LEARNING TO ANSWER FROM CORRECT DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of learning to generate an answer (or completion) to a question (or prompt), where there could be multiple correct answers, any one of which is acceptable at test time. Learning is based on demonstrations of some correct answer to each training question, as in Supervised Fine Tuning (SFT). We formalize the problem as offline imitation learning in contextual bandits, with demonstrations from some optimal policy, without explicitly observed rewards. Prior work assumes that the demonstrator belongs to a low-complexity policy class, which motivates maximum likelihood estimation (i.e., log-loss minimization). In contrast, we propose relying only on the reward model (specifying which answers are correct) being in a low-cardinality class, which we argue is a weaker assumption. We show that likelihood maximization methods can fail in this case, and instead devise an alternative novel approach that learns with sample complexity logarithmic in the cardinality of the reward class. Our approach and guarantees are robust and apply even when learning from arbitrary demonstrators and to the relaxed pass@k error setting. Our work motivates looking beyond likelihood maximization when learning from demonstrations.

1 INTRODUCTION

Many real-world problems involve generating answers to a question, where there may be many *equally good* responses. A math question can have millions of equally valid but differently written solutions, a coding task can admit many different perfectly working implementations, and a recommendation query can be satisfied by multiple items. The learner's challenge is *not* to reproduce all correct responses, but to generate a *single good answer*.

This recurring structure can be formalized as *contextual bandits*. Each question corresponds to a context $x \in \mathcal{X}$, each candidate response corresponds to an action $y \in \mathcal{Y}$, and there is an unknown reward function $r_*(x, y)$ (which we assume is binary for simplicity)[1], indicating whether a response is good (equivalently correct when binary rewards). We consider the problem of learning from demonstrations given by *some optimal demonstrator* π_*, i.e., training set of $x_i \sim \mathcal{D}, y_i \sim \pi_*(\cdot \mid x_i)$, where $\pi_*(\cdot \mid x)$ is supported on the set of optimal actions for the context x, given by

$$\sigma_*(x) := \{y \in \mathcal{Y} : r_*(x, y) = 1\}\,.$$

Note that $|\sigma_*(x)|$ can be huge and there may be many (non-unique) optimal demonstrators. Our goal is to learn a predictor policy $\widehat{\pi}$ that produces a *good* action (response) to an unseen context (question) sampled from $\mathcal{D}$, i.e., it has low error of outputting actions that are incorrect, as captured by the following loss function.

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}) = \mathbb{E}_{x\sim\mathcal{D},\widehat{y}\sim\widehat{\pi}(\cdot|x)}\left[\mathbb{1}\{\widehat{y} \notin \sigma_*(x)\}\right] = \mathbb{E}_{x\sim\mathcal{D},\widehat{y}\sim\widehat{\pi}(\cdot|x)}\left[\mathbb{1}\{r_*(x, \widehat{y}) \neq 1\}\right]\,. \tag{1}$$

This is *offline imitation learning* (Rajaraman et al., 2020; Rashidinejad et al., 2021) in contextual bandits with an *optimal* demonstrator. In modern large language models (LLMs), this is exactly the problem addressed during the *supervised fine-tuning* (SFT) phase,[2] where the model is trained on

[1] We will return to the more general bounded reward functions later in the presentation.

[2] In LLMs, the actual response y consists of a sequence of tokens, and so it is a general Markov Decision Process (MDP). However, the transitions of this MDP are deterministic and rewards are generally sparse, and hence, contextual bandits provide with a useful abstraction for the purpose of our study. Indeed, this setup exactly corresponds to the prompt-completion formulation for LLMs (Ouyang et al., 2022; Rafailov et al., 2023; Huang et al., 2025; Wu et al., 2025).

curated datasets of prompt–completion pairs. Importantly, these demonstrations are *high-quality* but *not exhaustive*.

An important remark about our objective (Eq. (1)) is that it is entirely reward-driven, focusing solely on **validity**, which aligns with the *reward maximization* view of LLMs. We do not require matching the distribution of π_* (see Appendix A for more discussion on this issue). The typical situation we consider in our setup is during SFT, where the goal is to improve the model's performance on a specific task, and there are many ways to achieve good performance—for example, producing gold-winning IMO solutions (reward = 1, i.e., $\sigma(\text{question}) = \{\text{all completely correct solution texts to the problem}\}$). This set is unfathomably large, due both to variation in solution approaches and details, and to differences in the text itself, down to word choices, spacing, and punctuation. Rather than matching how a distribution over experts π_*, would write their solution, producing any single correct solution is sufficient to win a gold medal at the IMO.

Policy Class Assumption and Likelihood Maximization. A common approach when learning from demonstrations is to assume that the demonstrator π_* belongs to a low-capacity policy class Π. This motivates maximum likelihood estimation (MLE), or equivalently log-loss minimization, as a natural learning rule; see, e.g., Foster et al. (2024).[3] Indeed, in SFT for LLMs, one fits a model by minimizing log-loss on prompt–completion pairs. Theoretically, prior work shows that under the assumption $\pi_* \in \Pi$, MLE enjoys sharp $O(\frac{\log|\Pi|}{m})$ convergence guarantees with m samples—both in terms of the loss in Eq.(1) and in matching the distribution of the demonstrator π_* under a certain notion of distance, ensuring both validity and coverage at the same time (Cohen et al., 2024; Foster et al., 2024). However, for this guarantee to be meaningful, $\log|\Pi|$ must be small. But in learning from demonstrations, there may be many equally good demonstrations, and the policy class modeling them can be extremely large (e.g., modeling the generative process of all optimal demonstrators, such as graduate students producing solutions to math problems).

1.1 Our Contributions

In this work, departing from prior work, we instead propose an alternative to the above—we only assume that the underlying *reward model class* has low cardinality (Section 2), i.e., the set-valued function $\sigma_* : \mathcal{X} \to 2^{\mathcal{Y}}$ defining correct answers, comes from a low-cardinality model class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ and the demonstrator π_* is supported on σ_*. This is a strictly weaker and more realistic assumption: when creating QA datasets, often high-quality responses to prompts are hand-picked, without any attempt to enumerate all valid responses or generate from them according to a fixed specified apriori distribution. Thus, imposing any further assumptions on this generative sampling may be overly restrictive, as done in the low-capacity policy class assumption.

- **Simple Failures of MLE Under the Low-Cardinality Reward Class (Section 3).** We demonstrate that MLE, which is minimax optimal under the low-cardinality policy class assumption and also enjoys convergence in distribution (Foster et al., 2024; Cohen et al., 2024), *can fail to generalize* for low-cardinality *reward model classes* even in simple situations (Theorems 1 and 7).
- **New Learning Algorithms (Section 4).** We first present a simple voting-based learner that ensures learnability for finite $\mathcal{S}$ (Theorem 2). This already goes beyond MLE, though its sample complexity can be as large as $\Omega(|\mathcal{S}|)$ (Theorem 11). We then present a *simple* and *novel* learner that succeeds with only $O(\log|\mathcal{S}|)$ samples (Theorem 4), and thus optimal, showing that exponential improvement in sample complexity is achievable with the right inductive bias. This sample complexity has no dependence on $|\mathcal{Y}|$ or $|\sigma(x)|$ which can be huge.
- **Extensions (Section 5).** (1) We generalize our optimal learner for general bounded (not necessarily binary) reward model classes as long as the demonstrator is optimal (Section 5.1). (2) We then also see how this algorithm is robust and can be generalized to situations when the demonstrator π_* is not necessarily optimal, i.e., not supported on

[3] Even beyond imitation learning, where other forms of feedback may be available, it is routine to take the view that $\pi_* \in \Pi$ for some low-capacity Π (Yun et al., 2025; Zhan et al., 2023; Xie et al., 2024; Zhang et al., 2025; Agarwal et al., 2025; Huang et al., 2024).

Rule \ Assump.	low-cardinality Π	low-cardinality $\mathcal{S}$
MLE	$\propto \log \lvert\Pi\rvert$	May not learn
Our Learner	$\propto \log \lvert\Pi\rvert$	$\propto \log \lvert\mathcal{S}\rvert$

Table 1: Comparison of MLE and our learner under the low-cardinality assumptions on Π and $\mathcal{S}$. See Section 5 for the extensions of our optimal learner, for (1) general bounded reward classes (2) non-optimal demonstrator (3) pass@k error.

σ_* (Theorem 5). In this case, one can still compete with the loss of the demonstrator $L_{\mathcal{D},\sigma_*}(\pi_*)$ up to a small constant blowup (e.g., $1.5\, L_{\mathcal{D},\sigma_*}(\pi_*)$). (3) Given recent interest in the relaxed notion of pass@k error (a relaxation of the objective in Eq. (1)), we show that $\Theta(\log|\mathcal{S}|/\log k)$ samples are both sufficient for our learner and, in the worst case, necessary for any learner, yielding at most a $\log k$ improvement in sample complexity from an information-theoretic perspective (Theorem 6).

We defer a broader discussion of our results in the context of other recent works to Section 6.

2 Setting and Problem Definition

We now formalize our setup. Let $\mathcal{X}$ and $\mathcal{Y}$ respectively be countable sets of all plausible contexts and actions. Recall the terminology introduced in Section 1 of a ground-truth support function $\sigma_* : \mathcal{X} \to 2^{\mathcal{Y}}$, a marginal distribution $\mathcal{D} \in \Delta(\mathcal{X})$ and the expert demonstrator $\pi_* : \mathcal{X} \to \Delta(\mathcal{Y})$. We use $(x, y) \sim \mathcal{D} \times \pi$ to denote a joint distribution where $x \sim \mathcal{D}$ and $y \sim \pi(\cdot \mid x)$.

Two Function Approximations. Note that our setup makes a distinction between the *support* of optimal answers σ_*, and the demonstrator's policy π_*. Keeping the distinction in mind, there are two natural types of function approximations even in the realizable case:

- **Model class approximation.** There is a class of support functions $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ modeling optimal actions under different contexts. The demonstrator π_* is optimal and supported on some $\sigma_* \in \mathcal{S}$ (i.e., $\operatorname{supp}\pi_*(\cdot \mid x) \subseteq \sigma_*(x)$).
- **Policy class approximation.** There is a policy class $\Pi \subseteq (\Delta(\mathcal{Y}))^{\mathcal{X}}$ such that the unknown $\pi_* \in \Pi$.[4] Note that this does not directly specify σ_*. It is natural to consider $\sigma_*(x) = \sigma_{\pi_*}(x) = \operatorname{supp}\pi_*(\cdot \mid x)$ as the ground-truth support function for evaluating the loss (Eq.(1)).

We note that while the two views are closely related, they differ significantly under *cardinality-based capacity control* on Π and $\mathcal{S}$. In particular, assuming small $|\mathcal{S}|$ is much weaker, since the class of all optimal policies supported on it (defined below) can be huge, especially when $|\sigma(x)|$ are large.

$$\Pi_{\mathcal{S}} := \bigcup_{\sigma\in\mathcal{S}} \Pi_\sigma \ , \text{ where } \ \Pi_\sigma := \{\text{Any } \pi : \mathcal{X} \to \Delta(\mathcal{Y}) \text{ s.t. } \operatorname{supp}\pi(\cdot \mid x) \subseteq \sigma(x)\,, \forall x \in \mathcal{X}\}\,. \quad (2)$$

On the other hand, for any policy class Π, the associated model class $\mathcal{S}_\Pi := \bigcup_{\pi\in\Pi}\{\sigma_\pi \mid \sigma_\pi(x) = \operatorname{supp}\pi(\cdot \mid x), \forall x \in \mathcal{X}\}$ has cardinality no larger than Π (i.e., $|\mathcal{S}_\Pi| \le |\Pi|$).

In either cases, the goal is to find an ε-suboptimal policy $\widehat{\pi}(S) : \mathcal{X} \to \Delta(\mathcal{Y})$, for the loss in Eq.(1) from i.i.d. samples $S = \{(x_i, y_i) \sim_{iid} (\mathcal{D} \times \pi_*) : i \in [m]\}$, as formalized below.

Definition 1 (Learning from Correct Demonstrations)**.** *We say that the reward model class $\mathcal{S}$ is learnable from correct demonstrations by an estimator $\widehat{\pi} : (\mathcal{X} \times \mathcal{Y})^* \to (\Delta(\mathcal{Y}))^{\mathcal{X}}$ with sample*

[4] We note that Foster et al. (2024) considered a more general setting than ours, aiming to compete with π_* (not necessarily optimal) under arbitrary bounded rewards, while assuming $|\Pi| < \infty$. To relax this assumption, we instead rely on the reward model class, focusing first on the special case where π_* is guaranteed to be optimal. We will return to the case when this may not happen in Section 5.

complexity $m_{\mathcal{S},\widehat{\pi}} : (0,1) \times (0,1) \to \mathbb{N}$*, if for any* $\varepsilon, \delta \in (0,1)$*, for any sample size* $m \geq m_{\mathcal{S},\widehat{\pi}}(\varepsilon, \delta)$*, for any* $\mathcal{D}, \sigma_*, \pi_*$*, we have*

$$\mathbb{P}_{S \sim (\mathcal{D} \times \pi_*)^m} \left[L_{\mathcal{D},\sigma_*}(\widehat{\pi}(S)) \leq \varepsilon \right] \geq 1 - \delta \,.$$

The learnability for Π *follows exactly the same definition after replacing* $\sigma_*(x) = \operatorname{supp} \pi_*(\cdot \mid x)$ *for measuring the loss.*

3 Why MLE is Good for Low-Cardinality Π but Fails for $\mathcal{S}$?

Given a policy class $\Pi \subseteq (\Delta(\mathcal{Y}))^{\mathcal{X}}$, the (conditional) Maximum Likelihood Estimator (MLE) (or the log-loss minimizer) is defined as

$$\mathrm{MLE}_{\Pi}(S) = \arg\max_{\pi \in \Pi} \prod_{i=1}^{m} \pi(y_i \mid x_i) = \arg\min_{\pi \in \Pi} -\sum_{i=1}^{m} \log \pi(y_i \mid x_i) \,. \tag{MLE}$$

Low-cardinality Π. We first discuss the case when we assume Π has a low cardinality. In this case, MLE is actually *minimax optimal* with respect to $|\Pi|$. The proposition below formalizes $O(\frac{\log|\Pi|}{m})$ convergence rate (in similar spirit to Cohen et al. (2024); Foster et al. (2021)) for our setup.

Proposition 1. *For any finite* $\Pi \subseteq (\Delta(\mathcal{Y}))^{\mathcal{X}}$*, for any* $\mathcal{D}$ *and* $\pi_* \in \Pi$*, with probability at least* $1 - \delta$ *over* $S \sim (\mathcal{D} \times \pi_*)^m$*, any* $\widehat{\pi}_{\mathrm{mle}}(S) \in \mathrm{MLE}_{\Pi}(S)$ *enjoys the following guarantee:*

$$L_{\mathcal{D},\sigma_{\pi_*}}(\widehat{\pi}_{\mathrm{mle}}(S)) \leq D^2_{\mathsf{H}}(\widehat{\pi}_{\mathrm{mle}}(S), \pi_*) \leq \frac{6 \log(2|\Pi|/\delta)}{m} \,.$$

Thus Π *is learnable with* $\widehat{\pi}_{\mathrm{mle}}$ *(cf. Definition 1) with sample complexity:* $m_{\Pi,\widehat{\pi}_{\mathrm{mle}}}(\varepsilon, \delta) = \frac{6}{\varepsilon} \log(2|\Pi|/\delta)$ *.*

In the above, $D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) := \sum_{z \in \mathcal{Z}} (\sqrt{\mathbb{P}(z)} - \sqrt{\mathbb{Q}(z)})^2$ is the squared Hellinger distance between distributions $\mathbb{P}$ and $\mathbb{Q}$ over a discrete domain $\mathcal{Z}$. The proof (in Appendix C) uses the ideas from Foster et al. (2024); one can first use the standard guarantees in density estimation to establish the convergence in the squared Hellinger distance for $D^2_{\mathsf{H}}(\widehat{\pi}_{\mathrm{mle}}(S), \pi_*)$, followed by controlling the loss (Eq. (1)) in terms of $D^2_{\mathsf{H}}(\widehat{\pi}_{\mathrm{mle}}(S), \pi_*)$. However, this guarantee is only meaningful when $\log|\Pi|$ is small. In the context of LLMs, this assumption may be somewhat unrealistic, since post-SFT models are not as perfect as this guarantee suggests–namely, matching the distribution of the demonstrator (see Section 6 for further discussion).

Low-Cardinality $\mathcal{S}$: Simple Failures of MLE over Natural Policy Classes. We now ask what can we say about the performance of MLE when only $\mathcal{S}$ is finite. It is important to note that our problem (cf. Definition 1) is only specified in terms of the model class $\mathcal{S}$. To use likelihood maximization, we need to consider a policy class modeling the underlying generative process. In what follows, we describe two natural ways of defining the policy class based on the model class $\mathcal{S}$.

First, given the promise that π_* is supported on some $\sigma_* \in \mathcal{S}$, one natural choice is to consider the class of $\Pi_{\mathcal{S}}$ from Eq. (2) of all policies supported on $\mathcal{S}$. Note that $|\Pi_{\mathcal{S}}|$ may be infinite now making the guarantee in Proposition 1 vacuous. We shall see that MLE fails to produce even correct responses. It is simple to observe that, for any x observed in the training set, $\mathrm{MLE}_{\Pi_{\mathcal{S}}}(S)$ always outputs an action according to the empirical distribution of observed y_i's for $x_i = x$. However, on unseen contexts, it may output any action $y \in \sigma(x)$ for some consistent σ. Thus, this MLE on unseen contexts is exactly identical to consistency-based learner:

$$\widehat{\pi}_{\mathrm{Con}}(S)(x) = \widehat{y} \text{ for some } \widehat{y} \in \sigma(x) \text{ where } \sigma \in V(S) := \{\sigma \in \mathcal{S} : y_i \in \sigma(x_i), \forall (x_i, y_i) \in S\}. \tag{3}$$

While such a consistency-based learner is known to be optimal (for finite classes) in supervised learning problems like binary or multiclass classification, it fails for our problem and thus, MLE over $\Pi_{\mathcal{S}}$ also fails. Consider $\mathcal{S} = \{\sigma_0, \sigma_{01}\}$, where $\sigma_0(x) = \{0\}$ and $\sigma_{01}(x) = \{0, 1\}$ for all x. If the true hypothesis is $\sigma_* = \sigma_0$, then all observed labels are 0. However, σ_{01} also remains consistent, and thus $\mathrm{MLE}_{\Pi_{\mathcal{S}}}(S)$ may output 1 at test time—*failing to generalize and go beyond memorization*. Thus, for an input distribution $\mathcal{D}$ such that the *missing mass* (i.e., unobserved contexts) is arbitrarily close to 1, we get the following failure.

Theorem 1 (Failure of MLE over $\Pi_{\mathcal{S}}$)**.** *There exists $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ with $|\mathcal{S}| = |\mathcal{Y}| = 2$ and $\mathcal{X} = \mathbb{N}$ and a choice of (σ_*, π_*) such that for every sample size m and $\gamma \in (0,1)$, there exists a marginal distribution $\mathcal{D}$ such that for $S \sim (\mathcal{D} \times \pi_*)^m$, some $\widehat{\pi}_{\mathrm{mle}}(S) \in \mathrm{MLE}_{\Pi_{\mathcal{S}}}(S)$ has the following guarantee:*

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}\left(L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{mle}}(S)) \geq 1-\gamma\right) = 1\,.$$

In the above example, MLE over $\Pi_{\mathcal{S}}$ essentially *overfits*: it achieves zero error on the training data but fails to generalize to unseen data. A simple cause of it is just the fact that the induced policy class $\Pi_{\mathcal{S}}$ is too rich. As a remedy, we may consider a restricted policy class $\Pi_{\text{unif},\mathcal{S}}$ with size $|\mathcal{S}|$:

$$\Pi_{\text{unif},\mathcal{S}} := \{\pi_{\text{unif},\sigma} : \sigma \in \mathcal{S}\} \quad \text{where} \quad \pi_{\text{unif},\sigma}(\cdot \mid x) = \mathrm{Unif}(\sigma(x)).$$

This is a natural candidate for a class with restricted capacity, where the learner only knows π_* is supported on some $\sigma_* \in \mathcal{S}$. However, the optimal demonstrator's policy π_* need not coincide with the canonical choice $\pi_{\mathrm{unif},\sigma_*}$, although it is perfectly supported on σ_*. So $\Pi_{\text{unif},\mathcal{S}}$ is actually misspecified, i.e., it may not contain π_*. This mismatch suffices to make the MLE fail again on the restricted capacity class $\Pi_{\text{unif},\mathcal{S}}$; see Theorem 7 and its proof in Appendix D.

Remark 3.1 (MLE achieves overlap)**.** *An interesting property of MLE over the restricted class $\Pi_{\text{unif},\mathcal{S}} = \{\pi_{\mathit{unif},\sigma} : \sigma \in \mathcal{S}\}$ is that it still achieves an overlap with ground-truth correct answers for any sample size $m \geq \frac{1}{\varepsilon}\big(\log|\mathcal{S}| + \log(1/\delta)\big)$. For any $\widehat{\pi}_{\mathrm{mle}}(S) \in \mathrm{MLE}_{\Pi_{\text{unif},\mathcal{S}}}(S)$, we have*

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}\left(\mathbb{P}_{x\sim\mathcal{D}}[\,\mathrm{supp}\,\widehat{\pi}_{\mathrm{mle}}(S)(\cdot \mid x) \cap \sigma_*(x) = \emptyset\,] \;\leq\; \varepsilon\right) \geq 1-\delta\,.$$

Thus, its predictions overlap with the ground-truth responses on all but an ε-fraction of inputs, though it may still generate answers outside the support with some non-trivial probability. See Appendix D.1 for further discussion on this.

4 Learning Algorithms

In this section, we present our main result: a sample-efficient learner (Section 4.2). En route to this, we first see how natural approaches, though sufficient to ensure learnability under finite $\mathcal{S}$, may have a disappointing linear dependence on $|\mathcal{S}|$ in the sample complexity (Section 4.1).

4.1 Warm-up

Our rule is simple: output from the common intersection of consistent hypotheses if it is non-empty, and otherwise output an arbitrary y within the support of some consistent σ. This suffices to ensure learnability.

Input: Sample $S = \{(x_i, y_i) : i \in [m]\}$ and a finite model class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$.

- Let $V(S) := \{\sigma \in \mathcal{S} : y_i \in \sigma(x_i), \forall (x_i, y_i) \in S\}$
- Return the predictor $\textsc{Common-Intersection}(S) = \widehat{\pi}_{\mathrm{CI}}(S) : \mathcal{X} \to \mathcal{Y}$ as follows:

$$\widehat{\pi}_{\mathrm{CI}}(S)(x) = \begin{cases} y \in \bigcap_{\sigma\in V(S)} \sigma(x)\,, & \text{if } \bigcap_{\sigma\in V(S)} \sigma(x) \neq \emptyset\,; \\ \text{arbitrary } y & \text{otherwise.} \end{cases}$$

Theorem 2. *Any finite $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ is learnable (cf. Definition 1) using the rule $\widehat{\pi}_{\mathrm{CI}}$ with sample complexity $m_{\mathcal{S},\widehat{\pi}_{\mathrm{CI}}}(\varepsilon,\delta) = \varepsilon^{-1}|\mathcal{S}|\left(\log|\mathcal{S}| + \log(1/\delta)\right)$.*

The proof is in Appendix E.2. The dependence on $|\mathcal{S}|$ in Theorem 2 is $O(|\mathcal{S}|\log|\mathcal{S}|)$, in contrast to the logarithmic dependence in standard supervised learning. We show this dependence is tight in the worst case (up to a log factor) for this rule, and even for a seemingly stronger variant that outputs by majority vote over consistent hypotheses: $\widehat{\pi}_{\mathrm{Maj}}(S)(x) = \arg\max_{y\in\mathcal{Y}} |\{\sigma \in V(S) : y \in \sigma(x)\}|$. See Theorem 11 in Appendix E.3.

4.2 Sample Efficient Learner

We now design a sample-efficient learner that achieves optimal *logarithmic dependence* on $|\mathcal{S}|$—even under the finite cardinality assumption on $\mathcal{S}$ and without assuming anything further about the

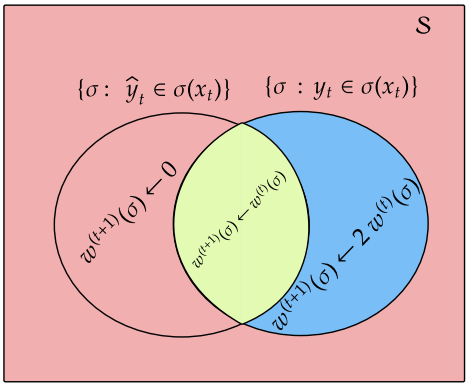

Figure 1: A visualization of the update rule of Algorithm 1 during t^{th} round. The weight of the hypotheses in the red, green, and blue regions are respectively set to zero, unchanged, and doubled.

optimal demonstrator's policy, we obtain sample complexity $\propto \log|\mathcal{S}|$, independent of $|\mathcal{X}|$, $|\mathcal{Y}|$, or $\sup_{\sigma,\,x}|\sigma(x)|$. To achieve this, we first turn our attention to the even more challenging online version of the problem. The statistical estimator will be designed by doing online-to-batch conversion.

Online Version. The adversary chooses $\sigma_* \in \mathcal{S}$. In each round t:

- The adversary chooses $x_t \in \mathcal{X}$. The learner predicts $\widehat{y}_t \in \mathcal{Y}$.
- The adversary shows some $y_t \in \sigma_*(x_t)$.
  (Importantly, the feedback does not inform the learner whether $\widehat{y}_t$ was a mistake or not.)

The algorithm maintains weight function $w^{(t)} : \mathcal{S} \to \mathbb{R}$ in each round.

Algorithm 1 MISTAKE-UNAWARE-WEIGHT-UPDATE

Input: A finite support class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$.

- Initialize $w^{(1)}(\sigma) = 1$ for all $\sigma \in \mathcal{S}$.
- In every round, receiving x_t:
  1. Output $\widehat{y}_t = \arg\max_{y\in\mathcal{Y}} \sum_{\sigma\in\mathcal{S}} w^{(t)}(\sigma)\,\mathbf{1}\{y \in \sigma(x_t)\}$.
  2. On receiving y_t, update the weights

$$w^{(t+1)}(\sigma) \leftarrow \begin{cases} 0 & \text{for all } \sigma \in \mathcal{S} \text{ with } y_t \notin \sigma(x_t)\,; \\ w^{(t)}(\sigma) & \text{for all } \sigma \in \mathcal{S} \text{ with both } y_t, \widehat{y}_t \in \sigma(x_t)\,; \\ 2\,w^{(t)}(\sigma) & \text{for all } \sigma \in \mathcal{S} \text{ with } y_t \in \sigma(x_t) \text{ and } \widehat{y}_t \notin \sigma(x_t)\,; \end{cases}$$

The key distinction from the standard *Majority* is in the final step of upweighting certain hypotheses: even without knowing whether $\widehat{y}_t$ was a mistake, the algorithm doubles the weight of all hypotheses that exclude $\widehat{y}_t$ but include the observed response y_t (see also Figure 1).

Theorem 3 (Online Guarantee). *On any sequence $((x_t, y_t))_{t\in\mathbb{N}}$ realizable by some $\sigma_* \in \mathcal{S}$, Algorithm 1 makes at most $\log_2|\mathcal{S}|$ mistakes.*

Proof. Letting $W_{t+1} = \sum_{\sigma\in\mathcal{S}} w^{(t+1)}(\sigma)$ be the total weight in the system after completion of t rounds, we first note that the sequence $\{W_t\}_t$ is non-increasing. This is because of the property of the algorithm that, during every round t, the weight added to the system is at most the weight eliminated from it. Formally,

$$W_{t+1} = \sum_{y_t\in\sigma(x_t) \text{ and } \widehat{y}_t\notin\sigma(x_t)} 2w^{(t)}(\sigma) + \sum_{y_t,\widehat{y}_t\in\sigma(x_t)} w^{(t)}(\sigma) \;\le\; \sum_{y_t\in\sigma(x_t) \text{ or } \widehat{y}_t\notin\sigma(x_t)} w^{(t)}(\sigma) \;\le\; W_t\,,$$

where the first inequality follows from the property of the algorithm that it always chooses $\widehat{y}_t = \arg\max_{y\in\mathcal{Y}} \sum_{\sigma\in\mathcal{S}} w^{(t)}(\sigma)\mathbf{1}\{y \in \sigma(x_t)\}$ (see also Figure 1). Now if the algorithm made M mistake on a realizable sequence for some $\sigma_* \in \mathcal{S}$ at the end some t number of rounds, then it must be that

$$w^{(t+1)}(\sigma_*) = 2^M \leq W_{t+1} \leq W_1 = |\mathcal{S}|\,, \text{ which implies } M \leq \log_2 |\mathcal{S}|\,.$$

□

We use the online-to-batch conversion of outputting a randomized predictor based on a random stopping time (for every test context).

Input: Sample $S = \{(x_i, y_i) : i \in [m]\}$ and a finite model class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$.

- Run Algorithm 1 once over S, and record $w^{(t)}$ before each round. Define

$$\widehat{\pi}_t(x) \;=\; \arg\max_{y\in\mathcal{Y}} \sum_{\sigma\in\mathcal{S}} w^{(t)}(\sigma)\,\mathbf{1}\{y \in \sigma(x)\}. \tag{4}$$

- On a test $x \in \mathcal{X}$, sample $I \sim \mathrm{Unif}\{1,\ldots,m\}$, and return $\widehat{\pi}_{\mathrm{o2b}}(S)(x) := \widehat{\pi}_I(x)$, i.e.,

$$\widehat{\pi}_{\mathrm{o2b}}(S)(x) = \frac{1}{m}\sum_{t=1}^{m} \widehat{\pi}_t(x). \tag{5}$$

Theorem 4 (Statistical Guarantee). *For any finite $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$, the estimator $\widehat{\pi}_{\mathrm{o2b}}$ in Eq. (5) achieves the following guarantee for any joint distribution $(\mathcal{D} \times \pi_*)$ supported on $\sigma_* \in \mathcal{S}$:*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S))] \leq \frac{\log_2 |\mathcal{S}|}{m}\,,$$

and, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S)) \;\leq\; \frac{1 + 2\log|\mathcal{S}| + 12\log\left(\frac{\log m}{\delta}\right)}{m}\,.$$

This implies that $\mathcal{S}$ is learnable (cf. Definition 1) using the estimator $\widehat{\pi}_{\mathrm{o2b}}$ with sample complexity $m_{\mathcal{S},\widehat{\pi}_{\mathrm{o2b}}}(\varepsilon,\delta) = O\left(\varepsilon^{-1}(\log|\mathcal{S}| + \log(1/\varepsilon\delta)\right)$.

See Appendix E for the proof. It follows from concentration for martingale difference sequences. To obtain a sharper dependence on $1/\varepsilon$ than in standard analyses (Cesa-Bianchi et al., 2004; Tewari & Kakade, 2008), we need a tighter control on the sum of martingale difference sequence, for which we apply Freedman's inequality (Lemma 2) with some algebraic manipulations.

5 Extensions

We now provide generalization of our optimal learner to important variants of our main setting.

5.1 General Bounded Reward Classes

Consider a finite reward class $\mathcal{R}$, consisting of functions $r : (\mathcal{X}\times\mathcal{Y}) \to [0,1]$. The demonstrator π_* is still optimal with respect to some unknown $r_* \in \mathcal{R}$, and is supported on

$$\sigma_*(x) := \arg\max_{y\in\mathcal{Y}} r_*(x,y)\,.$$

Both our optimal learner (and its pass@k variant, discussed in Section 5.3) can be implemented on the corresponding support class

$$\mathcal{S}_{\mathcal{R}} = \left\{\sigma_r \mid \sigma_r(x) = \arg\max_{y\in\mathcal{Y}} r(x,y),\ \forall x \in \mathcal{X}\right\}.$$

Since π_* is optimal, the reward sub-optimality of our learner can be directly controlled by the loss function in Eq. (1). In particular, we obtain the guarantee that the value of our predictor policy $V(\widehat{\pi}) \geq V(\pi_*) - \varepsilon$ with high probability, at a sample complexity determined by the size of $\mathcal{S}_{\mathcal{R}}$, which satisfies $|\mathcal{S}_{\mathcal{R}}| \leq |\mathcal{R}|$.

5.2 Learning from Arbitrary Demonstrator

We now discuss what happens if π_* is not necessarily optimal (again consider only binary rewards for simplicity). Note that in our Algorithm 1, we assign zero weight to all hypotheses that do not contain y_t—here we critically relied on the promise that the demonstrator is optimal. However, this strategy may be detrimental when it is not the case. In order to find an estimator, one needs to minimize a certain notion of *regret* of outputting bad responses with respect to any potential $\sigma \in \mathcal{S}$ in the online case, simply from demonstrations. To do so, we again output $\widehat{y}_t$ that is most voted, but upon receiving y_t, do a *"softer"* update:

$$w^{(t+1)}(\sigma) \leftarrow w^{(t)}(\sigma) \cdot \alpha^{\mathbf{1}[\widehat{y}_t \notin \sigma(x_t)]} \cdot \beta^{-\mathbf{1}[y_t \notin \sigma(x_t)]}\,, \text{ for some } \alpha, \beta \geq 1\,.$$

Then (α, β) are set so that, it ensures a certain *regret minimization* in the online case. Applying the standard online-to-batch conversion yields the following guarantee:

Theorem 5 (Learning from arbitrary demonstrator)**.** *For any finite class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$, there is an estimator $\widehat{\pi}$ such that for any unknown joint distribution $\mathcal{D} \times \pi_*$ (where π_* is not necessarily optimal) and reference $\sigma_* \in \mathcal{S}$, with probability $1-\delta$ over $S \sim (\mathcal{D} \times \pi_*)^m$, we have*

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}(S)) \leq 1.41\, L_{\mathcal{D},\sigma_*}(\pi_*) + \frac{1 + 6\log_2|\mathcal{S}| + 40\log\left(\frac{2\log 6m}{\delta}\right)}{m}.$$

See Appendix F for the proof and details of the algorithm.

5.3 pass@k Error Minimization

In modern practice, pass@k accuracy is often used as a benchmark (e.g., Chen et al. (2021); Orlanski et al. (2022); Dalal et al. (2025)). This relaxes the original goal by allowing a (stochastic) policy $\widehat{\mu} : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ to output k answers. The pass@k loss is defined as

$$L_{\mathcal{D},\sigma_*}(\widehat{\mu}) = \mathbb{E}_{x\sim\mathcal{D}}, \mathbb{E}_{\boldsymbol{y}=(y^{(1)},\ldots,y^{(k)})\sim\widehat{\mu}(\cdot|x)}\left[\mathbb{1}\{y^{(i)} \notin \sigma_*(x); \forall i \in [k]\}\right]. \tag{6}$$

The policy $\widehat{\mu}$ allows for any joint distribution over the set of k responses, which also includes adaptive sampling and need not be a product distribution (i.e., repeated independent sampling from a policy $\widehat{\pi} : \mathcal{X} \to \Delta(\mathcal{Y})$).

Our goal is to understand how the parameter k impacts sample complexity. We show that, under this relaxation, the worst-case improvement is only a $\log k$ factor in terms of the cardinality parameter. Again we return to the case where the demonstrator π_* is optimal. We extend our Algorithm 1 to this setting: the learner outputs a greedily picked set of responses $\{\widehat{y}_t^{\,(1)}, \ldots, \widehat{y}_t^{\,(k)}\}$ that covers most "uncovered mass" each time. Upon receiving y_t, the update is

$$w^{(t+1)}(\sigma) \leftarrow (k+1)\, w^{(t)}(\sigma)\,,$$

for all hypotheses that contain y_t but neither of $\{\widehat{y}_t^{\,(1)}, \ldots, \widehat{y}_t^{\,(k)}\}$. (The other updates are same.)

This leads to a sharp mistake bound of $\log_{k+1}|\mathcal{S}|$ in the online case, which then translates to a sample complexity in the statistical case, by the same online-to-batch conversion. See Appendix G for the details and formalization of the following informal result.

Theorem 6 (Informal: pass@k error)**.** *The minimax mistake bound in the online case as well as sample complexity bound in the statistical case are $\Theta(\log_k|\mathcal{S}|)$ for the family of finite classes of fixed size, when learning from optimal demonstrator.*

6 Discussion, Broader Conclusions, and Future Work

We studied the problem of learning to answer from correct demonstrations (aka imitation learning in contextual bandits from optimal demonstrator). Our goal was to move away from the common assumption that the demonstrator lies in a low-cardinality policy class Π, where MLE guarantees strong convergence to distribution of the demonstrator, under a certain notion of distance, thereby it can match the demonstrator also in performance. However, supervised fine-tuned models trained

via log-loss are not perfect (Ji et al., 2023; OpenAI, 2023)—they often fail to produce even correct responses. Even theoretically, the above results are in contrast to the alternative perspective of (Kalai & Vempala, 2024; Kalai et al., 2025), which argues that calibrated models (a coverage related notion) must necessarily hallucinate (i.e., incur high loss under our objective). The conclusion is that these guarantees, under the assumption $\pi_* \in \Pi$ with a low-capacity Π can be considered too crude to capture the current situations with LLMs (see also Section 6.2 of Foster et al. (2024)). Instead, we proposed a low-cardinality *reward model* class, which we argued is strictly weaker and more realistic assumption for current LLMs. This shift lets us expose simple failure cases of likelihood maximization and motivates new estimators that go beyond it in order to learn.

Broader Implications. Our new learner highlights that demonstrations available in SFT can carry much more information about the reward structure than what likelihood maximization—the standard approach for LLM fine-tuning—can extract. Another implication of our analysis concerns *hallucinations*. Our explanation is new: they arise naturally in the *prompted* question–answering scenario when learning is carried out via likelihood maximization. This departs from other recent theoretical perspectives (Kalai & Vempala, 2024; Kalavasis et al., 2025; Kalai et al., 2025). The closest to our work is (Kalai et al., 2025, Section 3.2), which matches our contextual bandit setting and similarly shows that hallucinations arise for statistical reasons. However, our failure cases are simpler and tied directly to MLE.

While their work attempts to bound hallucination rates in terms of a certain notion of "calibration"—encouraged by log-loss objective—our work goes further and asks: if *validity* and *coverage* are in tension, can we at least ensure correctness (reward maximization), when coverage may not be the primary goal anyway. We show that indeed it can be possible within our framework. There may, of course, be limited situations where coverage is important for fairness-related reasons, but even in such cases, the model should perhaps give a more comprehensive response indicating its uncertainty (e.g., (Kirchhof et al., 2025)). Thus the focus should perhaps be on validity/goodness/usefulness, with appropriate notion of utility, rather than attempting to learn the more difficult task of matching the distribution, cf. the quote from Vladimir Vapnik:

"When solving a problem of interest, do not solve a more general problem as an intermediate step."
—Vladimir Vapnik

Continuous Classes. This motivates the question what classes $\mathcal{S}$ should we use. In practice, we generally use parametric function classes, and thus a natural question is whether we can handle natural parametric continuous classes. For example, given a parameterized function class $h_w : (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ assigning real-valued scores, we can consider the associated support function $\sigma_w : x \mapsto \{y : h_w(x, y) \geq 0\}$. Natural choices for h_w include linear classes $h_w(x, y) = \langle w, \phi(x, y)\rangle$, low-rank score matrices $W \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ with $h_W(x, y) = W_{x,y}$, or even transformers $f_w(x, y) \in \mathbb{R}$ assigning scores to (x, y). Do the gaps we observed also manifest in these natural classes? What is the sample complexity of offline imitation learning for them? Can we characterize the complexity of the support class $\mathcal{S}_\mathcal{F} = \{\sigma_f : x \mapsto \{y : f(x, y) \geq 0\} \mid f \in \mathcal{F}\}$ in terms of abstract properties of $\mathcal{F}$, both for general classes and for simple parametric ones? What algorithms achieve this? One option is discretization and running our algorithm, yielding a log-cardinality bound statistically, but the algorithm's memory is exponential in the number of parameters and thus impractical. It is an important question to see if there are exact, approximate, or heuristic methods for learning with continuous classes that scale more favorably with the number of parameters (i.e., model size), and whether the inductive biases about the reward structure can be directly incorporated when learning from demonstrations, without modeling the policy of the demonstrator.

Technical Open Questions. Even for finite classes, important technical and conceptual gaps remain in our understanding. Our main learning rule in Section 4.2 is *randomized* and *improper* (see definition in Appendix A), whereas the simpler rule COMMON-INTERSECTION in Section 4.1, which is deterministic and proper, faces an $\Omega(|\mathcal{S}|)$ barrier. Is this gap fundamental? Can we design simpler rules even for finite classes that achieves $O(\log |\mathcal{S}|)$ sample complexity? Moreover, when the demonstrator is suboptimal, can we compete with the demonstrator's loss $L_{\mathcal{D},\sigma_*}(\pi_*)$ without incurring blow-up, either with our rule or any other?

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*, pp. 1638–1646. PMLR, 2014.

Alekh Agarwal, Christoph Dann, and Teodor V Marinov. Design considerations in offline preference-based rl. *arXiv preprint arXiv:2502.06861*, 2025.

DeepSeek AI. Deepseek models overview. https://www.deepseek.com, 2025.

Anthropic. Claude 3 model family. https://www.anthropic.com/news/claude-3, 2024.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems. *arXiv preprint arXiv:1209.2355*, 2012.

Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre Allauzen. Exploring precision and recall to assess the quality and diversity of llms. *arXiv preprint arXiv:2402.10693*, 2024.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Sandhini Puri, Gretchen Krueger, Mikhail Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Michael Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Phil Tillet, Felipe Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Szymon Jain, William Saunders, Jesse Hsu, Ryan Crowder, Anjali Srinivasan, Andrew Ho, Tom Gray, Nicholas Ryder, Dario Amodei, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Lee Cohen, Yishay Mansour, Shay Moran, and Han Shao. Probably approximately precision and recall learning. *arXiv preprint arXiv:2411.13029*, 2024.

Uri Dalal, Meirav Segal, Zvika Ben-Haim, Dan Lahav, and Omer Nevo. Leveraging llm inconsistency to boost pass@k performance. *arXiv preprint arXiv:2505.12938*, 2025.

Xingyu Dang, Christina Baek, Kaiyue Wen, Zico Kolter, and Aditi Raghunathan. Weight ensembling improves reasoning in language models. *arXiv preprint arXiv:2504.10478*, 2025. URL https://arxiv.org/abs/2504.10478.

Google DeepMind. Introducing gemini: Our largest and most capable ai model. https://blog.google/technology/ai/google-gemini-ai/, 2023.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Eric Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. Technical report, OpenAI, September 2025.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On the limits of language generation: Trade-offs between hallucination and mode-collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1732–1743, 2025.

Michael Kirchhof, Luca Füger, Adam Goliński, Eeshan Gunesh Dhekane, Arno Blaas, and Sinead Williamson. Self-reflective uncertainties: Do llms know their internal answer distribution? *arXiv preprint arXiv:2505.20295*, 2025.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Gabriel Orlanski, Seonhye Yang, and Michael Healy. Evaluating how fine-tuning on bimodal data affects code generation. *arXiv preprint arXiv:2211.07842*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025. doi: 10.48550/arXiv.2509.06941.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International conference on machine learning*, pp. 814–823. PMLR, 2015.

Ambuj Tewari and Sham Kakade. Online-to-batch conversions. Lecture notes, CMSC 35900: Learning Theory, Toyota Technological Institute at Chicago, 2008. https://home.ttic.edu/~tewari/lectures/lecture13.pdf.

Hugo Touvron et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025. URL https://arxiv.org/abs/2507.14843.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. URL https://arxiv.org/abs/2504.13837.

Jihun Yun, Juno Kim, Jongho Park, Junhyuck Kim, Jongha Jon Ryu, Jaewoong Cho, and Kwang-Sung Jun. Alignment as distribution learning: Your preference model is explicitly a language model. *arXiv preprint arXiv:2506.01523*, 2025.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Yudi Zhang, Lu Wang, Meng Fang, Yali Du, Chenghua Huang, Jun Wang, Qingwei Lin, Mykola Pechenizkiy, Dongmei Zhang, Saravan Rajmohan, et al. Distill not only data but also rewards: Can smaller language models surpass larger ones?, 2025. *URL https://arxiv. org/abs/2502.19557*, 2025.

Appendix Table of Contents

A Miscellaneous Deferred Discussions

Validity vs Coverage. For LLMs, we discuss several important reasons why our objective in Eq. (1) is important and worth studying in its own right, despite not demanding coverage or matching the distribution of the demonstrator. (1) First, it mirrors real-world usage of deployed LLMs (OpenAI, 2023; Anthropic, 2024; DeepMind, 2023; AI, 2025; Touvron et al., 2023), where the feedback signal is based solely on the quality of the single output shown to the user. The models are never directly evaluated using distributional distances, nor do we believe such evaluation is even feasible. (2) The objective may tolerate mode collapse, which might seem limiting to those who view coverage as important for fairness-related reasons. Even in such situations, the model should perhaps indicate its uncertainty (Kirchhof et al., 2025). And there are many other situations—typically during SFT—where coverage is not the primary goal. (3) Achieving both "validity" (precision) and "coverage" (roughly recall or calibration) is challenging and often in tension; we have both empirical (Bronnec et al., 2024) and theoretical (Kalai & Vempala, 2024; Kalai et al., 2025) evidence of this. In such situations, it is important to at least target the validity. (4) Finally, given the extensive alignment and post-training that modern LLMs undergo, which are completely reward-driven, it is unclear whether they preserve any coverage guarantees over correct responses (Song et al., 2025; Dang et al., 2025; Yue et al., 2025; Wu et al., 2025).

Proper vs Improper Learning. The model class $\mathcal{S}$ consists of set-valued functions σ, whereas the prediction is a single label. Thus, we define a notion of proper learning that is natural for our problem.

Definition 2 (Proper Learning)**.** *We call a learning rule $\widehat{\pi} : (\mathcal{X} \times \mathcal{Y})^* \to (\Delta(\mathcal{Y}))^{\mathcal{X}}$ proper if for any $S \in (\mathcal{X} \times \mathcal{Y})^*$, the policy $\widehat{\pi}(S)$ is supported on σ for some $\sigma \in \mathcal{S}$, i.e. $\mathrm{supp}(\widehat{\pi}(S)(\cdot \mid x)) \subseteq \sigma(x)$ for all $x \in \mathcal{X}$.*

Our optimal learner is randomized and improper. On the other hand, the rule COMMON-INTERSECTION is *deterministic* and also *proper*, after making a choice when outputting an arbitrary y—in the case where the common intersection is empty, we output a y that always belongs to $\sigma(x)$ for some fixed $\sigma \in V(S)$. The proofs of Theorems 2 and 9 still hold because we always treat an arbitrary y as a mistake anyway.

Related Work in Learning in contextual bandits. Contextual bandits have a long history in statistics. Early work includes Woodroofe (1979); Yang & Zhu (2002). The term contextual bandits was popularized by the paper Langford & Zhang (2007). A major line of work focuses online learning in contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agarwal et al., 2014); we refer interested readers to the recent book Lattimore & Szepesvári (2020) for more details. A parallel line of work focuses on the offline setting, where one learns from logged bandit feedback without active exploration. Influential contributions in this direction include Dudík et al. (2011); Bottou et al. (2012); Swaminathan & Joachims (2015). Most related to our work is imitation learning in contextual bandits, where the learner has access only to demonstrations from an expert. This perspective has been explored in recent theoretical studies (Rajaraman et al., 2020; Rashidinejad et al., 2021; Foster et al., 2024). They predominantly work under the low-capacity policy class assumption, which is different from our low-capacity reward class assumption.

B TECHNICAL PRELIMINARY LEMMAS

We start by a technical lemma about the one-sided change of measure bound on an expectation of a bounded function in terms of the Hellinger distance (e.g. Lemma A.11 from Foster et al. (2021)). We will use the exact variant from (Foster et al., 2024, Lemma 3.11).

Lemma 1 (Change-of-measure bound via Hellinger distance,Foster et al. (2024))**.** *Let $(\mathcal{Z}, \mathcal{F})$ be a measurable space and let $\mathbb{P}, \mathbb{Q}$ be probability measures on it. For every measurable function $h : \mathcal{Z} \to \mathbb{R}$:*

$$\left|\mathbb{E}_{\mathbb{P}}[h] - \mathbb{E}_{\mathbb{Q}}[h]\right| \;\le\; \sqrt{\frac{\mathbb{E}_{\mathbb{P}}[h^2] + \mathbb{E}_{\mathbb{Q}}[h^2]}{2}}\, D_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{7}$$

In particular for $h : \mathcal{Z} \to [0, R]$,

$$\mathbb{E}_{\mathbb{P}}[h] \;\le\; 2\,\mathbb{E}_{\mathbb{Q}}[h] \;+\; R\, D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})\;. \tag{8}$$

We now specify Freedman's inequality that provides us with a non-asymptotic bound on the sum of martingale difference sequence.

Lemma 2 (Freedman's inequality, Theorem 3 from Li et al. (2021))**.** *Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and write $\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_i]$. Let*

$$Y_m = \sum_{i=1}^{m} X_i$$

where (X_i) is a real-valued scalar sequence satisfying:

$$|X_i| \le R, \qquad \mathbb{E}_{i-1}[X_i] = 0 \quad \textit{for all } i \ge 1,$$

for some constant $R < \infty$. Define the predictable variance process

$$W_m := \sum_{i=1}^{m} \mathbb{E}_{i-1}[X_i^2],$$

and assume deterministically that $W_m \le \sigma^2$ for some constant $\sigma^2 < \infty$. Then for any integer $n \ge 1$, with probability at least $1 - \delta$,

$$\left|Y_m\right| \;\le\; \sqrt{8\,\max\left\{W_m,\; \frac{\sigma^2}{2^n}\right\}\log\left(\frac{2n}{\delta}\right)} \;+\; \frac{4}{3}R\log\left(\frac{2n}{\delta}\right).$$

Maximum Likelihood Estimation for Distribution Learning. We now state guarantee for the maximum likelihood estimator (MLE) for density estimation, exactly similar to (Foster et al., 2024, Section B.4). Given a class of candidate densities $\mathcal{G}$ and i.i.d. samples $z_1, \ldots, z_m \sim g_*$ (possibly not in $\mathcal{G}$), we define the empirical negative log-likelihood (log-loss) of $g \in \mathcal{G}$ as

$$L_{\log}(g) \;=\; -\sum_{i=1}^{m} \log g(z_i).$$

The maximum likelihood estimator is then

$$\hat{g}_{\mathrm{mle}} \;\in\; \arg\min_{g \in \mathcal{G}} \; L_{\log}(g). \tag{9}$$

Definition 3 (Log-loss covering number)**.** *For a class $\mathcal{G} \subseteq \Delta(\mathcal{Z})$, we say that a subset $\mathcal{G}' \subseteq \Delta(\mathcal{Z})$ is an ε-*cover with respect to the log-loss *if for all $g \in \mathcal{G}$ there exists $g' \in \mathcal{G}'$ such that $\sup_{z \in \mathcal{Z}} \log(g(z)/g'(z)) \le \varepsilon$. We denote the size of the smallest such cover by $\mathcal{N}_{\log}(\mathcal{G}, \varepsilon)$.*

We have the following property of MLE's convergence in the squared Hellinger distance with high probability.

Proposition 2. *With probability $1-\delta$ over m i.i.d. samples from any $g_* \in \mathcal{G}$,*

$$D^2_{\mathsf{H}}(g_*, \hat{g}_{\mathrm{mle}}) \;\le\; \inf_{\varepsilon > 0} \left\{ \frac{6 \log\big(2\,\mathcal{N}_{\log}(\mathcal{G}, \varepsilon)/\delta\big)}{m} + 4\varepsilon \right\} \;+\; 2 \inf_{g \in \mathcal{G}} \log\big(1 + D_{\chi^2}(g_* \,\|\, g)\big) \;+\; 2\,\varepsilon_{\mathrm{opt}}.$$

In particular, if $\mathcal{G}$ is finite and $\varepsilon_{opt} = 0$, the maximum likelihood estimator satisfies

$$D^2_{\mathsf{H}}(g_*, \hat{g}_{\mathrm{mle}}) \;\le\; \frac{6 \log\big(2\,|\mathcal{G}|/\delta\big)}{m} \;+\; 2 \inf_{g \in \mathcal{G}} \log\big(1 + D_{\chi^2}(g_* \,\|\, g)\big)\,.$$

Note that the term $\inf_{g \in \mathcal{G}} \log\big(1 + D_{\chi^2}(g_* \,\|\, g)\big)$ corresponds to the misspecification error, and is zero if $g_* \in \mathcal{G}$.

We note that the proof of (Foster et al., 2024, Proposition B.1) contains a couple of minor typographical errors. Namely, in Eq.(20) therein, the authors aim to compare $\widetilde{g}$ and $\widehat{g}$, but ended up comparing $\widetilde{g}$ and g_*. A similar mistake is repeated a couple of more times later in the proof without affecting the correctness of the argument.

C Proof of Proposition 1: MLE is Good for Low-Cardinality Π

In this, we present the proof of Proposition 1. As discussed in Section 3, the argument builds on a result from density estimation (Proposition 1) showing that MLE achieves convergence in squared Hellinger distance. We begin by providing intuition for why, even in this setting, MLE can obtain a guarantee expressed explicitly in terms of $\log|\Pi|$, without any dependence on $|\mathcal{X}|$, $|\mathcal{Y}|$, or the support size $\sup_{\sigma,x}|\sigma(x)|$, in a special case of the problem.

Intuition with the special cases of Π: We provide a more transparent and direct proof for the special case when for every $\pi \in \Pi, x \in \mathcal{X}$, the conditional density $\pi(\cdot \mid x)$ puts a uniform distribution over exactly s members of $\mathcal{Y}$ for some large but finite integer s. First, observe that in this special case we have a *dichotomy*; any hypothesis that does not contradict the data has the same likelihood as any other, so any $\pi \in \Pi$ that does not contradict with the data is MLE. For the unknown $\mathcal{D} \times \pi_*$, we now consider any π such that

$$L_{\mathcal{D},\sigma_*}(\pi) = \mathbb{P}_{x \sim \mathcal{D}, \widehat{y} \sim \pi(\cdot|x)}\left(\widehat{y} \notin \sigma_{\pi_*}(x)\right) > \varepsilon\,.$$

Then, due to the symmetry of the loss in the special case where each $\pi \in \Pi$ puts a uniform distribution on exactly s responses, we have

$$\mathbb{P}_{x \sim \mathcal{D}, \widehat{y} \sim \pi(\cdot|x)}\Big(\widehat{y} \notin \sigma_{\pi_*}(x)\Big) \;=\; \mathbb{P}_{x \sim \mathcal{D}, y \sim \pi_*(\cdot|x)}\Big(y \notin \sigma_{\pi}(x)\Big) > \varepsilon,$$

where the key fact used is the ability to change the order of randomness between $\widehat{y} \sim \pi(\cdot \mid x)$ and $y \sim \pi_*(\cdot \mid x)$.

This shows that when we sample $(x, y) \sim \mathcal{D} \times \pi_*$, the probability that (x, y) does not fall in the support $\sigma_\pi(x)$ exceeds ε. Hence, for any fixed $\pi \in \Pi$, after m i.i.d. draws

$$\mathbb{P}_S(\pi \text{ survives}) \le (1-\varepsilon)^m \le e^{-\varepsilon m}.$$

Therefore, by a standard union bound,

$$\mathbb{P}_S\big(\exists \text{ bad } \pi \in \Pi \text{ that survives}\big) \;\le\; |\Pi|\, e^{-\varepsilon m}.$$

The proposition follows by choosing $m \ge m_{\Pi,\widehat{\pi}_{\mathrm{mle}}}(\varepsilon, \delta) = \frac{\log|\Pi| + \log(1/\delta)}{\varepsilon}$.

Proof for any general Π: Consider any unknown but fixed marginal distribution $\mathcal{D} \in \Delta(\mathcal{X})$. For any conditional law $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$, let $\mathbb{P}_{(\mathcal{D},\pi)}$ denote the joint law over $(\mathcal{X} \times \mathcal{Y})$ given by the marginal distribution $\mathcal{D}$ and the conditional law $\pi(\cdot \mid x)$. First observe that for any $S \in (\mathcal{X} \times \mathcal{Y})^*$, the joint law $\mathbb{P}_{(\mathcal{D},\widehat{\pi}_{\mathrm{mle}}(S))}$ is the MLE of among all joint distribution $\{\mathbb{P}_{(\mathcal{D},\pi)} : \pi \in \Pi\}$. Using Proposition 2, for $S \sim (\mathcal{D} \times \pi_*)^m$

$$\mathbb{P}_S\left(D_{\mathsf{H}}^2\left(\mathbb{P}_{(\mathcal{D},\pi_*)}, \mathbb{P}_{\mathcal{D},\widehat{\pi}_{\mathrm{mle}}(S)}\right) \le \frac{6\log(2|\Pi|/\delta)}{m}\right) \ge 1-\delta\,. \tag{10}$$

Now let $\sigma_* : \mathcal{X} \to 2^{\mathcal{Y}}$ be the associated support set valued function of valid responses for $\sigma_*(x) = \mathrm{supp}(\pi_*(\cdot \mid x))$. Let us define a function $\mathrm{err} : (\mathcal{X} \times \mathcal{Y}) \to \{0, 1\}$ as

$$\mathrm{err}(x, y) = \begin{cases} 1 & \text{if } y \notin \sigma_*(x), \\ 0 & \text{otherwise.} \end{cases}$$

Then using Lemma 1, we have for any conditional law $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$

$$L_{\mathcal{D},\sigma_*}(\pi) = \mathbb{E}_{\mathbb{P}_{(\mathcal{D},\pi)}}[\mathrm{err}] \le D_{\mathsf{H}}^2(\mathbb{P}_{(\mathcal{D},\pi_*)}, \mathbb{P}_{(\mathcal{D},\pi)})\,,$$

where we used the fact that $L_{\mathcal{D},\sigma_*}(\pi_*) = \mathbb{E}_{\mathbb{P}_{\mathcal{D},\pi_*}}[\mathrm{err}] = 0$ and that err is a bounded function in $[0, 1]$. Combining this with (10), we obtain that with probability at least $1-\delta$ over $S \sim (\mathcal{D} \times \pi_*)^m$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{mle}}(S)) \le \frac{6\log(2|\Pi|/\delta)}{m}\,.$$

D Simple MLE Failures for Low-Cardinality $\mathcal{S}$

We first show that there is a simple instance of a support class, where some MLE over the entire class $\Pi_{\mathcal{S}} := \bigcup_\sigma \Pi_\sigma$ fails.

Proof of Theorem 1. Fix any $\gamma \in (0, 1)$. Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \mathbb{N}$. Define a support class $\mathcal{S} = \{\sigma_0, \sigma_{01}\} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ by

$$\sigma_0(x) = \{0\} \qquad \text{and} \qquad \sigma_{01}(x) = \{0, 1\} \quad \forall x \in \mathcal{X}.$$

Choose the ground-truth support $\sigma_* = \sigma_0$ and let the data-generating conditional be the point mass $\pi_*(\cdot \mid x) = \delta_0(\cdot)$ for all x; thus every observed label equals 0. Let $\Pi_{\mathcal{S}}$ be the class of all policies supported on $\mathcal{S}$ (see (2)). Now fix a sample size $m \in \mathbb{N}$. Set

$$q \;:=\; \left\lceil \frac{m}{\gamma} \right\rceil,$$

and define the marginal $\mathcal{D}$ to be the uniform distribution on $[q] = \{1, 2, \ldots, q\}$, i.e., $\mathcal{D}(\{x\}) = 1/q$ for $x \in [q]$ and 0 otherwise.

For any dataset $S = \{(x_i, y_i)\}_{i=1}^m \sim (\mathcal{D} \times \pi_*)^m$, write $S_{\mathrm{dis}} := \{x_i : i \in [m]\}$ for the set of distinct unlabeled inputs in S (so $S_{\mathrm{dis}} \le m$). Consider the predictor $\widehat{\pi}$ defined by

$$\widehat{\pi}(\cdot \mid x) = \begin{cases} \delta_0(\cdot), & x \in S_{\mathrm{dis}}, \\ \delta_1(\cdot), & x \notin S_{\mathrm{dis}}. \end{cases} \tag{11}$$

We claim that $\widehat{\pi} \in \mathrm{MLE}_{\Pi_{\mathcal{S}}}(S)$. Indeed, the log-likelihood is

$$\ell_{\log}(\pi; S) = \sum_{i=1}^{m} \log \pi(0 \mid x_i) = \sum_{x \in S_{\mathrm{dis}}} N_x(S) \, \log \pi(0 \mid x),$$

where $N_x(S) := |\{i : x_i = x\}|$. This expression depends only on the values $\pi(0 \mid x)$ for $x \in S_{\mathrm{dis}}$ and is maximized by setting $\pi(0 \mid x) = 1$ for every $x \in S_{\mathrm{dis}}$. For $x \notin S_{\mathrm{dis}}$ the likelihood does not constrain $\pi(\cdot \mid x)$, so any choice (for tie-braking) is a valid maximizer; in particular, (11) yields a valid MLE in $\Pi_{\mathcal{S}}$.

We next evaluate its population loss against the support σ_*. Since $\sigma_*(x) = \{0\}$ for all x,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}) = \mathbb{P}_{x\sim\mathcal{D},\, \widehat{y}\sim\widehat{\pi}(\cdot|x)}\big(\widehat{y} \notin \sigma_*(x)\big) = \mathbb{P}_{x\sim\mathcal{D}}\big(x \notin S_{\mathrm{dis}}\big) = 1 - \frac{|S_{\mathrm{dis}}|}{q}.$$

Using $|S_{\mathrm{dis}}| \le m$ and $q \ge m/\gamma$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}) \;\ge\; 1 - \frac{m}{q} \;\ge\; 1 - \gamma.$$

The bound holds deterministically for every sample S, hence

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}(L_{\mathcal{D},\sigma_*}(\widehat{\pi}) \ge 1 - \gamma) = 1\,.$$

□

We now show that the attempt to restrict the capacity of the class via another natural choice of considering $\Pi_{\mathrm{unif},\mathcal{S}} = \bigcup_{\sigma\in\mathcal{S}}\{\pi_{\mathrm{unif},\sigma}\}$ also does not work, when the expert demonstrations π_* does not necessarily follow the distribution $\pi_{\mathrm{unif},\sigma_*}$ while still showing optimal demonstrations.

Theorem 7 (MLE Failure 2)**.** *Fix $\gamma \in (0,1)$. There exists $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ with $|\mathcal{S}| = 2, |\mathcal{X}| = 1, |\mathcal{Y}| = 2\lceil 1/\gamma \rceil$, such that for some choice of $\mathcal{D} \times \pi_*$ for some π_* supported on $\mathcal{S}$ such that, for every sample size m, for $S \sim (\mathcal{D} \times \pi_*)^m$, there is a unique $\widehat{\pi}_{\mathrm{mle}}(S) \in \mathrm{MLE}_{\Pi_{\mathrm{unif},\mathcal{S}}}(S)$ that suffers from the following loss:*

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}\left(L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{mle}}(S)) \ge 1 - \gamma\right) = 1\,.$$

Proof of Theorem 7. Fix $\gamma \in (0,1)$ and let $s := \lceil 1/\gamma \rceil$. Take $\mathcal{X} = \{x\}$ and

$$\mathcal{Y} = \{y^\star\} \;\cup\; \{a_1, \ldots, a_{s-1}\} \;\cup\; \{b_1, \ldots, b_s\},$$

so $|\mathcal{Y}| = 1 + (s-1) + s = 2s = 2\lceil 1/\gamma \rceil$. Define $\sigma_1, \sigma_2 \in (2^{\mathcal{Y}})^{\mathcal{X}}$ by

$$\sigma_1(x) = \{y^\star, a_1, \ldots, a_{s-1}\} \quad (\text{size } s), \qquad \sigma_2(x) = \{y^\star, b_1, \ldots, b_s\} \quad (\text{size } s{+}1),$$

so $\sigma_1(x) \cap \sigma_2(x) = \{y^\star\}$ and they are otherwise disjoint. Let $\mathcal{S} = \{\sigma_1, \sigma_2\}$ and

$$\Pi_{\mathrm{unif},\mathcal{S}} := \{\pi_{\mathrm{unif},\sigma} : \sigma \in \mathcal{S}\}, \qquad \pi_{\mathrm{unif},\sigma}(y \mid x) = \begin{cases} \frac{1}{|\sigma(x)|}, & y \in \sigma(x), \\ 0, & \text{otherwise.} \end{cases}$$

Set $\mathcal{D}$ to be the point mass at x and choose the ground-truth support $\sigma_* = \sigma_2$ with data-generating conditional $\pi_* = \delta_{y^\star}$ (always emit $y^\star$). For any m, every dataset $S \sim (\mathcal{D}\times\pi_*)^m$ equals $\{(x, y^\star)\}^m$.

It is simple to see that $\pi_{\mathrm{unif},\sigma_1} \in \mathrm{MLE}_{\Pi_{\mathrm{unif},\mathcal{S}}}(S)$ is the unique maximum likelihood estimator. This is because

$$\prod_{i=1}^{m} \pi_{\mathrm{unif},\sigma_1}(y_i \mid x_i) = \left(\tfrac{1}{s}\right)^m, \qquad \prod_{i=1}^{m} \pi_{\mathrm{unif},\sigma_2}(y_i \mid x_i) = \left(\tfrac{1}{s+1}\right)^m.$$

However, with $\sigma_*(x) = \sigma_2(x)$, the estimator $\widehat{\pi}_{\mathrm{mle}}(S) = \pi_{\mathrm{unif},\sigma_1}$ has the error

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{mle}}(S)) = \mathbb{P}_{\widehat{y}\sim\pi_{\mathrm{unif},\sigma_1}(\cdot|x)}\big(\widehat{y} \notin \sigma_2(x)\big) = 1 - \pi_{\mathrm{unif},\sigma_1}(y^\star \mid x) = 1 - \frac{1}{s} \;=\; 1 - \frac{1}{\lceil 1/\gamma \rceil} \;\ge\; 1 - \gamma.$$

All bounds are deterministic given S, hence

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}(L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{mle}}(S)) \ge 1 - \gamma) = 1\,,$$

for every m, completing the proof. □

D.1 Overlap of MLE

We now show that MLE over the restricted capacity class $\Pi_{\text{unif},\mathcal{S}}$ attains a *non-trivial overlap* at the statistical limit of Theorem 4, though its failure to directly optimize the objective of interest (cf. Section 3). Note that it is trivial to achieve overlap, by just using a policy which puts mass on one good response from each support function class. However, this would produce the responses outside the support with overwhelming probability. Thus, for the purpose for overlap to be meaningful, we call it a *non-trivial* one if the output policy is only supported on one of $\sigma \in \mathcal{S}$ (some notion *properness*). The MLE over the class $\Pi_{\text{unif},\mathcal{S}} = \bigcup_{\sigma\in\mathcal{S}}\{\pi_{\text{unif},\sigma}\}$ will be of this form and achieves a non-trivial overlap.

Theorem 8. *For any finite class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ and an unknown joint realizable distribution $\mathcal{D} \times \pi_*$, where π_* is supported on some $\sigma_* \in \mathcal{S}$, for any estimator $\widehat{\pi}_{\text{mle}}(S) \in \text{MLE}_{\Pi_{\text{unif},\mathcal{S}}}(S)$, we have the following guarantee: for any sample size $m \geq \varepsilon^{-1}\left(\log|\mathcal{S}| + \log(1/\delta)\right)$, we have*

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}\left(\mathbb{P}_{x\sim\mathcal{D}}\left(\text{supp}(\widehat{\pi}_{\text{mle}}(S)(\cdot \mid x)) \cap \sigma_*(x) = \emptyset\right) \leq \varepsilon\right) \geq 1-\delta\,.$$

Proof of Theorem 8. The proof is simple. First note that $\pi_{\text{unif},\sigma_*}$ has non-zero likelihood. Therefore, any policy in the set $\text{MLE}_{\Pi_{\text{unif},\mathcal{S}}}(S)$ must have non-zero likelihood. Thus, for any σ for which $\pi_{\text{unif},\sigma} \in \text{MLE}_{\Pi_{\text{unif},\mathcal{S}}}(S)$, we must have that $\sigma \in V(S) := \{\sigma \in \mathcal{S} : y_i \in \sigma(x_i)\,\forall (x_i, y_i) \in S\}$. Therefore, in order to establish

$$\mathbb{P}_{S\sim(\mathcal{D}\times\pi_*)^m}\left(\mathbb{P}_{x\sim\mathcal{D}}\left(\text{supp}(\widehat{\pi}_{\text{mle}}(S)(x)) \cap \sigma_*(x) = \emptyset\right) \leq \varepsilon\right) \geq 1-\delta\,,$$

it suffices to establish

$$\mathbb{P}_S\left(\forall\sigma \in V(S) : \mathbb{P}_{x\sim\mathcal{D}}\left(\sigma(x) \cap \sigma_*(x) = \emptyset\right) \leq \varepsilon\right) \geq 1-\delta\,. \tag{12}$$

Consider any bad $\sigma \in \mathcal{S}$ such that $\mathbb{P}_{x\sim\mathcal{D}}(\sigma(x) \cap \sigma_*(x) = \emptyset) > \varepsilon$. With each draw $(x_i, y_i) \sim (\mathcal{D}\times\pi_*)$, we have that σ gets knocked-out of version space with probability at least ε, i.e. $\mathbb{P}_{(x_i,y_i)\sim(\mathcal{D}\times\pi_*)}(y_i \notin \sigma(x_i)) > \varepsilon$. Therefore, for any fixed σ, after sample $S \sim (\mathcal{D}\times\pi_*)^m$

$$\mathbb{P}_S(\sigma \in V(S)) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}.$$

Therefore, by a standard union bound,

$$\mathbb{P}_S\big(\exists \text{ bad } \sigma \in V(S) \text{ that survives}\big) \;\leq\; |\mathcal{S}|\, e^{-\varepsilon m} \leq |\mathcal{S}|\, 2^{-\varepsilon m}.$$

The theorem follows by noting that the $|\mathcal{S}|\, 2^{-\varepsilon m} \leq \delta$ for any $m \geq \frac{\log|\mathcal{S}|+\log(1/\delta)}{\varepsilon}$. □

Remark D.1 (Comparison with multi-class classification)**.** *Note that for multiclass classification when $\mathcal{S} \subseteq \mathcal{Y}^{\mathcal{X}}$ (i.e. all $|\sigma(x)| = 1$, the guarantee captured in (12) is enough to ensure learnability by just outputting a single predictor from $\widehat{\sigma} \in V(S)$ (i.e. consistent / ERM). This happens because the overlap implies that that labels are the same and so no error. However, for our problem despite this overlap, it is unclear how to output a single label so that it belongs to the support of σ_*. What would be sufficient for our problem is the following guarantee, where the quantifier $\forall\sigma \in V(S)$ is taken inside the randomness of test point sampling $x \sim \mathcal{D}$:*

$$\mathbb{P}_S\left(\mathbb{P}_{x\sim\mathcal{D}}\left(\forall\sigma \in V(S) : \sigma(x) \cap \sigma_*(x) = \emptyset\right) \leq \varepsilon\right) \geq 1-\delta\,.$$

However, we know that this provably requires the sample size where there is $\Omega(|\mathcal{S}|)$ dependence on cardinality—see the lower bound for Common-Intersection *estimator (Theorem 11).*

It may be possible to turn this into a predictor that directly starts to produce good responses, depending on the overlap among hypotheses and other types of feedback available in post-training (e.g., whether a generated response is good or not). This overlap can be captured by a parameter that reflects the need for repeated sampling and the number of feedback that must be queried, which in turn allows for a more quantitative understanding of how many feedbacks are required to guarantee performance in terms of this parameter. For example, this parameter would be maximum in the case of multi-class classification (Remark D.1) and no additional feedback is required. However, we leave it open to formulate an interesting setup that enables a study of both types of feedbacks together for our problem, and we do not attempt to investigate this any further.

E PROOFS FROM SECTION 4.2

We first show our proof of Theorem 4 and then return to the proofs for the rules COMMON-INTERSECTION and MAJORITY.

E.1 ONLINE-TO-BATCH ANALYSIS VIA FREEDMAN'S INEQUALITY

We have an online learner MISTAKE-UNAWARE-WEIGHT-UPDATE (Algorithm 1) that makes at most $\log_2 |\mathcal{S}|$ mistakes (Theorem 3). We now show, how using online-to-batch conversion via the estimator $\widehat{\pi}_{\mathrm{o2b}}$ (Eq. (5)), we can enjoy a similar sample complexity. The main difference from standard online-to-batch analysis is the use of Freedman's inequality (Lemma 2), which yields a better dependence on $1/\varepsilon$.

Proof of Theorem 4. Let $\ell_t = \mathbb{1}\{\widehat{\pi}_t(x_t) \notin \sigma_*(x_t)\}$. Because $\widehat{\pi}_t$ is a deterministic function of $S_{<t} = \{(x_i, y_i) : i < t\}$, we have

$$\mathbb{E}[\ell_t \mid S_{<t}] = L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)\,.$$

Hence

$$\mathbb{E}_S\left[L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S))\right] = \mathbb{E}_S\left[\frac{1}{m}\sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)\right] = \mathbb{E}_S\left[\frac{1}{m}\sum_{t=1}^{m}\ell_t\right] \le \frac{\log_2|\mathcal{S}|}{m}\,,$$

where in the last inequality we used Theorem 3, which guarantees $\sum_{t=1}^{m}\ell_t \le \log_2|\mathcal{S}|$.

For the high-probability statement, define the martingale differences

$$Z_t := L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) - \ell_t, \qquad \text{where } |Z_t| \le 1 \text{ almost surely.}$$

Then $\mathbb{E}[Z_t \mid S_{<t}] = 0$, and

$$\mathbb{E}[Z_t^2 \mid S_{<t}] = \mathbb{E}[(L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) - \ell_t)^2 \mid S_{<t}] = \mathrm{Var}(\ell_t \mid S_{<t}) = L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)(1 - L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)) \le L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)\,.$$

And taking $W_m = \sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)$ and $\sigma^2 = m$ suffices, thus, using Lemma 2 with $n = \log m$ inequality gives us with probability $1-\delta$

$$\begin{aligned}
\sum_{t=1}^{m} Z_t &\le \sqrt{8\left(1 + \sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)\right)\log\left(\frac{\log m}{\delta}\right)} + \frac{4}{3}\log\left(\frac{\log m}{\delta}\right) \\
&\le \frac{1}{2}\left(1 + \sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)\right) + 4\log\left(\frac{\log m}{\delta}\right) + \frac{4}{3}\log\left(\frac{\log m}{\delta}\right) \qquad \text{(GM} \le \text{AM)}
\end{aligned}$$

Substituting Z_t and rearranging terms,

$$\sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) \le 1 + 2\sum_{t=1}^{m}\ell_t + \frac{32}{3}\log\left(\frac{\log m}{\delta}\right)$$

Finally noting that $L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S)) = \frac{1}{m}\sum_{t=1}^{m} L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t)$ and that $\sum_{t=1}^{m}\ell_t \le \log_2|\mathcal{S}|$ (by Theorem 3), we obtain that with probability $1-\delta$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S)) \le \frac{1 + 2\log_2|\mathcal{S}| + 12\log\left(\frac{\log m}{\delta}\right)}{m}$$

□

E.2 UPPER BOUNDS FOR COMMON INTERSECTION AND MAJORITY

We start by analyzing the COMMON-INTERSECTION rule in the more difficult online setting which helps for the intuition for the statistical setting.

Theorem 9 (Online Guarantee for COMMON-INTERSECTION)**.** *On any sequence $((x_t, y_t))_{t\in\mathbb{N}}$ realizable by some $\sigma_* \in \mathcal{S}$, the rule* COMMON-INTERSECTION *(applied to the sequence seen so far) makes at most $|\mathcal{S}| - 1$ mistakes.*

Proof of Theorem 9. Consider any round t in which there was a mistake made by the rule. It must be that the set of consistent hypothesis V_t in that round, it must be that there was no *common intersection* in that round, i.e., $\bigcap_{\sigma \in V_t} \sigma(x_t) = \emptyset$. That means even though we would not know whether we made a mistake in that round, observing y_t will eliminate at least one hypothesis from the version space (i.e. $|V_{t+1}| \leq |V_t| - 1$). Therefore, the rule cannot make $|V_1| - 1 = |\mathcal{S}| - 1$ mistakes on any realizable sequence. □

We now analyze the performance of this rule in the statistical version.

Proof of Theorem 2. Partition the m examples into $K := |\mathcal{S}|$ consecutive blocks $B_1, \ldots, B_K$, each of length $n \geq \frac{1}{\varepsilon}(\log|\mathcal{S}| + \log(1/\delta))$. Let V_t denote the version space just before block B_t begins. I.e. define the restricted dataset $S_t = B_1 \cup \cdots \cup B_{t-1}$ and

$$V_t = \{\sigma \in \mathcal{S} : y_i \in \sigma(x_i)\ \forall (x_i, y_i) \in S_t\}$$

with $V_1 = \mathcal{S}$. Define the region of x, where we do not have a common intersection among V_t.

$$A_t := \Big\{x \in \mathcal{X} : \bigcap_{\sigma \in V_t} \sigma(x) = \emptyset\Big\}.$$

Note that $A_{t+1} \subseteq A_t$ for all $t \in [K]$ because $V_{t+1} \subseteq V_t$. Moreover, we never make an error when outputting from common-intersection region. Now, using these facts, we have

$$\begin{aligned}
\mathbb{P}_S\left(L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{CI}}(S)) > \varepsilon\right) &\leq \mathbb{P}_S\left(\mathbb{P}_{x\sim\mathcal{D}}(A_{K+1}) > \varepsilon\right) \\
&= \mathbb{P}_S\left(V_{K+1} \neq \emptyset \cap \mathbb{P}_{x\sim\mathcal{D}}(A_{K+1}) > \varepsilon\right) \quad (V_{K+1} \neq \emptyset \text{ always happens}) \\
&\leq \mathbb{P}_S\left(\exists t \in [K],\ V_{t+1} = V_t\ \cap\ \mathbb{P}_{x\sim\mathcal{D}}(A_{K+1}) > \varepsilon\right) \\
&\leq \mathbb{P}_S\left(\exists\, t \in [K], V_{t+1} = V_t\ \cap\ \mathbb{P}_{x\sim\mathcal{D}}(A_t) > \varepsilon\right) \\
&\leq \sum_{t=1}^{K} \mathbb{P}_{B_t}\left(\exists\, t \in [K], V_{t+1} = V_t\ \mid \mathbb{P}_{x\sim\mathcal{D}}(A_t) > \varepsilon\right) \\
&\leq \sum_{t=1}^{K} (1-\varepsilon)^{|B_t|} = K(1-\varepsilon)^n \\
&\leq |\mathcal{S}|\, 2^{-\varepsilon n} \leq |\mathcal{S}| \cdot \frac{\delta}{|\mathcal{S}|} = \delta\,.
\end{aligned}$$

The above calculation formalizes the following argument. There are two cases to consider:
Case 1. If $\mathbb{P}_{x\sim\mathcal{D}}(A_t) > \varepsilon$, then the probability that no $x \in A_t$ appears in block B_t is at most $(1-\varepsilon)^n \leq e^{-\varepsilon n} \leq 2^{-\varepsilon n} \leq \delta/|\mathcal{S}|$. Otherwise, some $x \in A_t$ appears; since $\bigcap_{\sigma \in V_t} \sigma(x) = \emptyset$, the observed label $y \in \sigma_*(x)$ excludes at least one $\sigma \in V_t$, so $|V_{t+1}| \leq |V_t| - 1$.

Case 2. If $\mathbb{P}_{x\sim\mathcal{D}}(A_t) \leq \varepsilon$, then on A_t^c the intersection is nonempty, and because $\sigma_* \in V_t$ it follows that the COMMON-INTERSECTION prediction is always correct there. Moreover, since $V_{t+1} \subseteq V_t$, the intersections can only grow and hence $A_{t+1} \subseteq A_t$. Therefore, once Case 2 holds, the final error remains below ε.

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{CI}}(S)) \leq \mathbb{P}_{\mathcal{D}}(A_t) < \varepsilon\,.$$

Putting these together, with probability at least $1 - K \cdot (\delta/|\mathcal{S}|) \geq 1 - \delta$, each block in Case 1 eliminates at least one hypothesis, and there are at most $K = |\mathcal{S}|$ such eliminations are even possible. Hence either Case 2 occurs in some block (giving final error $< \varepsilon$), or Case 1 occurs in all K blocks, which is not possible in the realizable setting, arriving at a contradiction. □

E.3 LOWER BOUNDS FOR COMMON INTERSECTION AND MAJORITY

For the lower bound, we will show the lower bound on a stronger rule which outputs according to a majority vote among the consistent hypothesis formalized below.

Input: Sample $S = \{(x_i, y_i) : i \in [m]\}$ and a finite support hypothesis class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$.

- Let $V(S) := \{\sigma \in \mathcal{S} : y_i \in \sigma(x_i), \forall (x_i, y_i) \in S\}$
- Return the predictor $\textsc{Majority}(S) = \widehat{\pi}_{\mathrm{Maj}}(S) : \mathcal{X} \to \mathcal{Y}$ defined as follows:

$$\widehat{\pi}_{\mathrm{Maj}}(S)(x) = \arg\max_{y \in \mathcal{Y}} |\{\sigma \in V(S) : y \in \sigma(x)\}|$$

Corollary 1. *Note that $\widehat{\pi}_{\mathrm{Maj}}$ always outputs from the common-intersection whenever it is non-empty, and thus, it enjoys the same online as well as the statistical guarantees as in Theorem 9 and Theorem 2 respectively.*

The lower bounds will hold for the following instance of the of the class.

Description of the class. Fix $d \in \mathbb{N}$. Let

$$\mathcal{Y} = \{0, 1\}, \qquad q := \left\lfloor \tfrac{d-1}{2} \right\rfloor, \qquad \mathcal{X} := \{1, 2, \ldots, q\}.$$

We define a hypothesis class $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_d\} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ as follows.

- **Distinguished hypothesis.** Set $\sigma_1(x) = \{1\}$ for every $x \in \mathcal{X}$. This will serve as the ground-truth hypothesis.
- **Adversarial hypotheses.** For each $i \geq 2$, require that $0 \in \sigma_i(x)$ for all $x \in \mathcal{X}$. Moreover, for each $t \in \{1, \ldots, q\}$ we designate a *pair* of hypotheses, $\sigma_{2t}, \sigma_{2t+1}$, that both exclude label 1 at coordinate t:
$$1 \notin \sigma_{2t}(t), \qquad 1 \notin \sigma_{2t+1}(t).$$
For all other coordinates $x \neq t$, these hypotheses include both labels, e.g.
$$\sigma_{2t}(x) = \sigma_{2t+1}(x) = \{0, 1\} \quad \text{for } x \neq t.$$
If $(d-1$ is odd, then there is one remaining index pairing. In that case, define $\sigma_d(x) = \{0, 1\}$ for all $x \in \mathcal{X}$.

Thus $\mathcal{S}$ has size exactly d, uses $2q \leq d-1$ adversarial hypotheses to plant two "anti–1" voters at each coordinate $t \in \mathcal{X}$, and possibly one additional "neutral" hypothesis if $d-1$ is odd. We are now ready to show the online lower bound.

Theorem 10 (Online Lower Bounds for COMMON-INTERSECTIONand MAJORITY)**.** *For every d, there exists a hypothesis class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ with $|\mathcal{S}| = d, |\mathcal{X}| = \lfloor (d-1)/2 \rfloor$ and $|\mathcal{Y}| = 2$ such that both the rules make $|\mathcal{X}|$ mistakes (i.e. a mistake on every round).*

Proof of Theorem 10. Note that it suffices to show the lower bound of simply the MAJORITY rule, which also implies the lower bound on COMMON-INTERSECTION. Consider the hypothesis class $\mathcal{S}$ constructed above, and let the ground truth be $\sigma_* = \sigma_1$. Present the sequence of instances $x_t = t$ for $t = 1, \ldots, q = |\mathcal{X}|$. Then $y_t = 1$ for all t under σ_*.

At each round t, the version space V_{t-1} contains σ_1 together with all adversarial hypotheses that have not yet been eliminated. By construction, every adversarial hypothesis other than σ_1 always includes 0, while at coordinate t at least two of them exclude 1. Hence

$$N_0(x_t; V_{t-1}) = |V_{t-1}| - 1 \quad \text{and} \quad N_1(x_t; V_{t-1}) \leq |V_{t-1}| - 2,$$

so the majority rule predicts 0 (which is an error according to $\sigma_* = \sigma_1$) and errs, therefore the rule makes an error on every round, completing the proof. □

Remark E.1. *Note that the rule* COMMON-INTERSECTION *and* MAJORITY *respectively recover the textbook rules Consistent and Halving in the standard realizable online classification Shalev-Shwartz & Ben-David (2014). However, both the rules have a mistake bound of $\Omega(|\mathcal{S}|)$ in our setup even when the labels are binary in the worst case (cf. Theorem 10). This is in sharp contrast with the standard classification where Halving enjoys $\log_2 |\mathcal{H}|$ mistake bound. This failure is due to the set-valued nature of the support functions.*

We now show the statistical lower bound in a similar spirit.

Theorem 11 (Statistical Lower Bounds for COMMON-INTERSECTION and MAJORITY)**.** *For every d, there exists a problem instance $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ with $|\mathcal{S}| = d, |\mathcal{X}| = \lfloor (d-1)/2 \rfloor, |\mathcal{Y}| = 2$ and some choice of realizable joint distribution $(\mathcal{D} \times \pi_*)$ where π_* is supported on $\sigma_* \in \mathcal{S}$ such that for any sample size $m \leq |\mathcal{X}|/2$, letting $\widehat{\pi}(S)$ to be either* COMMON-INTERSECTION *or* MAJORITY *have the following guarantee:*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}, \sigma_*}(\widehat{\pi}(S)) \geq 1/2 \right) = 1.$$

Proof of Theorem 11. It suffices to prove the claim for MAJORITY again; the bound for the other follows since MAJORITY is just a special instantiation of COMMON-INTERSECTION.

Consider the same class $\mathcal{S}$ constructed above with $|\mathcal{X}| = q$ and ground truth $\sigma_* = \sigma_1$ (so the realizable label is always 1). Let $\mathcal{D}$ be the uniform distribution on $\mathcal{X}$. Take π_* to be the only conditional distribution supported on σ_*, so the joint $(\mathcal{D} \times \pi_*)$ is realizable.

Fix any sample size $m \leq q/2$ and draw $S \sim \mathcal{D}^m$. Let $S_{\text{unseen}} \subseteq \mathcal{X}$ be the set of coordinates *unseen* in S; then $|S_{\text{unseen}}| \geq q - m \geq q/2$. Let $V(S) \subseteq \mathcal{S}$ be the version space of hypotheses consistent with S (with respect to the labels of σ_* label, which are always 1).

Again, by construction, for each $t \in S_{\text{unseen}}$ there are two designated adversarial hypotheses in $V(S)$. At such a point $t \in S_{\text{unseen}}$:

- Every hypothesis in $V(S)$ includes label 0, except σ_*, so
$$N_0(t; V(S)) = |V(S)| - 1.$$
- Every hypothesis in $V(S)$ includes label 1, except $\sigma_{2t}, \sigma_{2t+1}$, so
$$N_1(t; V(S)) \leq |V(S)| - 2.$$

Thus $N_0(t; V(S)) > N_1(t; V(S))$, and the majority rule outputs 0 (which is an error according to σ_*). Thus, $\sigma_* = \sigma_1$ is $y_t = 1$, MAJORITY errs on every unseen $t \in S_{\text{unseen}}$.

With $\mathcal{D}$ uniform on $\mathcal{X}$,

$$L_{\mathcal{D}, \sigma_*}(\widehat{\pi}(S)) \;\geq\; \mathbb{P}_{x \sim \mathcal{D}}\big[x \in S_{\text{unseen}}\big] \;=\; \frac{|S_{\text{unseen}}|}{q} \;\geq\; 1 - \frac{m}{q} \;\geq\; \frac{1}{2}.$$

Since this lower bound holds for every realization of S with $m \leq q/2$, we have $\mathbb{P}_{S \sim (\mathcal{D} \times \pi_*)^m}\big(L_{\mathcal{D}, \sigma_*}(\widehat{\pi}(S)) \geq \frac{1}{2}\big) = 1$. This proves the theorem. □

Because $d = (2k)^q$, we have $q = \log_{2k} d$, so the bound implies $m = \Omega\big(\log_k d\big)$ under $\mathcal{D} = \text{Unif}(\mathcal{X})$.

Remark E.2. *Both online (Theorem 15) and statistical lower bounds (Theorem 15) for* pass@k *essentially demonstrate that one cannot do better than memorization below $\Omega(\log_k d)$ barrier in the worst-case, even for the special case of the problem of multiclass classification $\mathcal{S} \subseteq \mathcal{Y}^{\mathcal{X}}$ which is isomorphic to $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ with $|\sigma(x)| = 1$.*

F ALGORITHM AND ANALYSIS FOR LEARNING FROM ARBITRARY DEMONSTRATOR

We will first begin by online learner which makes a softer update, and then show how the online-to-batch conversion enjoys the guarantee given in Theorem 5.

F.1 ONLINE LEARNER WITH SOFTER UPDATE

Our goal is to bound the **regret** against the best hypothesis in hindsight in a certain sense. Let's define two types of mistakes over T rounds for any given hypothesis $\sigma \in \mathcal{S}$:

Algorithm 2 Softer MISTAKE-UNAWARE-WEIGHT-UPDATE

Input: A finite support class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ and the penalty parameters $\alpha, \beta > 1$.

- Initialize $w^{(1)}(\sigma) = 1$ for all $\sigma \in \mathcal{S}$.
- In every round $t = 1, \ldots, T$, receiving x_t:
  1. Output
  $$\widehat{y}_t = \arg\max_{y \in \mathcal{Y}} \sum_{\sigma \in \mathcal{S}} w^{(t)}(\sigma)\, \mathbf{1}\{y \in \sigma(x_t)\}\,.$$
  2. On receiving y_t, update the weights:
  $$w^{(t+1)}(\sigma) \leftarrow w^{(t)}(\sigma) \cdot \alpha^{\mathbf{1}\{\widehat{y}_t \notin \sigma(x_t)\}} \cdot \beta^{-\mathbf{1}\{y_t \notin \sigma(x_t)\}} \quad \forall \sigma \in \mathcal{S}.$$
  Here, $\alpha > 1$ and $\beta > 1$ are penalty parameters. Note that this update rule ensures that after T rounds, the weight of any hypothesis σ is exactly $w^{(T+1)}(\sigma) = \alpha^{M_T^{\text{alg}}(\sigma)} \beta^{-M_T(\sigma)}$.

- **Algorithm's Mistakes relative to σ:** This is the number of times our algorithm's prediction $\widehat{y}_t$ was not in σ's valid set absolutely, irrespective of what y_t was.
$$M_T^{\text{alg}}(\sigma) := \sum_{t=1}^{T} \mathbf{1}[\widehat{y}_t \notin \sigma(x_t)]$$
- **Hypothesis σ's Mistakes:** This is the number of times the true label y_t was not in σ's valid set.
$$M_T(\sigma) := \sum_{t=1}^{T} \mathbf{1}[y_t \notin \sigma(x_t)]$$

The total number of mistakes made by our algorithm with respect to σ is $M_T^{\text{alg}}(\sigma)$, and the number of mistakes made by σ itself is $M_T(\sigma)$. Note that in some rounds, both types of mistake could occur. We now analyze the algorithm with the first simple lemma that the total weight of the hypotheses in the system is again non-increasing under a minor condition on (α, β).

Lemma 3 (Non-increasing Total Weight)**.** *As long as $\alpha, \beta > 1$ satisfies $\alpha \leq 2 - 1/\beta$, the total weight $\{W_t\}_t$ in the system is non-increasing.*

Proof of Lemma 3. Let $W_t = \sum_{\sigma \in \mathcal{S}} w^{(t)}(\sigma)$ be the total weight at the start of round t. Define the following disjoint sets of hypotheses:

$$\begin{aligned}
S_1 &= \{\sigma \in \mathcal{S} \mid y_t \in \sigma(x_t),\ \widehat{y}_t \in \sigma(x_t)\},\\
S_2 &= \{\sigma \in \mathcal{S} \mid y_t \in \sigma(x_t),\ \widehat{y}_t \notin \sigma(x_t)\},\\
S_3 &= \{\sigma \in \mathcal{S} \mid y_t \notin \sigma(x_t),\ \widehat{y}_t \in \sigma(x_t)\},\\
S_4 &= \{\sigma \in \mathcal{S} \mid y_t \notin \sigma(x_t),\ \widehat{y}_t \notin \sigma(x_t)\}.
\end{aligned}$$

By the update rule

$$w^{(t+1)}(\sigma) = w^{(t)}(\sigma)\, \alpha^{\mathbf{1}\{\widehat{y}_t \notin \sigma(x_t)\}}\, \beta^{-\mathbf{1}\{y_t \notin \sigma(x_t)\}},$$

we have:

$$w^{(t+1)}(\sigma) = w^{(t)}(\sigma) \begin{cases} 1, & \sigma \in S_1,\\ \alpha, & \sigma \in S_2,\\ 1/\beta, & \sigma \in S_3,\\ \alpha/\beta, & \sigma \in S_4. \end{cases}$$

Therefore:

$$W_{t+1} = W_t(S_1) + \alpha W_t(S_2) + \frac{1}{\beta} W_t(S_3) + \frac{\alpha}{\beta} W_t(S_4)\,.$$

The change in total weight is:

$$W_t - W_{t+1} = (1-\alpha) W_t(S_2) + \left(1 - \frac{1}{\beta}\right) W_t(S_3) + \left(1 - \frac{\alpha}{\beta}\right) W_t(S_4).$$

By the prediction rule:

$$W_t(y_t) \le W_t(\widehat{y}_t) \quad \Rightarrow \quad W_t(S_2) \le W_t(S_3).$$

Thus:

$$W_t - W_{t+1} \ge (1-\alpha)W_t(S_2) + \left(1 - \frac{1}{\beta}\right) W_t(S_2) + \left(1 - \frac{\alpha}{\beta}\right) W_t(S_4).$$

A sufficient condition for $W_{t+1} \le W_t$ is:

$$1 - \alpha + 1 - \frac{1}{\beta} \ge 0 \quad \text{and} \quad 1 - \frac{\alpha}{\beta} \ge 0,$$

which simplifies to:

$$\alpha \le 2 - \frac{1}{\beta} \quad \text{and} \quad \alpha \le \beta.$$

Since $\alpha \le \beta$ is implied by the first condition in relevant regimes, which completes the proof. □

Note that the realizable case is simply $\beta = \infty$ and $\alpha = 2$.

Final Regret Bound We are now ready to bound a certain type of regret of the algorithm.

Theorem 12. *Let $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$ be any finite model class, and let $\{w^{(t)}(\sigma)\}$ be the weights generated by Algorithm 2 when run with $\alpha, \beta > 1$ satisfying $\alpha \le 2 - 1/\beta$. Then for any $\sigma \in \mathcal{S}$, we have*

$$M_T^{\mathrm{alg}}(\sigma) \;\le\; \frac{\log|\mathcal{S}|}{\log(\alpha)} + M_T(\sigma) \cdot \frac{\log(\beta)}{\log(\alpha)}.$$

Proof. Let $W_t = \sum_{\sigma \in \mathcal{S}} w^{(t)}(\sigma)$ be the total weight at round t. By the update rule of Algorithm 2, we have for any $\sigma \in \mathcal{S}$:

$$w^{(T+1)}(\sigma) = \alpha^{M_T^{\mathrm{alg}}(\sigma)} \beta^{-M_T(\sigma)}.$$

Since $\alpha, \beta > 1$ satisfy $\alpha \le 2 - 1/\beta$, Lemma 3 ensures that $\{W_t\}$ is non-increasing, hence:

$$W_{T+1} \le W_1 = |\mathcal{S}|.$$

Therefore,

$$\alpha^{M_T^{\mathrm{alg}}(\sigma)} \beta^{-M_T(\sigma)} = w^{(T+1)}(\sigma) \le W_{T+1} \le |\mathcal{S}|,$$

and taking logarithms yields:

$$M_T^{\mathrm{alg}}(\sigma) \log(\alpha) - M_T(\sigma) \log(\beta) \le \log|\mathcal{S}|,$$

which rearranges to:

$$M_T^{\mathrm{alg}}(\sigma) \le \frac{\log|\mathcal{S}|}{\log(\alpha)} + M_T(\sigma) \cdot \frac{\log(\beta)}{\log(\alpha)}.$$

□

F.2 Online-to-Batch Conversion and Statistical Guarantee

Our learning rule for the statistical setting is again the same as Eq. (5), but by executing the softer Mistake-Unaware-Weight-Update (Algorithm 2) with choice of $(\alpha, \beta) = (4/3, 3/2)$. We again denote this predictor by $\widehat{\pi}_{\mathrm{o2b}}$. We show that Theorem 5 holds when we use this as an estimator.

Proof of Theorem 5. Let $\kappa := \log\beta / \log\alpha \le 1.41$ and consider any fixed but unknown reference σ_* and the demonstrator π_* (which may not be optimal). Define

$$Z_t := L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) - \mathbb{1}\{\widehat{\pi}_t(x_t) \notin \sigma_*(x_t)\} + \kappa \left(L_{\mathcal{D},\sigma_*}(\pi_*) - \mathbb{1}\{y_t \notin \sigma_*(x_t)\}\right),$$

Because $\widehat{\pi}_t$ is deterministic given $S_{<t} = \{(x_i, y_i) : i < t\}$, we have

$$\mathbb{E}[Z_t \mid S_{<t}] = L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) - L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) + \kappa \left(L_{\mathcal{D},\sigma_*}(\pi_*) - L_{\mathcal{D},\sigma_*}(\pi_*)\right) = 0,$$

so $\{Z_t\}_{t=1}^m$ forms a martingale difference sequence. Moreover, $|Z_t| \le 1+\kappa \le 3$ almost surely.

We have

$$\mathbb{E}[Z_t \mid S_{<t}] = 0, \quad \mathrm{Var}(Z_t \mid S_{<t}) \le 2L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) + 2\kappa^2 L_{\mathcal{D},\sigma_*}(\pi_*) \le 2L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) + 4\,L_{\mathcal{D},\sigma_*}(\pi_*).$$

Applying Freedman's inequality (Lemma 2) with $W_m = 2\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) + 4\sum_{t=1}^m L_{\mathcal{D},\sigma}(\pi_*)$ and $\sigma^2 = 6m$ yields that with probability $1-\delta$:

$$\sum_{t=1}^m Z_t \le \sqrt{8\,(1+W_m)\log\left(\frac{2\log 6m}{\delta}\right)} + \frac{4}{3}(1+\kappa)\log\left(\frac{2\log 6m}{\delta}\right).$$

By AM–GM inequality:

$$\sum_{t=1}^m Z_t \le \frac{(1+W_m)}{8} + 16\log\left(\frac{2\log 6m}{\delta}\right) + +4\log\left(\frac{2\log 6m}{\delta}\right).$$

Substituting Z_t and rearranging gives us:

$$\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t) \le \kappa L_{\mathcal{D},\sigma_*}(\pi_*) + 1 + 2\left(\sum_{t=1}^m \mathbf{1}\{\widehat{\pi}_t(x_t) \notin \sigma_*(x_t)\} - \kappa\sum_{t=1}^m \mathbf{1}\{y_t \notin \sigma_*(x_t)\}\right) + 40\log\left(\frac{2\log 6m}{\delta}\right),$$

Finally, using Theorem 12, and using the fact that

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S)) = \frac{1}{m}\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\pi}_t),$$

we obtain that, with probability at least $1-\delta$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}_{\mathrm{o2b}}(S)) \le \kappa\, L_{\mathcal{D},\sigma_*}(\pi_*) + \frac{1 + 6\log_2|\mathcal{S}| + 40\log\left(\frac{2\log 6m}{\delta}\right)}{m}.$$

□

G Algorithms and Proofs for pass@k-Error

In this, we provide our guarantees for pass@k error (and formalize Theorem 6). We first start by describing an online learner.

Algorithm 3 Online pass@k rule with greedy selection and mistake unaware updates

Input: A finite model class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$, and parameter $k \in \mathbb{N}$.

- Initialize $V_1 = \mathcal{S}$ and $w^{(1)}(\sigma) = 1$ for all $\sigma \in \mathcal{S}$.
- In every round, receiving x_t:
  1. For each $y \in \mathcal{Y}$, form the slice $A_y^t = \{\sigma \in V_t : y \in \sigma(x_t)\}$.
  2. *(Greedy top-k selection).* Let $\mathcal{Y}_0 = \emptyset$. For $i = 1, 2, \ldots, k$ set

$$\widehat{y}_t^{(i)} \in \arg\max_{y \in \mathcal{Y}\setminus\mathcal{Y}_{i-1}} w^{(t)}\Big(A_y^t \setminus \bigcup_{z\in\mathcal{Y}_{i-1}} A_z^t\Big), \qquad \mathcal{Y}_i \leftarrow \mathcal{Y}_{i-1} \cup \{\widehat{y}_t^{(i)}\}.$$

     (Break ties arbitrarily.) Define $U_t := \bigcup_{i=1}^k A^t_{\widehat{y}_t^{(i)}}$.
  3. Output the k labels $\widehat{y}_t^{(1)}, \ldots, \widehat{y}_t^{(k)}$.
  4. *(Weight update).* Upon receiving label y_t:

$$w^{(t+1)}(\sigma) \leftarrow \begin{cases} 0\,, & \text{for all } \sigma \notin A^t_{y_t}; \\ w^{(t)}(\sigma) & \text{for all } \sigma \in A^t_{y_t} \cap U_t; \\ (k+1)\,w^{(t)}(\sigma), & \text{for all } \sigma \in A^t_{y_t} \setminus U_t. \end{cases}$$

     So the version space: $V_{t+1} \leftarrow A^t_{y_t}$.

Theorem 13 (Online pass@k guarantee). *On any sequence $((x_t, y_t))_{t\in\mathbb{N}}$ realizable by some $\sigma_* \in \mathcal{S}$, Algorithm 3 makes at most $\log_{k+1}|\mathcal{S}|$ mistakes (i.e., rounds with $\{\widehat{y}_t^{(1)}, \dots, \widehat{y}_t^{(k)}\} \cap \sigma_*(x_t) = \emptyset$).*

Proof. Let V_t be the version space at the start of round t, and as in the algorithm write $A_y^t = \{\sigma \in V_t : y \in \sigma(x_t)\}$ and $U_t := \bigcup_{i=1}^k A^t_{\widehat{y}_t^{(i)}}$. Define the potential $W_t := w^{(t)}(V_t) = \sum_{\sigma\in V_t} w^{(t)}(\sigma)$, then we again have $\{W_t\}_t$ is non-increasing.

$$\begin{aligned} W_{t+1} = (k+1)w^{(t)}(A^t_{y_t} \setminus U_t) + w^{(t)}(A^t_{y_t}) &= k\, w^{(t)}(A^t_{y_t} \setminus U_t) + w^{(t)}(A^t_{y_t} \setminus U_t) + w^{(t)}(A^t_{y_t} \cap U_t) \\ &\le w^{(t)}(U_t \setminus A^t_{y_t}) + w^{(t)}(A^t_{y_t} \setminus U_t) + w^{(t)}(A^t_{y_t} \cap U_t) \\ &= w^{(t)}(U_t \cup A^t_{y_t}) \le W_t\,, \end{aligned}$$

where in the first inequality arises from the following key relation:

$$w^{(t)}(U_t \setminus A^t_{y_t}) \ge k\, w^{(t)}(A^t_{y_t} \setminus U_t). \tag{13}$$

We defer its proof to the end of this section.

Now suppose the algorithm makes M pass@k mistakes by the end of round t. On each mistake round we must have $\sigma_* \in A^t_{y_t} \setminus U_t$, so its weight is multiplied by $(k+1)$. Therefore

$$w^{(t+1)}(\sigma_*) = (k+1)^M \;\le\; W_{t+1} \;\le\; W_1 = |\mathcal{S}|,$$

which yields $M \le \log_{k+1}|\mathcal{S}|$. □

Proof of the key inequality (13). To simplify the notation, we remove the time index t and write U_t as U, and $A^t_{y_t}$ as A_{y_t}. Define the uncovered mass in A after selecting first i labels greedily as:

$$a_i := w^{(t)}\Big(A_{y_t} \setminus \bigcup_{z\in\mathcal{Y}_i} A_z\Big) \quad \text{so that} \quad a_0 = w^{(t)}(A_{y_t}),\;\; a_k = w^{(t)}(A_{y_t} \setminus U),$$

and $a_0 \ge a_1 \ge \cdots \ge a_k$. Correspondingly, define

$$s_i := a_{i-1} - a_i \;=\; w^{(t)}\Big(\big(A_{\widehat{y}_t^{(i)}} \cap A_{y_t}\big) \setminus \bigcup_{z\in\mathcal{Y}_{i-1}} A_z\Big).$$

In addition, we define the uncovered weight for which $\widehat{y}_t^{(i)}$ got picked as

$$m_i := w^{(t)}\Big(A_{\widehat{y}_t^{(i)}} \setminus \bigcup_{z\in\mathcal{Y}_{i-1}} A_z\Big).$$

By the design of the greedy selections, we have

$$m_i \;\ge\; a_{i-1} \qquad \text{for all } i \in [k]. \tag{14}$$

With these definitions and relations in place, we are ready to prove the desired inequality. Basic set inclusion / exclusion tells us that

$$\begin{aligned} w^{(t)}(U \setminus A_{y_t}) &= \sum_{i=1}^k w^{(t)}(A_{\widehat{y}_t^{(i)}} \setminus A_{y_t} \cup A_{\widehat{y}_t^{(1)}} \cdots \cup A_{\widehat{y}_t^{(i-1)}}) \\ &= \sum_{i=1}^k \left\{ w^{(t)}(A_{\widehat{y}_t^{(i)}} \setminus \bigcup_{z\in\mathcal{Y}_{i-1}} A_z) - w^{(t)}\Big(\big(A_{\widehat{y}_t^{(i)}} \cap A_{y_t}\big) \setminus \bigcup_{z\in\mathcal{Y}_{i-1}} A_z\Big) \right\} \\ &= \sum_{i=1}^k (m_i - s_i), \end{aligned}$$

where the last equality uses the definitions of m_i and s_i. Using (14), we obtain

$$m_i - s_i \;\ge\; a_{i-1} - (a_{i-1} - a_i) \;=\; a_i,$$

which allows us to further lower bound $w^{(t)}(U \setminus A)$ as

$$w^{(t)}(U \setminus A_{y_t}) \;\ge\; \sum_{i=1}^k a_i \;\ge\; k\, a_k.$$

Here we use the fact that $(a_i)_i$ is non-increasing. This proves the claim.

G.1 Statistical Upper Bound

Input: Sample $S = \{(x_i, y_i) : i \in [m]\}$ and a finite model class $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$.

- Run Algorithm 3 once over S, recording $(V_t, w^{(t)})$ at the *start* of each round $t \in [m]$.
- Find the deterministic predictor $\widehat{\mu}_t : \mathcal{X} \to \mathcal{Y}^k$ used by the online algorithm from the snapshot. I.e. for any $x \in \mathcal{X}$ and any $t \in [m]$, define slices (with respect to V_t)

$$A_y^t(x) \;:=\; \{\sigma \in V_t :\; y \in \sigma(x)\}\,.$$

  And greedily pick top k labels according to the rule described in Algorithm 3. Let $\mathcal{Y}_0(x) = \emptyset$. For $i = 1, \ldots, k$ set

$$\widehat{y}_t^{(i)}(x) \;\in\; \arg\max_{y \in \mathcal{Y} \setminus \mathcal{Y}_{i-1}(x)} w^{(t)}\Big(A_y^t(x) \setminus \bigcup_{z \in \mathcal{Y}_{i-1}(x)} A_z^t(x)\Big)$$

$$\mathcal{Y}_i(x) \leftarrow \mathcal{Y}_{i-1}(x) \cup \{\widehat{y}_t^{(i)}(x)\},$$

  breaking ties by arbitrary fixed rule.
- Then the deterministic predictor $\widehat{\mu}_t : \mathcal{X} \to \mathcal{Y}^k$ is given by:

$$\widehat{\mu}_t(x) \;:=\; (\widehat{y}_t^{(1)}(x), \ldots, \widehat{y}_t^{(k)}(x)).$$

- **Final batch predictor.** On a test input x, draw $I \sim \mathrm{Unif}\{1, \ldots, m\}$ and output

$$\widehat{\mu}_{\mathrm{o2b}}(S)(x) \;:=\; \widehat{\mu}_I(x). \tag{15}$$

  (Equivalently: $\widehat{\mu}_{\mathrm{o2b}}$ is the uniform mixture over $\{\widehat{\mu}_t\}_{t=1}^m$ for every test point.)

Below is the statistical guarantee for this estimator in similar spirit to Theorem 4.

Theorem 14 (Statistical Guarantee for pass@k)**.** *For any finite model class $\mathcal{S}$, the estimator $\widehat{\mu}_{\mathrm{o2b}}$ in Eq. (15) achieves the following guarantee for any joint distribution $(\mathcal{D} \times \pi_*)$ supported on $\sigma_* \in \mathcal{S}$:*

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D},\sigma_*}\big(\widehat{\mu}_{\mathrm{o2b}}(S)\big)\right] \;\le\; \frac{\log_{k+1}|\mathcal{S}|}{m}\,,$$

and, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$L_{\mathcal{D},\sigma_*}\big(\widehat{\mu}_{\mathrm{o2b}}(S)\big) \;\le\; \frac{1 + 2\log_{k+1}|\mathcal{S}| + 12\log\Big(\frac{\log m}{\delta}\Big)}{m}\,.$$

This implies that $\mathcal{S}$ is learnable (cf. Definition 1) using the estimator $\widehat{\mu}_{\mathrm{o2b}}$ with sample complexity

$$m_{\mathcal{S},\widehat{\mu}_{\mathrm{o2b}}}(\varepsilon,\delta) = O\left(\frac{1}{\varepsilon}\Big(\log_{k+1}|\mathcal{S}| + \log\big(\tfrac{1}{\varepsilon\delta}\big)\Big)\right).$$

Proof of Theorem 14. The proof is exactly similar to that of Theorem 4 and given for completeness. Let $\ell_t = \mathbb{1}\{\widehat{y}_t^{(i)}(x_t) \notin \sigma_*(x_t), \forall\, \widehat{y}_t^{(i)}(x_t) \in \widehat{\mu}_t(x_t)\}$. Because $\widehat{\mu}_t$ is a deterministic function of $S_{<t} = \{(x_i, y_i) : i < t\}$, we have

$$\mathbb{E}[\ell_t \mid S_{<t}] = L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\,.$$

Hence

$$\mathbb{E}_S\left[L_{\mathcal{D},\sigma_*}(\widehat{\mu}_{\mathrm{o2b}}(S))\right] = \mathbb{E}_S\left[\frac{1}{m}\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\right] = \mathbb{E}_S\left[\frac{1}{m}\sum_{t=1}^m \ell_t\right] \le \frac{\log_{k+1}|\mathcal{S}|}{m}\,,$$

where in the last inequality we used Theorem 13, which guarantees $\sum_{t=1}^m \ell_t \le \log_{k+1}|\mathcal{S}|$.

For the high-probability statement, define the martingale differences

$$Z_t := L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t) - \ell_t, \qquad \text{where } |Z_t| \le 1 \text{ almost surely.}$$

Then $\mathbb{E}[Z_t \mid S_{<t}] = 0$, and

$$\mathbb{E}[Z_t^2 \mid S_{<t}] = \mathbb{E}\big[(L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t) - \ell_t)^2 \mid S_{<t}\big] = \mathrm{Var}(\ell_t \mid S_{<t}) = L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\big(1 - L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\big) \;\le\; L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\,.$$

Taking $W_m = \sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)$ and $\sigma^2 = m$ suffices; thus, using Lemma 2 with $n = \log m$ gives, with probability $1-\delta$,

$$\begin{aligned}\sum_{t=1}^m Z_t &\le \sqrt{8\left(1 + \sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\right)\log\left(\frac{\log m}{\delta}\right)} + \frac{4}{3}\log\left(\frac{\log m}{\delta}\right) \\ &\le \frac{1}{2}\left(1 + \sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)\right) + 4\log\left(\frac{\log m}{\delta}\right) + \frac{4}{3}\log\left(\frac{\log m}{\delta}\right) \qquad \text{(GM} \le \text{AM)}\end{aligned}$$

Substituting Z_t and rearranging terms,

$$\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t) \;\le\; 1 + 2\sum_{t=1}^m \ell_t + \frac{32}{3}\log\left(\frac{\log m}{\delta}\right).$$

Finally, noting that $L_{\mathcal{D},\sigma_*}(\widehat{\mu}_{\mathrm{o2b}}(S)) = \frac{1}{m}\sum_{t=1}^m L_{\mathcal{D},\sigma_*}(\widehat{\mu}_t)$ and that $\sum_{t=1}^m \ell_t \le \log_{k+1}|\mathcal{S}|$ (by Theorem 13), we obtain that with probability $1-\delta$,

$$L_{\mathcal{D},\sigma_*}(\widehat{\mu}_{\mathrm{o2b}}(S)) \le \frac{1 + 2\log_{k+1}|\mathcal{S}| + 12\log\left(\frac{\log m}{\delta}\right)}{m}.$$

□

G.2 Lower Bounds for Online and Statistical Settings for k-pass Error

We next provide a lower bound that, information-theoretically, this dependence cannot be improved and we only gain a factor of $1/\log k$ in sample complexity as well as mistake bound in the worst-case.

Theorem 15 (Online $\Omega(\log_{k+1}|\mathcal{S}|)$ pass@k mistake bound even for multiclass classification)**.** *Fix integers $k \ge 1$ and $d \ge 2$. There exists a problem instance $\mathcal{S} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $|\mathcal{S}| \le d$, $|\mathcal{Y}| = k+1$, $|\mathcal{X}| = \lfloor \log_{k+1} d \rfloor$ such that for any deterministic online learning algorithm that outputs at most k labels, there exists a sequence $(x_t, y_t)_{t\in[|\mathcal{X}|]}$ realizable by some $\sigma_* \in \mathcal{S}$ such that it makes mistake on every round.*

Note that our instance is an instance of multiclass classification problem $\sigma : \mathcal{X} \to \mathcal{Y}$. This is isomorphic to an instance $\mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$, where $|\sigma(x)| = 1$ for all $x \in \mathcal{X}, \sigma \in \mathcal{S}$.

Proof of Theorem 15. Let $m := \lfloor \log_{k+1} d \rfloor$ and take $\mathcal{X} = \{1, \dots, m\}$ and $\mathcal{Y} = \{1, \dots, k+1\}$. Consider the full product class $\mathcal{S} = \mathcal{Y}^{\mathcal{X}}$, which has size $(k+1)^m \le d$. First of all, observe that in any round in which y_t does not belong to the list of $(\widehat{y}_t^{(1)}, \dots, \widehat{y}_t^{(k)})$, the mistake is made because we are in the multiclass classification setting.

For rounds $t \in [m]$, present a fresh coordinate $x_t = t$. Since $|\mathcal{Y}| = k+1$, there exists a label $y_t \in \mathcal{Y}$ that the learner failed to output in the set; $y_t \ne \widehat{y}_t^{(i)}$ for all $i \in [k]$. Reveal this y_t. This forces a mistake on every round. Moreover, this sequence is realizable since $\mathcal{S} = \mathcal{Y}^{\mathcal{X}}$ contains all functions from $\mathcal{X}$ to $\mathcal{Y}$. □

Theorem 16 (Statistical lower bound of $\Omega(\log_k|\mathcal{S}|)$)**.** *Fix integers $k \ge 1$ and $q \ge 1$. Let $\mathcal{X} = \{1, \dots, q\}$, $\mathcal{Y} = \{1, \dots, 2k\}$, and take the hypothesis class $\mathcal{S} = \mathcal{Y}^{\mathcal{X}}$ (all multiclass functions), so its cardinality is $d := |\mathcal{S}| = (2k)^q$. Let $\mathcal{D}$ be the uniform distribution on $\mathcal{X}$. Then for any estimators $\widehat{\mu} : (\mathcal{X}\times\mathcal{Y})^* \to \Delta(\mathcal{Y}^k)^{\mathcal{X}}$*

$$\inf_{\widehat{\mu}} \sup_{\sigma\in\mathcal{S}} \; \mathbb{E}_{S\sim(\mathcal{D}\times\sigma)^m}\, \mathbb{P}_{x\sim\mathcal{D}} \mathbb{P}_{\widehat{\boldsymbol{y}}(x)\sim\widehat{\mu}(\cdot|x)}\big[\,\sigma(x) \notin \widehat{\boldsymbol{y}}(x)\,\big] \;\;\ge\;\; \tfrac{1}{2}\Big(1 - \tfrac{1}{q}\Big)^m.$$

In particular, to ensure expected error at most $0 < \varepsilon < \frac{1}{2}$ for all $\sigma \in \mathcal{S}$, one needs

$$m \;\ge\; \frac{\ln(1/(2\varepsilon))}{-\ln(1-1/q)} \;\;\ge\;\; q\,\ln\Big(\frac{1}{2\varepsilon}\Big).$$

Proof. Fix any (possibly randomized) estimator $\widehat{\mu}$. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be the training sample drawn i.i.d. from $(\mathcal{D} \times \sigma)$ for $\sigma \sim \mathrm{Unif}(\mathcal{S})$, and let $U_S = \{x_i : 1 \le i \le m\} \subseteq \mathcal{X}$ be the set of distinct inputs seen in S. Draw $x \sim \mathcal{D}$ independently of S and then $\widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(S)(\cdot \mid x)$.

On any $x \notin U_S$, under the prior where σ is uniform over $\mathcal{S}$, for any (possibly randomized) k-list $\widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(\cdot \mid x)$,

$$\mathbb{P}_\sigma\big[\sigma(x) \in \widehat{\boldsymbol{y}}(x) \mid S,\, x \notin U_S\big] = \mathbb{E}\left[\frac{\#\text{ of distinct labels in } \widehat{\boldsymbol{y}}(x)}{|\mathcal{Y}|} \;\middle|\; S,\, x \notin U_S\right] \le \frac{k}{2k} = \frac{1}{2},$$

so $\mathbb{P}_\sigma[\sigma(x) \notin \widehat{\boldsymbol{y}}(x) \mid S,\, x \notin U_S] \ge \frac{1}{2}$. (Allowing duplicates in $\widehat{\boldsymbol{y}}(x)$ cannot decrease this probability.)

If $x \in U_S$, the learner can always include the observed label and incur zero error on that x. Therefore, for any estimator $\widehat{\mu}$,

$$\mathbb{P}_{\sigma, x, \widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(\cdot|x)}\big[\sigma(x) \notin \widehat{\boldsymbol{y}}(x) \mid S\big] \;\ge\; \frac{1}{2} \cdot \mathbb{P}[x \notin U_S].$$

Taking expectation over S and using $\mathcal{D} = \mathrm{Unif}(\mathcal{X})$ yields

$$\mathbb{E}_S\, \mathbb{P}_{\sigma, x, \widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(\cdot|x)}\big[\sigma(x) \notin \widehat{\boldsymbol{y}}(x)\big] \;\ge\; \frac{1}{2}\, \mathbb{E}_S\big[1 - |U_S|/q\big] \;=\; \tfrac{1}{2}\Big(1 - \tfrac{1}{q}\Big)^m,$$

since $\mathbb{E}[|U_S|] = q\big(1 - (1 - \frac{1}{q})^m\big)$. Finally, by minimax principle

$$\inf_{\widehat{\mu}} \sup_{\sigma \in \mathcal{S}} \mathbb{E}_S\, \mathbb{P}_{x \sim \mathcal{D}} \mathbb{P}_{\widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(\cdot|x)}\big[\,\sigma(x) \notin \widehat{\boldsymbol{y}}(x)\,\big] \ge \inf_{\widehat{\mu}}\; \mathbb{E}_{\sigma \sim \mathrm{Unif}(\mathcal{S})}\, \mathbb{E}_S\, \mathbb{P}_{x \sim \mathcal{D}} \mathbb{P}_{\widehat{\boldsymbol{y}}(x) \sim \widehat{\mu}(\cdot|x)}\big[\,\sigma(x) \notin \widehat{\boldsymbol{y}}(x)\,\big]$$
$$\ge\; \tfrac{1}{2}\Big(1 - \tfrac{1}{q}\Big)^m.$$

For the sample-complexity bound, solve $\frac{1}{2}(1 - \frac{1}{q})^m \le \varepsilon$ for m and use $-\ln(1 - 1/q) \le 1/q$. □