimer et al., 2009) (see their Lemma 5 and Corollary 6). Two standard technicalities are made explicit here: (i) softmax invariance to constant shifts (handled by an identifiability constraint), and (ii) the existence of a zero-cost expert at each (x, y) (which prevents the denominator from vanishing).

Setup and identifiability. Let

$$\ell_{\log}(r, x, k) = -\log\left(\frac{e^{r(x,k)}}{\sum_{j=1}^{n_e} e^{r(x,j)}}\right) = -r(x, k) + \log \sum_{j=1}^{n_e} e^{r(x,j)}$$

be the multinomial logistic loss. Throughout, we assume the logit vector $r(x, \cdot)$ lies in a compact set and satisfies a standard *identifiability constraint*, e.g.,

$$r(x, \cdot) \in \mathcal{B}_B := \left\{v \in \mathbb{R}^{n_e} : \|v\|_\infty \leq B, \sum_{j=1}^{n_e} v_j = 0\right\}. \quad (5)$$

(Equivalently, one may fix a reference class score to zero; any fixed gauge that removes the softmax invariance to constant shifts is acceptable.)

Lemma 7 (Lipschitz upper bound) Fix x and k . For any $r(x, \cdot), r'(x, \cdot) \in \mathcal{B}_B$,

$$|\ell_{\log}(r, x, k) - \ell_{\log}(r', x, k)| \leq 2\|r(x, \cdot) - r'(x, \cdot)\|_\infty.$$

Consequently,

$$\sum_{k=1}^{n_e} |\ell_{\log}(r, x, k) - \ell_{\log}(r', x, k)| \leq 2n_e \|r(x, \cdot) - r'(x, \cdot)\|_\infty.$$

Proof The gradient with respect to the logit vector has components

$$\frac{\partial \ell_{\log}}{\partial r(x, k)} = -1 + s_k, \quad \frac{\partial \ell_{\log}}{\partial r(x, j)} = s_j (j \neq k),$$

where $s = \text{softmax}(r(x, \cdot)) \in \Delta^{n_e-1}$. Thus $\|\nabla_{r(x, \cdot)} \ell_{\log}\|_1 = |-1 + s_k| + \sum_{j \neq k} s_j = 2(1 - s_k) \leq 2$. The mean value inequality yields the first claim; summing over k gives the second. $\blacksquare$

Lemma 8 (Coercivity on the zero-mean subspace) *Let $s \in \Delta^{n_e-1}$ and define the linear map $M_s := I - \mathbf{1}s^\top$. For any $v \in \mathbb{R}^{n_e}$ with $\sum_j v_j = 0$,*

$$\|M_s v\|_\infty \geq \frac{1}{2} \|v\|_\infty.$$

Proof Write $(M_s v)_i = v_i - s^\top v$. Since $s^\top v$ is a convex combination of the coordinates of v , it lies in $[\min_j v_j, \max_j v_j]$. Because $\sum_j v_j = 0$, we have $\min_j v_j \leq 0 \leq \max_j v_j$, hence $\max_j v_j - \min_j v_j \geq \|v\|_\infty$. For any i , $|(M_s v)_i| = |v_i - s^\top v| \geq \frac{1}{2}(\max_j v_j - \min_j v_j) \geq \frac{1}{2} \|v\|_\infty$, so taking the maximum over i gives the claim. $\blacksquare$

Lemma 9 (Pointwise lower bound on per-class differences) *Fix x and consider any $r, r' \in \mathcal{B}_B$. Then there exists a softmax vector s on the line segment between $r(x, \cdot)$ and $r'(x, \cdot)$ such that*

$$\ell_{\log}(r', x, \cdot) - \ell_{\log}(r, x, \cdot) = -(I - \mathbf{1}s^\top)(r'(x, \cdot) - r(x, \cdot)).$$

In particular,

$$\max_k |\ell_{\log}(r', x, k) - \ell_{\log}(r, x, k)| \geq \frac{1}{2} \|r'(x, \cdot) - r(x, \cdot)\|_\infty.$$

Proof By the mean value theorem applied to the smooth map $r \mapsto \ell_{\log}(r, x, \cdot)$, there exists $s \in \Delta^{n_e-1}$ on the segment connecting r and r' with $\nabla \ell_{\log} = I - \mathbf{1}s^\top$ (row-wise). Since both $r, r' \in \mathcal{B}_B$ satisfy $\sum_j r_j = \sum_j r'_j = 0$, their difference $v = r' - r$ also satisfies $\sum_j v_j = 0$. The identity then gives $\Delta \ell = -M_s v$ and Lemma 8 yields $\|\Delta \ell\|_\infty \geq \frac{1}{2} \|v\|_\infty$. $\blacksquare$

Bounded slope asymmetry: non-budgeted case. Suppose all experts have zero cost at (x, y) (non-budgeted systems). Then, for any $r, r' \in \mathcal{B}_B$ and fixed x ,

$$\frac{\max_c \sum_k (1 - c_k) |\Delta \ell_k|}{\min_c \sum_k (1 - c_k) |\Delta \ell_k|} = \frac{\sum_{k=1}^{n_e} |\Delta \ell_k|}{\sum_{k=1}^{n_e} |\Delta \ell_k|} = 1,$$

and trivially $K_{\ell_{\log}} = 1$. More meaningfully, combining Lemma 7 and Lemma 9 gives

$$\frac{\sum_{k=1}^{n_e} |\Delta \ell_k|}{\max_k |\Delta \ell_k|} \leq \frac{2n_e \|\Delta r\|_\infty}{\frac{1}{2} \|\Delta r\|_\infty} = 4n_e. \quad (6)$$

Thus, on compact domains with the standard identifiability (5), the slope asymmetry is explicitly bounded by a constant depending only on n_e (and the compactness B through the domain restriction, not through the constant).

Bounded slope asymmetry: budgeted case with zero-cost experts. Assume at each (x, y) there exists a nonempty index set $I(x, y) \subseteq [n_e]$ of zero-cost experts, i.e., $c_k(x, y) = 0 \iff k \in I(x, y)$. Then

$$\min_{\mathbf{c}} \sum_k (1-c_k) |\Delta \ell_k| = \sum_{k \in I(x, y)} |\Delta \ell_k|, \quad \max_{\mathbf{c}} \sum_k (1-c_k) |\Delta \ell_k| = \sum_{k \notin I(x, y)} |\Delta \ell_k| + \sum_{k \in I(x, y)} |\Delta \ell_k| = \sum_{k=1}^{n_e} |\Delta \ell_k|.$$

Therefore,

$$K_{\ell_{\log}} \leq \frac{\sum_{k=1}^{n_e} |\Delta \ell_k|}{\sum_{k \in I(x, y)} |\Delta \ell_k|} \leq \frac{2n_e \|\Delta r\|_{\infty}}{\sum_{k \in I(x, y)} |\Delta \ell_k|}, \quad (7)$$

using Lemma 7 for the numerator.

A simple uniform bound follows if the zero-cost set has size at least a fixed fraction: suppose there exists $\rho \in (0, 1]$ such that $|I(x, y)| \geq \rho n_e$ for all (x, y) . Then by averaging, $\sum_{k \in I(x, y)} |\Delta \ell_k| \geq \frac{|I(x, y)|}{n_e} \sum_{k=1}^{n_e} |\Delta \ell_k| \geq \rho \max_k |\Delta \ell_k|$. Combining with Lemma 9 gives

$$K_{\ell_{\log}} \leq \frac{2n_e \|\Delta r\|_{\infty}}{\rho \max_k |\Delta \ell_k|} \leq \frac{2n_e \|\Delta r\|_{\infty}}{\rho \frac{1}{2} \|\Delta r\|_{\infty}} = \frac{4}{\rho} n_e. \quad (8)$$

Hence, in the budgeted case with a uniform lower bound on the number of zero-cost experts, $K_{\ell_{\log}}$ is explicitly bounded by $C(n_e, \rho) = \frac{4}{\rho} n_e$.

Remarks. (i) The identifiability constraint (5) (or any equivalent gauge fixing) is standard in multiclass logistic models and is necessary to avoid the softmax invariance to constant shifts, under which $\Delta \ell \equiv 0$ can occur for $\Delta r \neq 0$. (ii) The constants above are *dimension-explicit* and do not depend on the particular scores beyond the compactness parameter B . (iii) The same proof pattern applies to other comp-sum losses whose per-class losses are Lipschitz in the logits on compact domains (e.g., exponential), with constants depending only on n_e and the chosen gauge. (iv) When $I(x, y) = [n_e]$ (non-budgeted systems), the clean bound (6) shows $K_{\ell_{\log}} \leq 4n_e$; when $|I(x, y)| \geq \rho n_e$ uniformly, the budgeted bound (8) applies.

Appendix D. Proofs of Budgeted Two-Stage Deferral

D.1. Lemma 10 and Proof

Lemma 10 For any $\delta > 0$, with probability at least $1 - \delta$, for all time steps T and all pairs $(r, r') \in \mathcal{R}_T^2$, we have

$$|\mathcal{E}_T(r) - \mathcal{E}_T(r') - (\mathcal{E}(r) - \mathcal{E}(r'))| \leq \Delta_T.$$

Proof Fix any time step T and any pair $(r, r') \in \mathcal{R}_T^2$. Define the deviation sequence $Z_t, t \in [T]$:

$$Z_t = \sum_{k=1}^{n_e} \frac{1_{\{k_t=k\}} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(r, x_t, k) - \ell(r', x_t, k)) - (\mathcal{E}(r) - \mathcal{E}(r')).$$

Here, the selection probabilities $p_{t,k}$ depend only on randomness up to time $t-1$, and the parameters $q_{t,k}$ are fixed before the algorithm begins. Thus, $(Z_t)_t$ forms a martingale difference sequence with respect to the natural filtration, since $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$.

Next, we bound the absolute value of Z_t . Using the definition of $p_{t,k}$ and the boundedness of the loss, we have:

$$|Z_t| \leq \frac{1}{q_{\min} p_{t,k}} |\ell(r, x_t, k) - \ell(r', x_t, k)| + |\mathcal{E}(r) - \mathcal{E}(r')| \leq \frac{1}{q_{\min}} + 1 = \bar{q}.$$

Applying Azuma's inequality (Mohri et al., 2018, Theorem D.7) with failure probability $\delta/(T(T+1)|\mathcal{R}|^2)$ gives the desired bound for a fixed T and pair (r, r') . Taking a union bound over all $t \in [T]$ and all pairs $(r, r') \in \mathcal{R}_T^2$ yields the result. $\blacksquare$

D.2. Proof of Theorem 1

Theorem 1 (Two-Stage Generalization Bound)

Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^{n_e}$, and let $\mathcal{R}$ be a hypothesis class. Assume that $r^* \in \mathcal{R}$ minimizes the expected surrogate loss $\mathcal{E}(r)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $T \geq 1$:

- The optimal hypothesis r^* belongs to the retained set $\mathcal{R}_T$;
- For all $r, r' \in \mathcal{R}_T$, the generalization gap satisfies

$$\mathcal{E}(r) - \mathcal{E}(r') \leq 2\Delta_{T-1}.$$

In particular, the learned hypothesis r_T at time T satisfies

$$\mathcal{E}(r_T) - \mathcal{E}(r^*) \leq 2\Delta_{T-1}.$$

Proof We prove by induction that $r^* \in \mathcal{R}_T$ for all T . The base case $T = 1$ holds trivially since $\mathcal{R}_1 = \mathcal{R}$. Suppose the claim holds for T .

Let r_T be the empirical risk minimizer in $\mathcal{R}_T$. Then, by Lemma 10:

$$\mathcal{E}_T(r^*) - \mathcal{E}_T(r_T) \leq \mathcal{E}(r^*) - \mathcal{E}(r_T) + \Delta_T \leq \Delta_T.$$

Thus $\mathcal{E}_T(r^*) \leq \mathcal{E}_T^* + \Delta_T$, implying $r^* \in \mathcal{R}_{T+1}$.

For the second part, let $r, r' \in \mathcal{R}_T$. Then:

$$\mathcal{E}(r) - \mathcal{E}(r') \leq \mathcal{E}_{T-1}(r) - \mathcal{E}_{T-1}(r') + \Delta_{T-1} \leq \mathcal{E}_{T-1}^* + \Delta_{T-1} - \mathcal{E}_{T-1}^* + \Delta_{T-1} \leq 2\Delta_{T-1}.$$

Applying this to r_T and r^* completes the proof. $\blacksquare$

D.3. Proof of Lemma 4

Lemma 11 For any distribution $\mathcal{D}$ and any multi-class loss function ℓ , we have $\rho(r, r^*) \leq K_\ell \cdot (\mathcal{E}(r) + \mathcal{E}^*)$ for all $r \in \mathcal{R}$.

Proof Expanding the definition of ρ and applying the triangle inequality:

$$\begin{aligned}
 \rho(r, r^*) &= \mathbb{E}_{(x,y) \sim D} \left[\max_{c \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(r, x, k) - \ell(r^*, x, k)| \right] \\
 &\leq K_\ell \mathbb{E}_{(x,y) \sim D} \left[\sum_{k=1}^{n_e} (1 - c_k) |\ell(r, x, k) - \ell(r^*, x, k)| \right] \\
 &\leq K_\ell \left(\mathbb{E}_{(x,y) \sim D} \left[\sum_{k=1}^{n_e} (1 - c_k) \ell(r, x, k) \right] + \mathbb{E}_{(x,y) \sim D} \left[\sum_{k=1}^{n_e} (1 - c_k) \ell(r^*, x, k) \right] \right) \\
 &= K_\ell (\mathcal{E}(r) + \mathcal{E}(r^*)).
 \end{aligned}$$

This completes the proof. ■

D.4. Proof of Theorem 6

Theorem 6 (Two-Stage Label Complexity Bound) *Let $\mathcal{D}$ be a two-stage deferral distribution and $\mathcal{R}$ a hypothesis set. Suppose the loss function ℓ has slope asymmetry K_ℓ and the disagreement coefficient of the problem is θ . Then, with probability at least $1 - \delta$, the expected number of cost queries made by the budgeted two-stage deferral algorithm over T rounds is bounded by.*

$$4\theta \cdot K_\ell \cdot \left(\mathcal{E}^* T + O\left((1/q_{\min} + 1)\sqrt{T \log(|\mathcal{R}|T/\delta)}\right) \right), \quad (3)$$

where the expectation is taken over the algorithm's randomness.

Proof Let r^* be the best-in-class minimizer. At time t , from Theorem 1, we have

$$\mathcal{R}_t \subset \{r \in \mathcal{R} : \mathcal{E}(r) \leq \mathcal{E}^* + 2\Delta_{t-1}\}.$$

By Lemma 4, this implies $\mathcal{R}_t \subset B(r^*, \epsilon)$ with $\epsilon = K_\ell(2\mathcal{E}^* + 2\Delta_{t-1})$. The expected number of cost queries at round t is:

$$\begin{aligned}
 \mathbb{E} \left[\sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k} \right] &= \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} p_{t,k} \right] \\
 &= \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} \max_{r, r' \in \mathcal{R}_t} |\ell(r, x, k) - \ell(r', x, k)| \right] \\
 &\leq 2 \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} \max_{r \in \mathcal{R}_t} |\ell(r, x, k) - \ell(r^*, x, k)| \right] \\
 &\leq 2 \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} \max_{r \in B(r^*, \epsilon)} |\ell(r, x, k) - \ell(r^*, x, k)| \right] \\
 &\leq 2\theta\epsilon \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} \right] \\
 &= 4\theta K_\ell (\mathcal{E}^* + \Delta_{t-1})
 \end{aligned}$$

Summing over $t = 1$ to T gives the claimed result. ■

Appendix E. Budgeted Single-Stage Multiple-Expert Deferral

We now consider a budgeted single-stage deferral setting, where the learner can either predict a label directly or defer to one of n_e predefined experts. The decision space is expanded to $\bar{\mathcal{Y}} = [n + n_e]$, with the original label space $\mathcal{Y} = [n]$ and deferral actions labeled $n + 1$ through $n + n_e$.

Similar to the two-stage setting, we decompose the decision into two parts: (1) deciding whether to defer and, if so, selecting an expert k and (2) determining the probability of querying $c_k(x_t, y_t)$ once k is chosen. For the query probability $p_{t,k}$, we carefully design its expression based on the surrogate loss scores for expert k computed by each routing function in the current version space, along with the maximum cost in next section. The distinction from the two-stage setting is that the single-stage setting integrates an additional "no deferral" option, characterized by its own query probability, $q_{t,0}$. This formulation allows us to derive strong theoretical guarantees on the label complexity.

E.1. Deferral and Surrogate Loss

Each hypothesis $h \in \mathcal{H}$ maps $(x, \bar{y}) \in \mathcal{X} \times \bar{\mathcal{Y}}$ to a real-valued score, with predictions made via $\hat{h}(x) = \arg\max_{\bar{y}} h(x, \bar{y})$. The *single-stage deferral loss* is:

$$L_{\text{def}}(h, x, y, \mathbf{c}) = \mathbb{1}_{h(x) \neq y} \mathbb{1}_{h(x) \in [n]} + \sum_{k=1}^{n_e} c_k(x, y) \mathbb{1}_{h(x) = n+k},$$

where $c_k(x, y)$ reflects the cost of deferring to expert k , typically defined as the expert's misclassification error (Verma et al., 2023), as in the two-stage setting. Direct minimization is intractable, thus we will use a surrogate loss proposed by Verma et al. (2023); Mao et al. (2024a):

$$L(h, x, y, \mathbf{c}) = \ell(h, x, y) + \sum_{k=1}^{n_e} (1 - c_k(x, y)) \ell(h, x, n+k),$$

where ℓ is multi-class surrogate loss (e.g., logistic). As before, we assume $L \in [0, 1]$ and define the generalization error $\mathcal{E}(h) = \mathbb{E}[L(h, x, y, \mathbf{c})]$.

E.2. Algorithm Overview

Algorithm 3 outlines the procedure. As with the two-stage setting, for any $t \in [T]$ and $k \in [0, n_e]$, we denote by $q_{t,k}$ the probability of selecting expert k at time t when $k \neq 0$ and that of making a direct prediction when $k = 0$. The optimal choice for the value of $q_t = (q_{t,0}, q_{t,1}, \dots, q_{t,n_e})$ is determined in Appendix E.4, using our theoretical bounds.

If deferring, a Bernoulli trial with success probability $p_{t,k}$, returned by *Sampling-Probs*, determines whether the cost $c_{t,k}$ is queried. Queried examples are then stored with corresponding importance weights.

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^{n_e}$. The generalization error of $h \in \mathcal{H}$ on $\mathcal{D}$ is given by $\mathcal{E}(h) = \mathbb{E}_{(x,y,\mathbf{c}) \sim \mathcal{D}}[L(h, x, y, \mathbf{c})] = \mathbb{E}_{(x,y,\mathbf{c}) \sim \mathcal{D}}[\ell(h, x, y) + \sum_{k=1}^{n_e} (1 - c_k(x, y)) \ell(h, x, n+k)]$. The importance weighted estimate of the generalization error of at time T is

$$\mathcal{E}_T(r) = \frac{1}{T} \sum_{t=1}^T \left(\frac{1_{k_t=0}}{q_{t,0}} \ell(h, x_t, y_t) + \sum_{k=1}^{n_e} \frac{1_{k_t=k} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) \ell(h, x_t, n+k) \right)$$

where $(k_t, Q_{t,k})$ is as defined in the algorithm. It is straightforward to see that $\mathbb{E}[\mathcal{E}_T(r)] = \mathcal{E}(r)$, with the expectation taken over all the random variables involved.

Algorithm 3 Budgeted Single-Stage Deferral with Multiple Experts (Subroutine SAMPLING-PROBS)

```

INITIALIZE  $S_0 \leftarrow \emptyset$ ;
for  $t = 1$  to  $T$  do
  RECEIVE( $x_t, y_t$ );
   $p_t \leftarrow \text{SAMPLING-PROBS}(x_t, y_t, \{x_s, y_s, \mathbf{c}_s, q_s, k_s, p_s, Q_s: 1 \leq s < t\})$ ;
   $k_t \leftarrow \text{SAMPLE}(n_e + 1, q_{t,k})$ ;
  if  $k_t = 0$  then
     $S_t \leftarrow S_{t-1} \cup \left\{ \left( x_t, y_t, 0, \frac{1}{q_{t,k_t}} \right) \right\}$ ;
  else
     $Q_{t,k_t} \leftarrow \text{BERNOULLI}(p_{t,k_t})$ ;
  end if
  if  $Q_{t,k_t} = 1$  then
     $c_{t,k_t} \leftarrow \text{QUERY-COST}(k_t, (x_t, y_t))$ 
     $S_t \leftarrow S_{t-1} \cup \left\{ \left( x_t, y_t, c_{t,k_t}, \frac{1}{q_{t,k_t} p_{t,k_t}} \right) \right\}$ ;
  else
     $S_t \leftarrow S_{t-1}$ ;
  end if
   $r_t \leftarrow \operatorname{argmin}_{r \in \mathcal{R}} \sum_{(x,y,c,w) \in S_t} w(1-c)\ell(r, x, n+k_t)1_{k_t \neq 0} + w \cdot (1-c) \cdot \ell(r, x, y)1_{k_t=0}$ .
end for

```

E.3. Sampling-PROBS Strategy and Generalization Guarantees

We apply a shrinking version space strategy (Algorithm 4) similar to the two-stage case. Let $\mathcal{H}_t$ be the version space at time t . Using standard uniform convergence arguments, we prune $\mathcal{H}_t$ to keep hypotheses whose empirical loss is within Δ_t of the minimum $\mathcal{E}_t^*$:

$$\mathcal{H}_{t+1} = \{h \in \mathcal{H}_t : \mathcal{E}_t(r) \leq \mathcal{E}_t^* + \Delta_t\}. \quad (9)$$

where $\Delta_t = \sqrt{\bar{q}^2 (8/t) \log(2t(t+1)|\mathcal{H}|^2/\delta)}$ and $\bar{q} = \frac{1}{q_{\min}} + 1$ with $q_{\min} = \min_{k \in 0 \cup [n_e]} q_{t,k} > 0$.

Algorithm 4 Sampling-PROBS Subroutine with Past History

```

INITIALIZE  $\mathcal{H}_0 \leftarrow \mathcal{H}$ ;
for  $t = 2$  to  $T$  do
   $\mathcal{E}_{t-1}(h) \leftarrow \frac{1}{t-1} \sum_{s=1}^{t-1} \left( \frac{1_{k_s=0}}{q_{s,0}} \ell(h, x_s, y_s) + \sum_{k=1}^{n_e} \frac{1_{k_s=k} Q_{s,k}}{q_{s,k} p_{s,k}} (1 - c_{s,k}(x_s, y_s)) \ell(h, x_s, n+k) \right)$ ;
   $\mathcal{E}_{t-1}^* \leftarrow \min_{h \in \mathcal{H}_{t-1}} \mathcal{E}_{t-1}(h)$ ;
   $\mathcal{H}_t \leftarrow \{h \in \mathcal{H}_{t-1} : \mathcal{E}_{t-1}(h) \leq \mathcal{E}_{t-1}^* + \Delta_{t-1}\}$ ;
   $p_{t,k} \leftarrow \max_{h, h' \in \mathcal{H}_t} (\ell(h, x_t, n+k) - \ell(h', x_t, n+k))$ .
end for

```

The sampling probability $p_{t,k}$ is based on the variability of the expert-specific component of the surrogate loss:

$$p_{t,k} = \max_{h, h' \in \mathcal{H}_t} (\ell(h, x_t, n+k) - \ell(h', x_t, n+k)).$$

Leveraging the decomposability of the surrogate loss, this design allocates the query budget adaptively, prioritizing experts and instances where the disagreement among remaining hypotheses is greatest, thus targeting high-uncertainty regions. We now establish high-probability performance guarantees for the predictors output by the algorithm.

Theorem 12 (Single-Stage Generalization Bound) *Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y} \times 0, 1^{n_e}$, and let $\mathcal{H}$ be a hypothesis set. Suppose $h^* \in \mathcal{R}$ minimizes the expected surrogate loss $\mathcal{E}(h)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $T \geq 1$:*

- The optimal hypothesis h^* remains in $\mathcal{H}_T$.
- For any $h, h' \in \mathcal{H}_T$, the difference in generalization error satisfies $\mathcal{E}(h) - \mathcal{E}(h') \leq 2\Delta_{T-1}$.

In particular, the learned hypothesis h_T at time T satisfies: $\mathcal{E}(h_T) - \mathcal{E}(h^*) \leq 2\Delta_{T-1}$.

We need the following lemma for the proof.

Lemma 13 *For all data distributions $\mathcal{D}$, for all hypothesis sets $\mathcal{H}$, for all $\delta > 0$, with probability at least $1 - \delta$, for all T and all $h, h' \in \mathcal{H}_T$,*

$$|\mathcal{E}_T(h) - \mathcal{E}_T(h') - \mathcal{E}(h) + \mathcal{E}(h')| \leq \Delta_T.$$

Proof Fix any time step T and any pair $(h, h') \in \mathcal{H}_T^2$. Define the deviation sequence Z_t , $t \in [T]$:

$$\begin{aligned} Z_t &= \frac{1_{k_t=0}}{q_{t,0}} (\ell(h, x_t, y_t) - \ell(h', x_t, y_t)) \\ &\quad + \sum_{k=1}^{n_e} \frac{1_{k_t=k} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(h, x_t, n+k) - \ell(h', x_t, n+k)) - (\mathcal{E}(h) - \mathcal{E}(h')). \end{aligned}$$

Here, the selection probabilities $p_{t,k}$ depend only on randomness up to time $t-1$, and the parameters $q_{t,k}$ are fixed before the algorithm begins. Thus, $(Z_t)_t$ forms a martingale difference sequence with respect to the natural filtration, since $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$. Next, we bound the absolute value of Z_t . Using the definition of $p_{t,k}$ and the boundedness of the loss, we have:

$$\begin{aligned} |Z_t| &\leq \frac{1}{q_{\min}} |\ell(h, x_t, y_t) - \ell(h', x_t, y_t)| 1_{k=0} \\ &\quad + \frac{1}{q_{\min} p_{t,k}} |\ell(h, x_t, n+k) - \ell(h', x_t, n+k)| 1_{k \neq 0} + |\mathcal{E}(h) - \mathcal{E}(h')| \\ &\leq \frac{1}{q_{\min}} + 1 = \bar{q}. \end{aligned}$$

Applying Azuma's inequality (Mohri et al., 2018, Theorem D.7) with failure probability $\delta/(T(T+1)|\mathcal{H}|^2)$, we have

$$\begin{aligned} &\mathbb{P}[|\mathcal{E}_T(h) - \mathcal{E}_T(h') - \mathcal{E}(h) + \mathcal{E}(h')| \geq \Delta_T] \\ &= \mathbb{P}\left[\left|\frac{1}{T} \sum_{t=1}^T \left(\frac{1_{k_t=0}}{q_{t,0}} (\ell(h, x_t, y_t) - \ell(h', x_t, y_t)) \right. \right. \right. \\ &\quad \left. \left. + \sum_{k=1}^{n_e} \frac{1_{k_t=k} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(h, x_t, n+k) - \ell(h', x_t, n+k)) - (\mathcal{E}(h) - \mathcal{E}(h')) \right)\right| \geq \Delta_T\right] \\ &= \mathbb{P}\left[\left|\sum_{t=1}^T Z_t\right| \geq T\Delta_T\right] \leq 2e^{-\frac{T\Delta_T^2}{2\bar{q}^2}} = \frac{\delta}{T(T+1)|\mathcal{H}|^2}. \end{aligned}$$

Since $\mathcal{H}_T$ is a random subset of $\mathcal{H}$, it suffices to take a union bound over all $h, h' \in \mathcal{H}$, and T . A union bound over T finishes the proof. $\blacksquare$

Proof [Proof of Theorem 12] We prove by induction that $h^* \in \mathcal{H}_T$ for all T . The base case $T = 1$ holds trivially since $\mathcal{H}_1 = \mathcal{H}$. Suppose the claim holds for T .

Let h_T be the empirical risk minimizer in $\mathcal{H}_T$. Then, by Lemma 10:

$$\mathcal{E}_T(h^*) - \mathcal{E}_T(h_T) \leq \mathcal{E}(h^*) - \mathcal{E}(h_T) + \Delta_T \leq \Delta_T.$$

Thus $\mathcal{E}_T(h^*) \leq \mathcal{E}_T^* + \Delta_T$, implying $h^* \in \mathcal{H}_{T+1}$.

For the second part, let $h, h' \in \mathcal{H}_T$. Then:

$$\mathcal{E}(h) - \mathcal{E}(h') \leq \mathcal{E}_{T-1}(h) - \mathcal{E}_{T-1}(h') + \Delta_{T-1} \leq \mathcal{E}_{T-1}^* + \Delta_{T-1} - \mathcal{E}_{T-1}^* + \Delta_{T-1} = 2\Delta_{T-1}.$$

Applying this to h_T and h^* completes the proof. $\blacksquare$

E.4. Label Complexity

We showed that the generalization error of the classifier output by budgeted single-stage deferral (Sampling-Probs) is similar to the generalization error of the classifier chosen passively after seeing all T labels. How many of those T labels does the active learner request?

To derive label complexity guarantees for our algorithm, we must adapt existing tools and definition in active learning to our single-stage deferral setting. In particular, we will define a new notion of slope asymmetry, hypothesis distance metric, generalized disagreement coefficient, based on experts' costs c_k and tailored to our setting. These tools allow us to demonstrate that our budgeted single-stage deferral algorithm can achieve a favorable label complexity, in fact lower than its fully supervised counterpart when the learning problem is approximately realizable and the disagreement coefficient of the hypothesis set is not too large.

We give label complexity upper bounds for a class of multi-class surrogate loss functions that includes multinomial logistic loss and a class of cost functions that satisfy natural assumptions. We require that the loss function has bounded *slope asymmetry*, defined below.

Definition 14 (Slope Asymmetry for Single-Stage Deferral) *The slope asymmetry of a multi-class loss function $\ell : \mathcal{H} \times \mathcal{X} \times [n + n_e] \rightarrow [0, \infty)$ is $K_\ell =$*

$$\sup_{h, h' \in \mathcal{H}, x \in \mathcal{X}, y \in \mathcal{Y}} \frac{|\ell(h, x, y) - \ell(h', x, y)| + \max_{\mathbf{c} \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(h, x, n+k) - \ell(h', x, n+k)|}{|\ell(h, x, y) - \ell(h', x, y)| + \min_{\mathbf{c} \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(h, x, n+k) - \ell(h', x, n+k)|}.$$

As in the two-stage case, this quantity is always well-defined and finite if, for every (x, y) , there exists at least one expert k^* with zero cost: $c_{k^*}(x, y) = 0$. In practice, K_ℓ is bounded for common convex surrogates such as the logistic loss, provided the range of score functions r is restricted to a compact interval (e.g., $[-B, B]$). Next, we define a distance measure over the hypothesis set that reflects variability in expert-specific loss components.

Definition 15 (Single-Stage Hypothesis Distance Metric) *For any $h, h' \in \mathcal{H}$, define $\rho(h, h') =$*

$$\mathbb{E}_{(x,y) \sim D} \left[|\ell(h, x, y) - \ell(h', x, y)| + \max_{\mathbf{c} \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(h, x, n+k) - \ell(h', x, n+k)| \right].$$

We define the ϵ -ball around h as $B(h, \epsilon) = \{h' \in \mathcal{H} : \rho(h, h') \leq \epsilon\}$.

Suppose $h^* \in \mathcal{H}$ minimizes the expected surrogate loss: $\mathcal{E}^* = \mathcal{E}(h^*) = \inf_{h \in \mathcal{H}} \mathcal{E}(h)$. At time t , the version space $\mathcal{H}_t$ contains only hypotheses with generalization error at most $\mathcal{E}^* + 2\Delta_{t-1}$. But how close are these hypotheses to h^* in ρ -distance? The following lemma provides an upper bound in terms of the slope asymmetry:

Lemma 16 *For any distribution $\mathcal{D}$ and any multi-class loss function ℓ , we have $\rho(h, h^*) \leq K_\ell \cdot (\mathcal{E}(h) + \mathcal{E}^*)$ for all $h \in \mathcal{H}$.*

Proof Expanding the definition of ρ and applying the triangle inequality:

$$\begin{aligned}
 \rho(h, h^*) &= \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[|\ell(h, x, y) - \ell(h^*, x, y)| + \max_{\mathbf{c} \in \{0,1\}^{n_e}} \sum_{k=1}^{n_e} (1 - c_k) |\ell(h, x, n+k) - \ell(h^*, x, n+k)| \right] \\
 &\leq K_\ell \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[|\ell(h, x, y) - \ell(h^*, x, y)| + \sum_{k=1}^{n_e} (1 - c_k) |\ell(h, x, n+k) - \ell(h^*, x, n+k)| \right] \\
 &\leq K_\ell \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell(h, x, y) + \sum_{k=1}^{n_e} (1 - c_k) \ell(h, x, n+k) \right] \right. \\
 &\quad \left. + \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell(h^*, x, y) + \sum_{k=1}^{n_e} (1 - c_k) \ell(h^*, x, n+k) \right] \right) \\
 &= K_\ell (\mathcal{E}(h) + \mathcal{E}(h^*)).
 \end{aligned}$$

This completes the proof. ■

The following extends the notion of disagreement (Hanneke, 2007) to the single-stage deferral setting.

Definition 17 *The disagreement coefficient θ is the smallest value such that, for all $\epsilon > 0$,*

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} \sup_{h \in B(h^*, \epsilon)} \sup_{k \in [n_e]} |\ell(h, x, n+k) - \ell(h^*, x, n+k)| \leq \theta \epsilon.$$

We now present an upper bound on the expected number of cost queries required by the algorithm.

Theorem 18 (Single-Stage Label Complexity Bound) *Let $\mathcal{D}$ be a single-stage deferral distribution and $\mathcal{H}$ a hypothesis set. Suppose the loss function ℓ has slope asymmetry K_ℓ and the disagreement coefficient of the problem is θ . Then, with probability at least $1 - \delta$, the expected number of cost queries made by the budgeted single-stage deferral algorithm over T rounds is bounded by:*

$$4\theta \cdot K_\ell \cdot \left(\mathcal{E}^* T + O\left((1/q_{\min} + 1) \sqrt{T \log(|\mathcal{H}|T/\delta)} \right) \right), \quad (10)$$

where the expectation is taken over the algorithm's randomness.

Proof Let h^* be the best-in-class minimizer. At time t , from Theorem 12, we have

$$\mathcal{H}_t \subset \{h \in \mathcal{H} : \mathcal{E}(h) \leq \mathcal{E}^* + 2\Delta_{t-1}\}.$$

By Lemma 16, this implies $\mathcal{H}_t \subset B(h^*, \epsilon)$ with $\epsilon = K_\ell(2\mathcal{E}^* + 2\Delta_{t-1})$. The expected number of cost queries at round t is:

$$\begin{aligned}
 \mathbb{E}\left[\sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k}\right] &= \mathbb{E}\left[\sum_{k=1}^{n_e} q_{t,k} p_{t,k}\right] \\
 &= \mathbb{E}\left[\sum_{k=1}^{n_e} q_{t,k} \max_{h, h' \in \mathcal{H}_t} |\ell(h, x, n+k) - \ell(h', x, n+k)|\right] \\
 &\leq 2 \mathbb{E}\left[\sum_{k=1}^{n_e} q_{t,k} \max_{h \in \mathcal{H}_t} |\ell(h, x, n+k) - \ell(h^*, x, n+k)|\right] \\
 &\leq 2 \mathbb{E}\left[\sum_{k=1}^{n_e} q_{t,k} \max_{h \in B(h^*, \epsilon)} |\ell(h, x, n+k) - \ell(h^*, x, n+k)|\right] \\
 &\leq 2\theta\epsilon \mathbb{E}\left[\sum_{k=1}^{n_e} q_{t,k}\right] \\
 &\leq 4\theta K_\ell(\mathcal{E}^* + \Delta_{t-1})
 \end{aligned}$$

Summing over $t = 1$ to T gives the claimed result. $\blacksquare$

The theorem establishes a label complexity bound for our budgeted single-stage deferral algorithm. As in the two-stage setting, in the realizable case, the bound scales as $\tilde{O}(\sqrt{T})$, significantly improving over the linear label complexity $n_e T$ incurred by standard two-stage methods that query all expert costs and, even in the agnostic setting, the bound remains favorable when the optimal surrogate loss $\mathcal{E}^*$ is small. For simplicity of exposition, the main body presents this square-root bound, while the stronger logarithmic bound (derived via Freedman’s inequality) is given in full detail in Appendix F.

The dependence on the generalized disagreement coefficient is, in general, unavoidable, as shown by Hanneke (2014). However, this coefficient has been shown to be bounded for many common hypothesis classes, enabling meaningful guarantees in practice.

Optimal q_t Finally, we note that both the generalization and label complexity bounds are minimized when the expert sampling probabilities are uniform: $q_{t,0} = q_{t,1} = \dots = q_{t,n_e} = \frac{1}{n_e+1}$, for all t , yielding $q_{\min} = 1/(n_e + 1)$ since all experts are treated symmetrically. Under this setting, the bounds simplify to:

$$\begin{aligned}
 \mathcal{E}(r_T) &\leq \mathcal{E}(r^*) + 2(n_e + 2)\sqrt{(8/(T-1)) \log(2(T-1)T|\mathcal{H}|^2/\delta)} \\
 \mathbb{E}\left[\sum_{t=1}^T \sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k}\right] &\leq 4\theta \cdot K_\ell \cdot \left(\mathcal{E}^* T + O\left((n_e + 2)\sqrt{T \log(|\mathcal{H}|T/\delta)}\right)\right), \tag{11}
 \end{aligned}$$

both of which depend on the number of experts. Note that the label complexity is significantly more favorable than in the standard deferral setting, where it is $n_e T$.

By leveraging Freedman’s inequality (Freedman, 1975) in place of Azuma’s (Mohri et al., 2018, Theorem D.7) in our analysis, we can in fact derive learning and sample complexity bounds that depend only logarithmically on T in the realizable case (see Appendix F). This significantly strengthens our theoretical guarantees and further highlights the advantages of our approach.

E.5. Practical Implementation

As in the two-stage setting, our single-stage algorithm admits an efficient implementation in common scenarios, particularly when the hypothesis set $\mathcal{H}$ is convex and the surrogate loss is a convex comp-sum loss (e.g., logistic loss). Each round then reduces to solving convex programs for empirical risk minimization and for estimating sampling probabilities, both of which are tractable using standard optimization solvers. The approach also extends to more expressive models such as neural networks, though at the cost of non-convex optimization.

Comp-sum losses. We consider *comp-sum losses* (Mao et al., 2023f) as the multi-class surrogate loss family ℓ , which includes many popular losses such as the multinomial logistic loss. A comp-sum loss is defined for any $(h, x, \bar{y}) \in \mathcal{H} \times \mathcal{X} \times [n + n_e]$ as

$$\ell_{\text{comp}}(h, x, \bar{y}) = \Psi\left(\frac{e^{h(x, \bar{y})}}{\sum_{y' \in [n+n_e]} e^{h(x, y')}}\right),$$

where $\Psi: [0, 1] \rightarrow \mathbb{R}_+ \cup \{+\infty\}$ is non-increasing. A notable instance is $\Psi(u) = -\log u$, which yields the *multinomial logistic loss* (Verhulst, 1838, 1845; Berkson, 1944, 1951). For suitable choices of Ψ , ℓ_{comp} is convex in h .

Convex feasible region. At each round t , Algorithm 4 requires solving two optimization problems over the restricted hypothesis set $\mathcal{H}_t$, defined as the intersection of convex constraints accumulated up to round t :

$$\mathcal{H}_t = \bigcap_{t' < t} \left\{ h \in \mathcal{H}: \frac{1}{t'} \sum_{i=1}^{t'} \left(\frac{1_{k_i=0}}{q_{i,0}} \ell_{\text{comp}}(h, x_i, y_i) + \sum_{k=1}^{n_e} \frac{1_{k_i=k} Q_{i,k}}{q_{i,k} p_{i,k}} (1 - c_{i,k}(x_i, y_i)) \ell_{\text{comp}}(h, x_i, n+k) \right) \leq \mathcal{E}_{t'}^* + \Delta_{t'} \right\}.$$

Since ℓ_{comp} is convex in h , each constraint defines a convex set, and thus $\mathcal{H}_t$ is convex.

First optimization. The first optimization at round T computes the minimal empirical loss:

$$\mathcal{E}_t^* = \min_{h \in \mathcal{H}_t} \frac{1}{t} \sum_{i=1}^t \left(\frac{1_{k_i=0}}{q_{i,0}} \Psi\left(\frac{e^{h(x_i, y_i)}}{\sum_{\bar{y} \in [n+n_e]} e^{h(x_i, \bar{y})}}\right) + \sum_{k=1}^{n_e} \frac{1_{k_i=k} Q_{i,k}}{q_{i,k} p_{i,k}} (1 - c_{i,k}(x_i, y_i)) \Psi\left(\frac{e^{h(x_i, n+k)}}{\sum_{\bar{y} \in [n+n_e]} e^{h(x_i, \bar{y})}}\right) \right),$$

a convex program in h over the feasible region $\mathcal{H}_t$.

Second optimization. The second optimization determines the sampling probability $p_{t,k}$ for each expert k by maximizing the difference in surrogate losses:

$$\max_{h, h' \in \mathcal{H}_t} \left\{ \Psi\left(\frac{e^{h(x, n+k)}}{\sum_{\bar{y} \in [n+n_e]} e^{h(x, \bar{y})}}\right) - \Psi\left(\frac{e^{h'(x, n+k)}}{\sum_{\bar{y} \in [n+n_e]} e^{h'(x, \bar{y})}}\right) \right\}.$$

Since Ψ is non-increasing, this expression is maximized when one term is minimized and the other maximized. Define

$$S_{\min}(x, k) \equiv \min_{h \in \mathcal{H}_t} \frac{e^{h(x, n+k)}}{\sum_{\bar{y} \in [n+n_e]} e^{h(x, \bar{y})}},$$

$$S_{\max}(x, k) \equiv \max_{h \in \mathcal{H}_t} \frac{e^{h(x, n+k)}}{\sum_{\bar{y} \in [n+n_e]} e^{h(x, \bar{y})}}.$$

Then the optimal loss variation equals $\Psi(S_{\min}(x, k)) - \Psi(S_{\max}(x, k))$.

Interpretation and extensions. This shows that for convex comp-sum losses and convex hypothesis sets $\mathcal{H}$, both optimization problems required by Algorithm 4 are convex and can be solved efficiently. As an example, when $\mathcal{H}$ is the linear class $h(x, \bar{y}) = \mathbf{w} \cdot \Phi(x, \bar{y})$ with $\|\mathbf{w}\| \leq B$, the optimizations reduce to convex programs in $\mathbf{w}$. For neural networks, the problems are no longer convex but can be tackled in practice with standard stochastic gradient descent (SGD).

Heuristics. As in the two-stage case (see also IWAL (Cortes et al., 2019b)), practical heuristics can be used to simplify implementation. For the first optimization, one can minimize over $\mathcal{H}$ instead of $\mathcal{H}_t$, and for the second optimization, it suffices to impose only the most recent constraint (from round $t - 1$) rather than all past constraints. With these choices, the optimal solution remains in the feasible set, while computational cost is reduced.

Appendix F. Improved Sample Complexity and Label Complexity Bounds

By leveraging Freedman’s inequality (Freedman, 1975) in place of Azuma’s (Mohri et al., 2018, Theorem D.7) in our analysis, we can in fact derive learning and sample complexity bounds that depend only logarithmically on T in the realizable case. Our derivation primarily follows the approach of Cortes et al. (2019b). For simplicity, we present these results for the two-stage setting; the results and proofs for the single-stage setting are essentially analogous.

Fix any time step T and any pair $(r, r') \in \mathcal{R}_T^2$. Define the deviation sequence $Z_t, t \in [T]$:

$$Z_t = \sum_{k=1}^{n_e} \frac{1_{\{k_t=k\}} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(r, x_t, k) - \ell(r', x_t, k)) - (\mathcal{E}(r) - \mathcal{E}(r')).$$

Let $q_{\min} = \min_{k \in [n_e]} q_{t,k} > 0$. The following result is adapted from (Kakade and Tewari, 2008, Lemma 3), which is derived from (Freedman, 1975).

Lemma 19 *For any $\delta > 0$, with probability at least $1 - \delta$, for all time steps $T \geq 3$ and all pairs $(r, r') \in \mathcal{R}_T^2$, we have*

$$\left| \sum_{t=1}^T Z_t \right| \leq \max \left\{ 2 \sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} \mathbb{E} \left[\frac{p_{t,k}}{q_{\min}^2} \mid \mathcal{F}_{t-1} \right]}, 6 \sqrt{\log \left(\frac{8 \log(T)}{\delta} \right)} \right\} \times \sqrt{\log \left(\frac{8 \log(T)}{\delta} \right)}.$$

Proof We apply [Kakade and Tewari \(2008, Lemma 3\)](#) and the fact that $|Z_t| \leq \bar{q}$. Furthermore,

$$\begin{aligned}
\text{var}[Z_t \mid \mathcal{F}_{t-1}] &= \text{var}\left[\sum_{k=1}^{n_e} \frac{1_{\{k_t=k\}} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(r, x_t, k) - \ell(r', x_t, k)) \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{k=1}^{n_e} \text{var}\left[\frac{1_{\{k_t=k\}} Q_{t,k}}{q_{t,k} p_{t,k}} (1 - c_{t,k}(x_t, y_t)) (\ell(r, x_t, k) - \ell(r', x_t, k)) \mid \mathcal{F}_{t-1}\right] \\
&\leq \sum_{k=1}^{n_e} \mathbb{E}\left[\frac{Q_{t,k}^2}{q_{t,k}^2 p_{t,k}^2} (1 - c_{t,k}(x_t, y_t))^2 (\ell(r, x_t, k) - \ell(r', x_t, k))^2 \mid \mathcal{F}_{t-1}\right] \\
&\leq \sum_{k=1}^{n_e} \mathbb{E}\left[\frac{Q_{t,k} p_{t,k}^2}{q_{\min}^2 p_{t,k}^2} \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{k=1}^{n_e} \mathbb{E}\left[\frac{Q_{t,k}}{q_{\min}^2} \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{k=1}^{n_e} \mathbb{E}\left[\frac{p_{t,k}}{q_{\min}^2} \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

A union bound over Z_t and $-Z_t$ concludes the proof. $\blacksquare$

Given Lemma 19 above, we can adapt ([Beygelzimer et al., 2009, Lemma 3](#)) to using the Bernstein-like inequality. Specifically, let us adopt the following threshold:

$$\Delta_T = \frac{2}{T q_{\min}} \left(\sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}} + 6 \sqrt{\log\left(\frac{(3 + n_e T) T^2}{\delta}\right)} \right) \times \sqrt{\log\left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta}\right)}.$$

We now establish high-probability performance guarantees for the predictors output by the algorithm.

Lemma 20 *Given any hypothesis set $\mathcal{R}$, for any $\delta > 0$, with probability at least $1 - \delta$, for all time steps $T \geq 3$ and all pairs $(r, r') \in \mathcal{R}_T^2$,*

$$|\mathcal{E}_T(r) - \mathcal{E}_T(r') - \mathcal{E}(r) + \mathcal{E}(r')| \leq \Delta_T.$$

In particular, if we let $r = r^$ and $r' = r_T$, it follows that*

$$\mathcal{E}(r_T) \leq \mathcal{E}(r^*) + \Delta_T.$$

Proof Apply Lemma 19 to time $T \geq 3$ and any pair $(r, r') \in \mathcal{R}_T^2$, with error probability $\delta/(T^2 |\mathcal{R}|^2)$ for round T . A union bound over $T \geq 3$ and (r, r') gives, with probability at least $1 - \delta$,

$$\begin{aligned}
&|\mathcal{E}_T(r) - \mathcal{E}_T(r') - \mathcal{E}(r) + \mathcal{E}(r')| \\
&\leq \frac{1}{T} \max \left\{ 2 \sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} \mathbb{E}\left[\frac{p_{t,k}}{q_{\min}^2} \mid \mathcal{F}_{t-1}\right]}, 6 \sqrt{\log\left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta}\right)} \right\} \times \sqrt{\log\left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta}\right)}.
\end{aligned} \tag{12}$$

Next, by (Cesa-Bianchi and Gentile, 2008, Proposition 2), with probability at least $1 - \delta$, for all $T \geq 3$, we can write

$$\sum_{t=1}^T \mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \leq \left(\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k} \right) + 36 \log \left(\frac{(3 + \sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}) T^2}{\delta} \right) \quad (13)$$

$$\begin{aligned} &+ 2 \sqrt{\left(\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k} \right) \log \left(\frac{(3 + \sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}) T^2}{\delta} \right)} \\ &\leq \left(\sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}} + 6 \sqrt{\log \left(\frac{(3 + n_e T) T^2}{\delta} \right)} \right)^2. \end{aligned} \quad (14)$$

Combining (12) and (13), we get with probability at least $1 - 2\delta$, for all $T \geq 3$,

$$\begin{aligned} &|\mathcal{E}_T(r) - \mathcal{E}_T(r') - \mathcal{E}(r) + \mathcal{E}(r')| \\ &\leq \frac{2}{T q_{\min}} \left(\sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}} + 6 \sqrt{\log \left(\frac{(3 + n_e T) T^2}{\delta} \right)} \right) \times \sqrt{\log \left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta} \right)}, \end{aligned}$$

as claimed. $\blacksquare$

We now present an upper bound on the expected number of cost queries required by the algorithm.

Lemma 21 *Given any hypothesis set $\mathcal{R}$, and distribution $\mathcal{D}$, with $\theta(\mathcal{D}, \mathcal{R}) = \theta$, for all $\delta > 0$, for all $T \geq 3$, with probability at least $1 - \delta$, we have*

$$\sum_{t=1}^T \mathbb{E} \left[\sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k} \mid \mathcal{F}_{t-1} \right] \leq 4n_e \theta K_L \left(\mathcal{E}^* T + O \left(\sqrt{\mathcal{E}^* T \log(T|\mathcal{R}|/\delta)} \right) \right) + O(\log^3(T|\mathcal{R}|/\delta)),$$

where K_L is a constant that depends on the loss function $\mathcal{L}$.

Proof By (Beygelzimer et al., 2009, Theorem 11), for $t \geq 3$, the following holds:

$$\mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \leq 4n_e \theta K_L (\mathcal{E}^* + \Delta_{t-1}),$$

where $\mathcal{E}^* = \mathcal{E}(r^*) = \inf_{r \in \mathcal{R}} \mathcal{E}(r)$ is the error of best-in-class. Plugging in the expression for Δ_{t-1} , and applying again a similar concentration inequality as before to relate $\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}$ to $\sum_{t=1}^T \mathbb{E}[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1}]$, we end up with a recursion on $\mathbb{E}[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1}]$:

$$\mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \leq 4n_e \theta K_L \mathcal{E}^* + \frac{4n_e \theta K_L c_1}{t-1} \sqrt{\sum_{s=1}^{t-1} \mathbb{E}_{(x_t, y_t)} [p_s \mid \mathcal{F}_{s-1}] + c_2 \left(\frac{\log[(t-1)|\mathcal{R}|/\delta]}{t-1} \right)}, \quad (15)$$

where $c_1 = 2 \sqrt{\log \left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta} \right)} = O \left(\sqrt{\log \left(\frac{T|\mathcal{R}|}{\delta} \right)} \right)$, and c_2 is a constant. For simplicity, denote by $4n_e \theta K_L = c_0$. We show by induction that for all $t \geq 3$,

$$\mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \leq c_0 \mathcal{E}^* + c_4 \sqrt{\frac{\mathcal{E}^*}{t-1}} + \frac{c_5}{t-1}, \quad (16)$$

for some constants c_4, c_5 . Assume by induction that (16) holds for all $s \leq t-1$. Thus, from (15), we have

$$\begin{aligned}
 & \mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \\
 & \leq c_0 \mathcal{E}^* + \frac{c_0 c_1}{t-1} \sqrt{c_0 \mathcal{E}^* (t-1) + 2c_4 \sqrt{\mathcal{E}^* (t-1)} + c_5 \log(t-1)} + c_2 \left(\frac{\log[(t-1)|\mathcal{R}|/\delta]}{t-1} \right) \\
 & \leq c_0 \mathcal{E}^* + \frac{c_0 c_1}{t-1} \left[\sqrt{c_0 \mathcal{E}^* (t-1) + 2c_4 \sqrt{\mathcal{E}^* (t-1)}} + \sqrt{c_5 \log(t-1)} \right] + c_2 \left(\frac{\log[(t-1)|\mathcal{R}|/\delta]}{t-1} \right) \\
 & \leq c_0 \mathcal{E}^* + \frac{c_0 c_1}{t-1} \left[\sqrt{c_0 \mathcal{E}^* (t-1)} + \frac{c_4}{\sqrt{c_0}} \right] + \frac{c_0 c_1 \sqrt{c_5 \log(t-1)} + c_2 \log[(t-1)|\mathcal{R}|/\delta]}{t-1} \\
 & = c_0 \mathcal{E}^* + \frac{c_0 c_1 \sqrt{c_0 \mathcal{E}^*}}{\sqrt{t-1}} + \frac{\sqrt{c_0} c_1 c_4 + c_0 c_1 \sqrt{c_5 \log(t-1)} + c_2 \log[(t-1)|\mathcal{R}|/\delta]}{t-1},
 \end{aligned}$$

where we use the fact that $\sqrt{a+b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$ for $a, b > 0$. To complete the induction, we need to show that

$$\frac{c_0 c_1 \sqrt{c_0 \mathcal{E}^*}}{\sqrt{t-1}} + \frac{\sqrt{c_0} c_1 c_4 + c_0 c_1 \sqrt{c_5 \log(t-1)} + c_2 \log[(t-1)|\mathcal{R}|/\delta]}{t-1} \leq c_4 \sqrt{\frac{\mathcal{E}^*}{t-1}} + \frac{c_5}{t-1}.$$

Thus, $c_4 = c_0 c_1 \sqrt{c_0} = O\left(\sqrt{\log\left(\frac{T|\mathcal{R}|}{\delta}\right)}\right)$, and

$$c_5 \geq c_0^2 c_1^2 + c_0 c_1 \sqrt{c_5 \log(t-1)} + c_2 \log[(t-1)|\mathcal{R}|/\delta] \implies c_5 = O(c_0^2 c_1^2 \log(T)) = O(\log^2(T|\mathcal{R}|/\delta)).$$

Thus,

$$\mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \leq c_0 \mathcal{E}^* + O\left(\sqrt{\log(T|\mathcal{R}|/\delta)}\right) \sqrt{\frac{\mathcal{E}^*}{(t-1)}} + \frac{O(\log^2(T|\mathcal{R}|/\delta))}{t-1}.$$

Finally,

$$\begin{aligned}
 \sum_{t=1}^T \mathbb{E} \left[\sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k} \mid \mathcal{F}_{t-1} \right] &= \sum_{t=1}^T \mathbb{E} \left[\sum_{k=1}^{n_e} q_{t,k} p_{t,k} \mid \mathcal{F}_{t-1} \right] \\
 &\leq \sum_{t=1}^T \mathbb{E} \left[\sum_{k=1}^{n_e} p_{t,k} \mid \mathcal{F}_{t-1} \right] \\
 &\leq 4n_e \theta K_L \left(\mathcal{E}^* T + O\left(\sqrt{\mathcal{E}^* T \log(T|\mathcal{R}|/\delta)}\right) \right) + O(\log^3(T|\mathcal{R}|/\delta)), \quad (17)
 \end{aligned}$$

■

Finally, we present the improved sample complexity and label complexity bounds for the budgeted two-stage deferral algorithm.

Theorem 22 *Let r_T denote the hypothesis returned by the budgeted two-stage deferral algorithm after T rounds and τ_T the total number of cost queries. Then, for all $\delta > 0$, with probability at least $1 - \delta$, for any $T > 0$, the following inequality holds:*

$$\mathcal{E}(r_T) \leq \mathcal{E}^* + \frac{2}{T q_{\min}} \left(\sqrt{\sum_{t=1}^T \sum_{k=1}^{n_e} p_{t,k}} + 6 \sqrt{\log \left(\frac{(3 + n_e T) T^2}{\delta} \right)} \right) \times \sqrt{\log \left(\frac{8 T^2 |\mathcal{R}|^2 \log(T)}{\delta} \right)}.$$

Moreover, with probability at least $1 - \delta$, for any $T > 0$, the following inequality holds:

$$\tau_T \leq 8 n_e \theta K_L \left(\mathcal{E}^* T + O \left(\sqrt{\mathcal{E}^* T \log(T |\mathcal{R}| / \delta)} \right) \right) + O(\log^3(T |\mathcal{R}| / \delta)),$$

where K_L is a constant that depends on the loss function L .

Proof The bound of generalization error $\mathcal{E}(r_T)$ follows from Lemma 20. To get the bound on the number of labels τ_T , we relate $\sum_{t=1}^T \mathbb{E}[\sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k} \mid \mathcal{F}_{t-1}]$ in Lemma 21 to $\tau_T = \sum_{t=1}^T \sum_{k=1}^{n_e} 1_{k_t=k} Q_{t,k}$ through a Bernstein-like inequality for martingales. Combining with (17) completes the proof. ■

Optimal q_t We note that the generalization bound is minimized when the expert sampling probabilities are uniform: $q_{t,k} = 1/n_e$ for all t and k , yielding $q_{\min} = 1/n_e$ since all experts are treated symmetrically. Thus, the learning and sample complexity bounds in Theorem 22 depend only logarithmically on T in the realizable case $\mathcal{E}^* = 0$ and improve upon the bounds given by (4) in Section 5.

Appendix G. Budgeted Deferral with ϵ -Cover

Our framework for budgeted deferral algorithms, along with its associated theoretical guarantees, can be effectively extended from finite to infinite hypothesis classes, using covering numbers. ϵ -Covers allow us to approximate an infinite hypothesis set with a carefully chosen finite one, while keeping the approximation error bounded.

Let's first formally define an ϵ -cover.

Definition 23 *Let $\mathcal{R}_\infty$ be an infinite hypothesis set equipped with a distance metric $\rho(\cdot, \cdot)$. A subset $\mathcal{G} \subset \mathcal{R}_\infty$ is termed an ϵ -cover of $\mathcal{R}_\infty$ if, for every hypothesis $r \in \mathcal{R}_\infty$, there exists a corresponding hypothesis $g \in \mathcal{G}$ such that their distance $\rho(r, g)$ is no more than ϵ . The covering number, denoted by $\mathcal{N}(\mathcal{R}_\infty, \epsilon)$, represents the cardinality of the smallest possible ϵ -cover for $\mathcal{R}_\infty$.*

The practical use of an ϵ -cover stems from its ability to ensure that the performance achievable within the finite cover $\mathcal{G}$ is close to the optimal performance within the larger infinite set $\mathcal{R}_\infty$. This relationship is captured by the following well-established lemma (see, e.g., (Cortes et al., 2019a, Lemma 4)).

Lemma 24 *The minimum error achievable within an infinite hypothesis set $\mathcal{R}_\infty$ and the minimum error achievable within its ϵ -cover $\mathcal{G}$ differ by at most ϵ . Specifically:*

$$\min_{r \in \mathcal{R}_\infty} \mathcal{E}(r) \leq \min_{g \in \mathcal{G}} \mathcal{E}(g) \leq \min_{r \in \mathcal{R}_\infty} \mathcal{E}(r) + \epsilon.$$

Proof The first inequality, $\min_{r \in \mathcal{R}_\infty} \mathcal{E}(r) \leq \min_{g \in \mathcal{G}} \mathcal{E}(g)$, is straightforward, as $\mathcal{G}$ is a subset of $\mathcal{R}_\infty$, meaning the best hypothesis in $\mathcal{R}_\infty$ must be at least as good as the best in $\mathcal{G}$.

For the second inequality, let $r^* = \operatorname{argmin}_{r \in \mathcal{R}_\infty} \mathcal{E}(r)$ denote a hypothesis that achieves the minimum error in $\mathcal{R}_\infty$. By the definition of an ϵ -cover, there must exist a hypothesis $g^* \in \mathcal{G}$ such that $\rho(g^*, r^*) \leq \epsilon$. Assuming a standard property where the absolute difference in errors is bounded by this distance metric, we have:

$$|\mathcal{E}(g^*) - \mathcal{E}(r^*)| \leq \rho(g^*, r^*) \leq \epsilon.$$

This implies that $\mathcal{E}(g^*) \leq \mathcal{E}(r^*) + \epsilon$. Since $\min_{g \in \mathcal{G}} \mathcal{E}(g)$ is, by definition, less than or equal to the error of any specific $g^* \in \mathcal{G}$, it follows that: $\min_{g \in \mathcal{G}} \mathcal{E}(g) \leq \mathcal{E}(g^*) \leq \mathcal{E}(r^*) + \epsilon$. This establishes the second inequality and completes the proof. $\blacksquare$

Lemma 24 provides a crucial insight: when the learner is faced with an infinite family of hypotheses $\mathcal{R}_\infty$, using a finite ϵ -cover $\mathcal{G}$ results in an approximation error (i.e., the potential increase in the minimum achievable error) of at most ϵ . Therefore, our budgeted deferral algorithm can be run using this finite hypothesis set $\mathcal{G}$ (ideally, the one with the smallest cardinality) as a proxy for $\mathcal{R}_\infty$. This strategy allows the algorithm to achieve favorable learning guarantees, ensuring that its performance is within ϵ of the optimal performance achievable within the original infinite hypothesis class.

Appendix H. High-Probability Label Complexity Bounds

Our main label complexity guarantees are stated in expectation form, as is standard in much of the prior active learning literature (Beygelzimer et al., 2009; Cortes et al., 2019a,b, 2020; Mohri et al., 2018). Here, we show how they can be strengthened to high-probability bounds.

Theorem 25 *Fix the sample S . Let $Q_t = \sum_{k=1}^{n_e} \mathbb{I}_{k_t=k} Q_{t,k} \in \{0, 1\}$ be the indicator variable that the algorithm queries at round t , with*

$$p_t = \mathbb{P}[Q_t = 1 \mid \mathcal{F}_{t-1}], \quad Q(S) = \sum_{t=1}^T Q_t, \quad \mu = \mathbb{E}[Q(S)] = \sum_{t=1}^T p_t.$$

Then, for any $\delta \in (0, 1/e)$ and $T \geq 3$, we have

$$\mathbb{P}\left[Q(S) - \mu > \max\left\{2\sqrt{\mu \log(1/\delta)}, 3\log(1/\delta)\right\}\right] \leq 4\log(T) \delta;$$

If $\mu \geq 4\log(1/\delta)$, then for any $\epsilon \in (0, 1]$,

$$\mathbb{P}[Q(S) \geq (1 + \epsilon)\mu] \leq 4\log(T) \exp\left(-\frac{\epsilon^2 \mu}{4}\right);$$

If $\mu < 4\log(1/\delta)$, then

$$\mathbb{P}[Q(S) \geq \mu + 3\log(1/\delta)] \leq 4\log(T) \delta.$$

Proof Define $X_t = Q_t - p_t$. Then $\{X_t, \mathcal{F}_t\}$ is a martingale-difference sequence with $|X_t| \leq 1$ and $\text{Var}(X_t | \mathcal{F}_{t-1}) = p_t(1 - p_t)$. Thus

$$V = \sum_{t=1}^T p_t(1 - p_t) \leq \mu.$$

Applying the version of Freedman's inequality from [Kakade and Tewari \(2008, Lemma 3\)](#) with $b = 1$ gives the first inequality.

For the second statement, if $\mu \geq 4 \log(1/\delta)$ then the $2\sqrt{V \log(1/\delta)}$ term dominates. Setting $\log(1/\delta) = \epsilon^2 \mu / 4$ yields the multiplicative form.

For the third statement, the additive regime $3 \log(1/\delta)$ dominates, which directly gives the claimed bound. $\blacksquare$

Corollary 26 *Assume that with probability at least $1 - \delta_1$ over the draw of S we have $\mu = \mathbb{E}[Q(S)] \leq B$. Fix $\epsilon > 0$ and let*

$$\delta_2 = 4 \log(T) \exp\left(-\frac{\epsilon^2 B}{4}\right).$$

Then, with probability at least $1 - (\delta_1 + \delta_2)$ (over both S and the algorithm's coin flips),

$$Q(S) \leq (1 + \epsilon)B.$$

Proof Condition on the event $\mu \leq B$, which occurs with probability at least $1 - \delta_1$. Applying Theorem 25 with $\mu \leq B$ gives

$$\mathbb{P}[Q(S) \geq (1 + \epsilon)B] \leq \delta_2.$$

A union bound yields the claim. $\blacksquare$

Discussion. Assume a fixed total failure probability $\delta_{\text{total}} = \delta$. By setting $\delta_1 = \delta/2$ and $\delta_2 = \delta/2$, we can solve for the required bound B on the expected queries μ . If the condition on the expected number of queries holds, $\mathbb{P}[\mu \leq B] \geq 1 - \delta/2$, and we choose B such that

$$B \geq \frac{4}{\epsilon^2} \log\left[\frac{8 \log(T)}{\delta}\right],$$

then with an overall probability of at least $1 - \delta$, the total number of queries $Q(S)$ is bounded by

$$Q(S) \leq (1 + \epsilon)B.$$

This shows that if the expected number of queries μ is logarithmic in both the time horizon T and inverse probability $1/\delta$, the realized number of queries $Q(S)$ is tightly concentrated around its expectation with high probability.