THE STATISTICAL BENEFITS OF MULTIPLE RESPONSES FOR LEARNING FROM DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many generative systems return multiple candidate responses and are evaluated according to the best one. Recent work shows that, when demonstrations are optimal, pass@k can reduce the sample complexity of learning from demonstrations by a logarithmic factor in k. We ask what happens when the demonstrator is not assumed to be optimal. We find that multiple responses provide a qualitatively stronger benefit in this setting. In a finite reward-class model with no reward feedback, moving from pass@1 to any pass@k with $k \geq 2$ changes the worst-case dependence on target accuracy from $1/\varepsilon^2$ to $1/\varepsilon$, uniformly over demonstrator quality. Under standard evaluation, where an unknown reward is fixed before training, increasing k provides an additional and distinct benefit: the optimal dependence on a reward class of size N improves from $\log N$ to $\log N/\log k$. We further show that these two effects can be separated. Under robust evaluation, where one learned policy must compete with the demonstrator simultaneously for every reward in the class, the fast $1/\varepsilon$ dependence persists, while the $1/\log k$ improvement can disappear. We establish matching upper and lower bounds in the corresponding regimes and give a greedy multiplicative-weights learner achieving the upper bounds without any assumption on demonstrator quality.

1 INTRODUCTION

Learning from demonstrations (Argall et al., 2009; Ouyang et al., 2022; Joshi et al., 2026) is a common paradigm for training modern predictive and generative systems: for each input or context, the learner observes a response produced by a human, an expert, or another model, and seeks to perform comparably well without direct access to the underlying reward that makes those responses desirable. This problem falls within apprenticeship learning (Abbeel & Ng, 2004; Syed & Schapire, 2007), a branch of imitation learning (Pomerleau, 1988; Ng et al., 2000) that studies how to match a demonstrator's value under an unknown reward. Most of the learning-theoretic literature has focused on the usual pass@1 setting, where performance is determined by a single returned response. Yet modern generative systems are often evaluated using pass@k: a model produces several candidate answers, programs, or solutions, and performance is determined by the best one (Chen et al., 2021; Li et al., 2022; Brown et al., 2024). Recent work also directly optimizes this pass@k metric during reinforcement learning (Tang et al., 2025; Walder & Karkhanis, 2026; Chen et al., 2025; Bagirov et al., 2025). Despite its practical relevance, the statistical effect of moving from pass@1 to pass@k remains much less understood.

Recent work by Joshi et al. (2026) initiated the learning-theoretic study of pass@k in this apprenticeship-learning framework. They considered the favorable setting of optimal demonstrations and showed that, for a reward class of cardinality N, pass@k improves the dependence of the sample complexity on N from order $\log N$ to order $\log N/\log k$, with a matching lower bound. In practice, however, demonstrations need not be optimal: human demonstrators may have limited expertise or make mistakes, and demonstrations generated by other models can themselves be imperfect. More generally, high-quality demonstrations need not maximize the reward ultimately used for evaluation. Without assuming that the demonstrator is optimal, we ask how allowing the learner to return multiple responses rather than a single response changes the sample complexity of learning from demonstrations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this question, we consider a finite reward-class model in which the learner observes m context–response pairs generated by an unknown demonstrator π_{demo} and knows a finite class $\mathcal{R}$ of bounded reward functions, but receives no reward feedback. At test time, a pass@k policy returns k possibly dependent responses and is evaluated according to the highest reward among them. We compare its value with the demonstrator's single-response value.

We study two natural evaluation criteria. The first is the *standard evaluation* setting of Joshi et al. (2026), in which an unknown reward $r^\star \in \mathcal{R}$ is fixed before training and performance is measured under that reward. Secondly, when several reward functions are plausible descriptions of the desired behavior, it may be undesirable to tie the guarantee to one particular reward. This motivates the stronger *robust evaluation* criterion, under which the learned policy must compete with the demonstrator simultaneously for every reward in $\mathcal{R}$ (Syed & Schapire, 2007). Throughout the introduction, we focus on arbitrary demonstrators, without any optimality or near-optimality assumption.[1]

Standard evaluation. Under standard evaluation, we obtain a sharp characterization for every $k \geq 2$. Write $N := |\mathcal{R}|$, and let ε and δ denote the target accuracy and failure probability, respectively. The worst-case sample complexity scales as

$$\frac{1}{\varepsilon}\left(\frac{\log N}{\log k} + \log\frac{1}{\delta}\right).$$

The upper bound holds for arbitrary demonstrators, while the matching lower bound already holds when the demonstrator is optimal for the selected evaluation reward.

Comparing this result with pass@1 reveals two statistically distinct benefits of multiple responses. For arbitrary demonstrators, the pass@1 upper bound of Joshi et al. (2026) has order

$$\frac{1}{\varepsilon^2}\left(\log N + \log\frac{1}{\delta}\right),$$

and we show that this dependence is unimprovable in general. Hence the first benefit is a qualitative slow-to-fast transition: moving from one response to any $k \geq 2$ changes the worst-case dependence on accuracy from $1/\varepsilon^2$ to $1/\varepsilon$. The mechanism is simple: when two responses are separated only by a small bias in the demonstrations, pass@1 must estimate which response is better, whereas pass@2 can cover both possibilities. Multiple responses therefore turn an estimation problem into a coverage problem. The second benefit concerns the dependence on the reward-class size. Joshi et al. (2026) showed, for optimal demonstrations, that increasing k improves this dependence from $\log N$ to $\log N/\log k$. Our result shows that this improvement persists even for arbitrarily suboptimal demonstrations.

Robust evaluation. The picture changes under robust evaluation. For every $k \geq 2$, the same learner achieves, for arbitrary demonstrators, sample complexity

$$\frac{1}{\varepsilon}\left(\log N + \log\frac{1}{\delta}\right),$$

up to universal constants, and we prove a matching lower bound in the high-accuracy regime, i.e., when $\varepsilon \ll 1/k$. Thus the slow-to-fast transition from pass@1 to pass@k survives the stronger robust guarantee. In contrast, the additional $1/\log k$ improvement in the reward-class dependence can disappear: in this regime, a full $\log N$ dependence is unavoidable regardless of k. Table 1 summarizes the comparison of optimal sample complexities for arbitrary demonstrators.

1.1 Related work

Learning from demonstrations. Apprenticeship learning aims to match a demonstrator's value under an unknown reward (Abbeel & Ng, 2004; Syed & Schapire, 2007). Behavior-cloning analyses often assume a restricted policy class (Foster et al., 2024); our learner instead uses a known reward class and does not fit the demonstrator's response distribution. Closest to our setting, the pass@k guarantees of Joshi et al. (2026) assume optimal demonstrations. We remove this assumption while

[1] In the formal development, this corresponds to the unrestricted suboptimality budget $\Delta = 1$. Section 2 gives results for general $\Delta \in [0, 1]$.

Table 1: Sample complexity for arbitrary demonstrators, up to universal constants. See Section 2 and Appendix C for precise statements.

	Standard evaluation	**Robust evaluation**
pass@1	$\dfrac{\log N + \log(1/\delta)}{\varepsilon^2}$	$\dfrac{\log N + \log(1/\delta)}{\varepsilon^2}$
pass@k, $k \geq 2$	$\dfrac{1}{\varepsilon}\left(\dfrac{\log N}{\log k} + \log\dfrac{1}{\delta}\right)$	$\dfrac{\log N + \log(1/\delta)}{\varepsilon}$

retaining fast $1/\varepsilon$ sample complexity for every $k \geq 2$, and also give guarantees that hold simultaneously over the reward class. Pour et al. (2026) characterize realizable online learning with multiple correct answers under different feedback models. Their realizable setting with one correct response and no feedback on the learner's correctness is close to our setting with binary rewards and optimal demonstrations. They study single-response prediction and use a different agnostic benchmark from our comparison with the demonstrator's value.

List learning. PAC and online list learning allow several predicted labels and count an error when the target label is absent (Charikar & Pabbaraju, 2023; Moran et al., 2023). PAC learning uses i.i.d. examples, while online learning receives label feedback after each prediction. Our pass@k objective also evaluates several responses, but compares their best reward with the demonstrator's value. Several responses may be acceptable, rewards need not be binary, and demonstrations may be suboptimal. Thus a successful prediction need not contain the demonstrated response.

Inference-time selection. Recent work studies how to select candidates from a given generator using an estimated reward model (Huang et al., 2025; Di et al., 2026). Their guarantees depend on the generator's ability to produce good responses, the accuracy of reward estimates, and the number of candidates sampled. We instead learn the policy that generates the candidates, using demonstrations and a known reward class. The learner receives no reward feedback during training. Our guarantees concern the number of demonstrations needed to match the demonstrator's value under pass@k evaluation.t

2 Problem Setup and Main Results

2.1 Setup and evaluation criteria

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the context and response spaces, and let $\mathcal{D} \in \Delta(\mathcal{X})$ be an unknown context distribution, where $\Delta(\mathcal{Z})$ denotes the set of probability distributions on $\mathcal{Z}$. The learner knows a finite, non-empty reward class $\mathcal{R} \subseteq [0,1]^{\mathcal{X}\times\mathcal{Y}}$, but observes neither the evaluation reward nor reward feedback. An unknown demonstrator $\pi_{\mathrm{demo}} : \mathcal{X} \to \Delta(\mathcal{Y})$ generates m i.i.d. demonstrations

$$S = ((x_1, y_1), \ldots, (x_m, y_m)), \qquad x_i \sim \mathcal{D}, \quad y_i \sim \pi_{\mathrm{demo}}(\cdot \mid x_i).$$

A possibly randomized learning rule $\mathcal{A}_k(\mathcal{R}, S)$ returns a pass@k policy $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S) : \mathcal{X} \to \Delta(\mathcal{Y}^k)$, whose k responses may be dependent. For $r \in \mathcal{R}$, define

$$V_r^k(\pi) := \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{(y^{(1)},\ldots,y^{(k)})\sim\pi(\cdot|x)}\left[\max_{\ell\in[k]} r(x, y^{(\ell)})\right],$$

and write $V_r := V_r^1$. We measure performance relative to the demonstrator through

$$\mathcal{E}_r^k(\pi, \pi_{\mathrm{demo}}) := V_r(\pi_{\mathrm{demo}}) - V_r^k(\pi).$$

The excess value may be negative when the learned policy outperforms the demonstrator.

We study two evaluation modes: standard evaluation at a fixed but unknown reward, and robust evaluation simultaneously over the reward class.

Standard evaluation. Under standard evaluation, a target reward $r^\star \in \mathcal{R}$ is fixed before training but is not revealed to the learner. The goal is to make $\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}})$ small.

The demonstrator need not be optimal for $r^\star$. Define the optimal value under reward r by

$$V_r^\star := \mathbb{E}_{x\sim\mathcal{D}}\left[\sup_{y\in\mathcal{Y}} r(x,y)\right],$$

and define the demonstrator's suboptimality under r as

$$\Delta_r := V_r^\star - V_r(\pi_{\mathrm{demo}}).$$

We use $\Delta \in [0,1]$ as a budget on this quantity. Thus $\Delta = 0$ restricts attention to demonstrators that are optimal for the target reward, whereas $\Delta = 1$ places no restriction on the demonstrator.

Definition 1 (Standard sample complexity)**.** For $\varepsilon, \delta \in (0,1)$ and $\Delta \in [0,1]$, the standard sample complexity $m^{(k)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta)$ is the smallest nonnegative integer m_0 for which there exists a learning rule such that, for every $m \geq m_0$, every context distribution $\mathcal{D}$, every target reward $r^\star \in \mathcal{R}$, and every demonstrator π_{demo} satisfying $\Delta_{r^\star} \leq \Delta$, the learned policy satisfies

$$\mathbb{P}\big\{\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}}) > \varepsilon\big\} \leq \delta.$$

Robust evaluation. Under robust evaluation, a single learned policy must compete with the demonstrator simultaneously for every reward in $\mathcal{R}$. Accordingly, the relevant notion of demonstrator quality is its worst-case suboptimality over the reward class,

$$\Delta_{\mathcal{R}} := \sup_{r\in\mathcal{R}} \Delta_r = \sup_{r\in\mathcal{R}} \big[V_r^\star - V_r(\pi_{\mathrm{demo}})\big].$$

Definition 2 (Robust sample complexity)**.** For $\varepsilon, \delta \in (0,1)$ and $\Delta \in [0,1]$, the robust sample complexity $m^{(k)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, \Delta)$ is the smallest nonnegative integer m_0 for which there exists a learning rule such that, for every $m \geq m_0$, every context distribution $\mathcal{D}$, and every demonstrator π_{demo} satisfying $\Delta_{\mathcal{R}} \leq \Delta$, the learned policy satisfies

$$\mathbb{P}\left\{\sup_{r\in\mathcal{R}} \mathcal{E}^k_r(\widehat{\pi}, \pi_{\mathrm{demo}}) > \varepsilon\right\} \leq \delta.$$

Both sample complexities above are defined for a fixed reward class $\mathcal{R}$. Our upper bounds hold for every finite $\mathcal{R}$, whereas the lower bounds below show that, for each prescribed cardinality N, there exist reward classes of size N (or at most N) with the stated sample complexity. We first study standard evaluation and then turn to robust evaluation to determine which benefits of multiple responses survive the stronger requirement of simultaneous validity.

2.2 Standard evaluation

Under standard evaluation, the learner needs to compete with the demonstrator under a single reward fixed before training, although that reward is never revealed. We first show that, once $k \geq 2$, the sample complexity no longer depends on the quality of the demonstrator.

Pass@k achieves a fast rate for arbitrary demonstrators. Our first result gives a high-probability guarantee that holds without any assumption on demonstrator quality. The learning rule achieving this guarantee is constructed and analyzed in Section 3.

Theorem 3 (Standard pass@k upper bound)**.** *For every $k \geq 2$, there exists a learning rule $\mathcal{A}_k$ with the following guarantee. For every $m \geq 1$, $\delta \in (0,1)$, context distribution $\mathcal{D}$, demonstrator π_{demo}, and evaluation reward $r^\star \in \mathcal{R}$ fixed before sampling, the policy $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S)$ satisfies*

$$\mathbb{P}_S\left\{\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{3\log|\mathcal{R}|}{m\log k} + \frac{3\log(1/\delta)+1}{m}\right\} \leq \delta.$$

Consequently, for every $\varepsilon > 0$, $\delta \in (0,1)$, and $\Delta \in [0,1]$,

$$m^{(k)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \leq \left\lceil \frac{3\log|\mathcal{R}|/\log k + 3\log(1/\delta)+1}{\varepsilon} \right\rceil.$$

Two features of the bound are worth separating. First, its $1/\varepsilon$ dependence is independent of Δ: no additional samples are required when the demonstrator is imperfect. Second, increasing k improves the dependence on the reward-class size from $\log|\mathcal{R}|$ to $\log|\mathcal{R}|/\log k$, while leaving the confidence term unchanged.

The upper bound leaves two questions: is the independence from demonstrator quality genuine, and is the $1/\log k$ dependence optimal? The next result answers both affirmatively. In fact, the matching lower bound already holds for demonstrators that are optimal for the evaluation reward.

Theorem 4 (Standard pass@k lower bound)**.** *There exist universal constants $c, c_0 > 0$ such that the following holds. For every pair of integers $k \geq 2$ and $N \geq 2k$, there exist finite spaces $\mathcal{X}, \mathcal{Y}$ and a binary reward class $\mathcal{R} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{Y}}$ with $|\mathcal{R}| = N$ such that, for every $\Delta \in [0,1]$, $0 < \varepsilon \leq c_0$, and $0 < \delta < 1/16$,*

$$m^{(k)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \geq \frac{c}{\varepsilon}\left(\frac{\log N}{\log k} + \log\frac{1}{\delta}\right).$$

The lower bound holds even when the demonstrator is optimal for the selected evaluation reward.

Together, Theorems 3 and 4 establish the optimal dependence on reward-class size, accuracy, and confidence in the stated parameter range. The matching bounds show that neither the fast rate nor the $1/\log k$ improvement relies on favorable demonstrations. The upper bound holds for arbitrary demonstrators, while the lower bound already holds for optimal ones. Thus, once multiple responses are allowed, demonstrator suboptimality creates no additional statistical cost. The proofs of Theorem 3 and Theorem 4 are given in the Appendices A.1 and A.2 respectively.

Comparison with pass@1. To isolate what is gained by allowing multiple responses, we now compare the preceding rate with the single-response problem. A result of Joshi et al. (2026) gives

$$m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \lesssim \frac{\log|\mathcal{R}| + \log(1/\delta)}{\varepsilon} + \frac{\Delta\,[\log|\mathcal{R}| + \log(1/\delta)]}{\varepsilon^2}.$$

For an optimal demonstrator ($\Delta = 0$), this bound has the same fast $1/\varepsilon$ dependence as pass@k. Positive suboptimality, however, introduces a potentially dominant $1/\varepsilon^2$ term. The following lower bound shows that this deterioration is intrinsic rather than an artifact of the analysis.

Theorem 5 (Standard pass@1 lower bound)**.** *There exists a universal constant $c > 0$ such that the following holds. For every integer $N \geq 2$, $\Delta \in [0,1]$, $0 < \varepsilon \leq 1/64$, and $0 < \delta \leq 1/64$, there exist finite spaces $\mathcal{X}, \mathcal{Y}$ and a finite binary reward class $\mathcal{R} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{Y}}$ with $|\mathcal{R}| \leq N$ such that*

$$m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \geq c\left[\frac{\log N + \log(1/\delta)}{\varepsilon} + \frac{\Delta}{\varepsilon^2}\left\{\log\frac{1}{\delta} + \log\left(1 + \frac{N\varepsilon^2}{(\Delta+\varepsilon)^2}\right)\right\}\right].$$

At $\Delta = 0$, the lower bound has the fast order $[\log N + \log(1/\delta)]/\varepsilon$. For positive suboptimality budgets, it also captures a slow contribution proportional to Δ/ε^2. In particular, if

$$N \geq (1 + \Delta/\varepsilon)^4 \qquad \text{or} \qquad \log(1/\delta) \geq \log N,$$

the lower bound matches the suboptimality-dependent upper bound up to universal constants. See the proof in Appendix A.3.

The preceding bounds make precise the two benefits of multiple responses highlighted in the introduction. First, there is a qualitative slow-to-fast transition at $k = 2$: for arbitrary demonstrators, pass@1 can require $1/\varepsilon^2$ samples, whereas every pass@k with $k \geq 2$ achieves a $1/\varepsilon$ rate. Second, within the regime $k \geq 2$, increasing k further improves the dependence on the reward-class size from $\log N$ to $\log N/\log k$, extending the optimal-demonstrator phenomenon of Joshi et al. (2026) to arbitrarily suboptimal demonstrations.

The first transition reflects a change in the underlying statistical task. This can already be seen with a single context and two responses a and b. Consider two possible instances: in the first, only a receives reward one and the demonstrator outputs a with probability $1/2 + 2\varepsilon$; in the second, only b receives reward one and the demonstrator outputs b with the same probability. A pass@1 learner must distinguish these two possibilities. Predicting a and b equally often achieves value $1/2$, while the demonstrator's value is $1/2 + 2\varepsilon$, so attaining excess at most ε requires identifying the sign of an $\mathcal{O}(\varepsilon)$ bias. This is a standard small-bias testing problem and requires order $1/\varepsilon^2$ samples.

A pass@2 learner can instead return both responses and obtain reward one under either instance, without estimating the bias at all. What remains is a coverage problem: the learner must ensure that relevant response possibilities are represented among its outputs. This is closely related to missing-mass phenomena (Rashidinejad et al., 2021; Rajaraman et al., 2020), where the error is controlled by the probability mass of possibilities not yet covered by the sample, leading to a $1/m$ rather than a $1/\sqrt{m}$ accuracy scale. Thus multiple responses remove the small-bias estimation bottleneck and replace it by a coverage-type difficulty.

2.3 Robust evaluation

Standard evaluation revealed two distinct benefits of multiple responses: moving from $k = 1$ to $k \geq 2$ removes the suboptimality-induced slow rate, while increasing k further improves the dependence on the reward-class size by a factor of $\log k$. Robust evaluation allows us to ask whether these two benefits persist when a single learned policy must perform well simultaneously for every reward in $\mathcal{R}$.

The $1/\log k$ improvement disappears. The same learning rule used under standard evaluation also gives a robust guarantee.

Theorem 6 (Robust pass@k upper bound)**.** *Let $k \geq 2$. The learning rule $\mathcal{A}_k$ from Theorem 3 satisfies the following guarantee: for every $m \geq 1$, $\delta \in (0,1)$, context distribution $\mathcal{D}$, and demonstrator π_{demo}, the policy $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S)$ obeys*

$$\mathbb{P}_S\left\{\sup_{r\in\mathcal{R}} \mathcal{E}_r^k(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{3\log|\mathcal{R}|}{m\log k} + \frac{3\log(|\mathcal{R}|/\delta)+1}{m}\right\} \leq \delta.$$

Consequently, for every $\varepsilon > 0$, $\delta \in (0,1)$, and $\Delta \in [0,1]$,

$$m_{\mathrm{rob}}^{(k)}(\mathcal{R}, \varepsilon, \delta, \Delta) \leq \left\lceil \frac{8\log(|\mathcal{R}|/\delta)+1}{\varepsilon} \right\rceil. \tag{1}$$

The same learner therefore retains the fast $1/\varepsilon$ dependence under robust evaluation. The difference is that simultaneous validity introduces an additional $\log|\mathcal{R}|$ term that is not reduced as k grows. From the upper bound alone, however, it is unclear whether this loss is fundamental or merely the price of the union-bound analysis. The next result shows that it is fundamental in the high-accuracy regime.

Theorem 7 (Robust pass@k lower bound)**.** *For every pair of integers $k \geq 2$ and $N \geq 2k^2$, there exist finite spaces $\mathcal{X}, \mathcal{Y}$ and a binary reward class $\mathcal{R} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{Y}}$ with $|\mathcal{R}| = N$ such that, for every $0 < \Delta \leq 1$, $0 < \delta < 1/16$, and $0 < \varepsilon \leq \frac{1}{384}\min\{\Delta, 1/k\}$, we have*

$$m_{\mathrm{rob}}^{(k)}(\mathcal{R}, \varepsilon, \delta, \Delta) > \left\lfloor \frac{\log N + \log(1/\delta)}{384\varepsilon} \right\rfloor. \tag{2}$$

Combining the two bounds, for $N \geq 2k^2$ and $\varepsilon \lesssim \min\{\Delta, \frac{1}{k}\}$, we obtain the sharp sample complexity $(\log N + \log(1/\delta))/\varepsilon$. Thus robust evaluation cleanly separates the two benefits identified under standard evaluation. The transition to a fast $1/\varepsilon$ rate survives, but the additional $1/\log k$ improvement does not: in the high-accuracy regime, a full $\log N$ dependence is unavoidable regardless of k. Since the same learner achieves the standard and robust upper bounds, this difference is caused by the stronger evaluation requirement rather than by a change in algorithm. The proofs of Theorem 6 and Theorem 7 are given in Appendices A.4 and A.5, respectively.

Comparison with pass@1. The disappearance of the $1/\log k$ gain raises a natural question: does allowing multiple responses still help at all under robust evaluation? The answer is yes. To see this, we compare the preceding rate with pass@1. Although the pass@1 result of Joshi et al. (2026) is stated under standard evaluation, their analysis actually gives a simultaneous high-probability guarantee over the reward class. It therefore implies the robust upper bound

$$m_{\mathrm{rob}}^{(1)}(\mathcal{R}, \varepsilon, \delta, \Delta) \lesssim \frac{\log(|\mathcal{R}|/\delta)}{\varepsilon} + \frac{\Delta\log(|\mathcal{R}|/\delta)}{\varepsilon^2}.$$

We show that the slow term is unavoidable.

Theorem 8 (Robust pass@1 lower bound)**.** *For every integer $N \geq 2$ and $0 < \delta < 1/16$, there exist finite spaces $\mathcal{X}, \mathcal{Y}$ and a finite binary reward class $\mathcal{R}$ with $|\mathcal{R}| \leq N$ such that, for every $0 < \Delta \leq 1$ and $0 < \varepsilon \leq \Delta/256$,*

$$m^{(1)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, \Delta) > \left\lfloor \frac{\Delta\,[\log N + \log(1/\delta)]}{262144\,\varepsilon^2} \right\rfloor. \tag{3}$$

When $\varepsilon \leq c_0\Delta$ for some sufficiently small constant c_0, the upper bound has order $\Delta[\log|\mathcal{R}| + \log(1/\delta)]/\varepsilon^2$. The lower bound shows that this dependence is optimal in general. Taking $\Delta = 1$ gives the arbitrary-demonstrator rate $[\log N + \log(1/\delta)]/\varepsilon^2$. Comparing this with the pass@k rate shows that the slow-to-fast transition survives robust evaluation, even though the additional $1/\log k$ improvement does not. See the proof in Appendix A.6.[2] All sample complexity results are summarized in the tables in Appendix C.

3 Learning algorithm and analysis

This section develops the learner underlying the pass@k upper bounds for $k \geq 2$ in Section 2. The key idea is to combine greedy coverage (Nemhauser et al., 1978) with multiplicative weights (Arora et al., 2012): thresholding turns bounded rewards into coverage objectives, and selecting k responses allows an asymmetric weight update that yields the $\log|\mathcal{R}|/\log k$ dependence. We first describe and analyze the online learner, then connect its online-to-batch conversion to the standard and robust statistical guarantees.

3.1 The online learner

At round t, the learner observes x_t, produces k responses, then receives the demonstration y_t. It knows $\mathcal{R}$ but receives no feedback identifying the evaluation reward. The protocol in Figure 1 allows arbitrary sequences, without independence or demonstrator-optimality assumptions.

Algorithm 1

Input: Finite and non-empty reward class $\mathcal{R} \subseteq [0,1]^{\mathcal{X}\times\mathcal{Y}}$, parameters $k \in \mathbb{N}\backslash\{1\}$, $L, T \in \mathbb{N}$.

Initialize $w^{(1)}(r,j) = 1$ for all $(r,j) \in \mathcal{S}_L := \mathcal{R} \times [L]$.
for round $t = 1, 2, \ldots, T$ **do**
  Receive context x_t.
  For $y \in \mathcal{Y}$, we set, $A^y_t = \{(r,j) \in \mathcal{S}_L \mid r(x_t, y) \geq u_j\}$, where $u_j := \frac{j}{L}$.
  Greedy top-k selection: Set $Y_0 = \emptyset$. For $i = 1, \ldots, k$:

$$\hat{y}^{(i)}_t \in \arg\max_{y \in \mathcal{Y}\backslash Y_{i-1}} \sum_{(r,j) \in A^y_t \backslash \cup_{z \in Y_{i-1}} A^z_t} w^{(t)}(r,j), \quad Y_i \leftarrow Y_{i-1} \cup \{\hat{y}^{(i)}_t\}.$$

  Output $(\hat{y}^{(1)}_t, \ldots, \hat{y}^{(k)}_t)$.
  Define $U_t := \bigcup_{i=1}^k A^{\hat{y}^{(i)}_t}_t$.
  Receive demonstration y_t.
  Define $D_t := A^{y_t}_t = \{(r,j) : r(x_t, y_t) \geq u_j\}$.
  Weight update: For each $(r,j) \in \mathcal{S}_L$:

$$w^{(t+1)}(r,j) \leftarrow \begin{cases} (1 + k/4)\, w^{(t)}(r,j), & (r,j) \in D_t \setminus U_t, \\ (3/4)\, w^{(t)}(r,j), & (r,j) \in U_t \setminus D_t, \\ w^{(t)}(r,j), & \text{otherwise.} \end{cases}$$

end for

For binary rewards, a response *covers* the rewards for which it scores one, and a tuple succeeds whenever at least one response covers the evaluation reward. For general rewards in $[0,1]$, we apply

[2] At $\Delta = 0$, a stronger conclusion holds: no demonstrations are needed. At each context, choose a response maximizing the sum of the known rewards. Under class-wide optimality, this policy is optimal for every reward on a set of contexts of probability one.

the same idea to reward-threshold pairs. Given an integer $L \geq 1$, define $u_j := j/L,\ j \in [L]$ and $\mathcal{S}_L := \mathcal{R} \times [L]$. A response y at context x_t covers the set

$$A_t^y := \{(r, j) \in \mathcal{S}_L : r(x_t, y) \geq u_j\}.$$

For each reward, the fraction of thresholds covered is $\lfloor Lr(x_t, y) \rfloor / L$, which approximates its value with error at most $1/L$. Thus, threshold coverage represents bounded rewards up to a controlled discretization error. Algorithm 1 maintains positive weights on $\mathcal{S}_L$. Write

$$w^{(t)}(A) := \sum_{(r,j)\in A} w^{(t)}(r, j), \qquad A \subseteq \mathcal{S}_L.$$

Each greedy step selects the response covering the largest remaining weight. Let U_t denote the union of reward-threshold pairs covered by the learner's tuple and D_t the set covered by the demonstration. All weights are initialized to one. After observing y_t, the algorithm updates each weight according to three cases:

- **Only the demonstration covers the pair.** If $(r, j) \in D_t \setminus U_t$, the demonstration clears u_j but none of the learner's responses does. Multiplying the weight by $1 + k/4$ gives this pair greater priority in future selections.
- **Only the learner covers the pair.** If $(r, j) \in U_t \setminus D_t$, a learner response clears u_j but the demonstration does not. Multiplying the weight by $3/4$ reduces this pair's priority.
- **Both cover the pair, or neither does.** The weight remains unchanged.

These updates prioritize pairs where the learner falls behind the demonstrator, without requiring optimal demonstrations. Ties follow a fixed rule. If $|\mathcal{Y}| < k$, return every response and pad with repetitions, attaining the optimal reward for every $r \in \mathcal{R}$.

3.2 Analysis of the online learner

The analysis combines the coverage advantage of greedy selection with control of the total weight. Lemma 18 in the appendix shows that, at every round,

$$w^{(t)}(U_t \setminus D_t) \geq k\, w^{(t)}(D_t \setminus U_t).$$

Thus, reducing the weights on $U_t \setminus D_t$ by a factor of $3/4$ removes at least as much total weight as multiplying the weights on $D_t \setminus U_t$ by $1 + k/4$ adds. Consequently, the potential $W_t := w^{(t)}(\mathcal{S}_L)$ is non-increasing, and $W_{T+1} \leq W_1 = L|\mathcal{R}|$ (Lemma 19 and Corollary 20).

For a fixed reward, averaging the log-weights over its L thresholds cancels the factor L in this potential bound, leaving a $\log|\mathcal{R}|$ term. A threshold covered only by the demonstration contributes $\log(1 + k/4)$ to its log-weight, while one covered only by the learner subtracts $\log(4/3)$. Since $\log(1 + k/4) \geq \log(4/3)$, the accumulated updates bound the cumulative discretized regret by $\log|\mathcal{R}|/\log(1 + k/4)$, see Lemma 10. Using $\log(1 + k/4) \geq \frac{1}{2}\log k$ and accounting for at most $1/L$ rounding error per round yields the following result.

Theorem 9 (Online regret)**.** *For integers $k \geq 2$ and $L, T \geq 1$, Algorithm 1 satisfies, for every $r \in \mathcal{R}$ and every sequence of contexts and demonstrations,*

$$\sum_{t=1}^{T} \left[r(x_t, y_t) - \max_{i\in[k]} r(x_t, \hat{y}_t^{(i)}) \right] \leq \frac{2\log|\mathcal{R}|}{\log k} + \frac{T}{L}.$$

The guarantee holds for every realized sequence, without independence or demonstrator-optimality assumptions. The $1/\log k$ improvement comes from the larger weight increase per missed threshold permitted by greedy coverage. For binary rewards, $L = 1$ gives exact thresholding, so the T/L term can be omitted. The proof is in Appendix B.1.

Comparison with pass@1. For $|\mathcal{R}| \geq 2$, tuning the learning rate in Theorem 5 of Joshi et al. (2026) gives, under i.i.d. demonstrations and for every fixed $r \in \mathcal{R}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \big(r(x_t, y_t) - r(x_t, \hat{y}_t)\big)\right] \leq \max\Big\{4\log|\mathcal{R}|,\, 2\sqrt{2T\Delta_{\mathcal{R}}\log|\mathcal{R}|}\Big\}.$$

Here $\hat{y}_t$ is the single response of the pass@1 learner, and $\Delta_{\mathcal{R}}$ is the demonstrator's worst-case suboptimality over $\mathcal{R}$. Theorem 9 bounds the pass@k regret by $2\log|\mathcal{R}|/\log k + T/L$ for every realized sequence and every $k \geq 2$. Thus, up to the discretization error T/L, this guarantee eliminates the suboptimality-dependent term and reduces the reward-class contribution from order $\log|\mathcal{R}|$ to $\log|\mathcal{R}|/\log k$. It also holds without independence assumptions or quality-dependent tuning.

3.3 Online-to-batch conversion

We now apply the online-to-batch conversion (Cesa-Bianchi et al., 2004) to our online learner to obtain the statistical upper bounds in Section 2. Let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ consist of $m \geq 1$ i.i.d. demonstrations, with $x_t \sim \mathcal{D}$ and $y_t \sim \pi_{\mathrm{demo}}(\cdot \mid x_t)$. Run Algorithm 1 on this sample with $T = m$. Let $\widehat{\pi}_t : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ be the policy used at round t, constructed from the first $t-1$ demonstrations. The learned policy is the uniform mixture

$$\widehat{\pi}_{\mathrm{o2b}}(\cdot \mid x) := \frac{1}{m}\sum_{t=1}^{m} \widehat{\pi}_t(\cdot \mid x), \qquad x \in \mathcal{X}. \tag{4}$$

At prediction time, draw one index $\tau \sim \mathrm{Unif}([m])$, independently of the training sample and test context, and generate the entire k-response tuple using $\widehat{\pi}_\tau$. The complete procedure is given in Algorithm 2 in the appendix. Because the mixture is over distributions on complete tuples, its population excess value is exactly the average of the snapshot excess values:

$$\mathcal{E}_r^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}}) = \frac{1}{m}\sum_{t=1}^{m}\big[V_r(\pi_{\mathrm{demo}}) - V_r^k(\widehat{\pi}_t)\big].$$

Moreover, $\widehat{\pi}_t$ is determined before the independent pair (x_t, y_t) is observed. Conditional on the preceding demonstrations, the expected reward shortfall at round t therefore equals $V_r(\pi_{\mathrm{demo}}) - V_r^k(\widehat{\pi}_t)$. This is the link between the online process and population performance.

Recovering the upper bounds in Section 2. Set $L = m$, giving discretization error at most $1/m$. For a reward $r^\star$ fixed before training, a one-step exponential bound relates threshold log-weight changes to conditional expected shortfalls. Combining these bounds across rounds gives an exponential supermartingale. The total-weight bound and Markov's inequality then yield Theorem 3, with reward-class term $3\log|\mathcal{R}|/(m\log k)$ and confidence term $3\log(1/\delta)/m$. See Appendix A.1 for the concentration argument. For robust evaluation, apply the same fixed-reward guarantee with failure probability $\delta/|\mathcal{R}|$ to each $r \in \mathcal{R}$ and take a union bound. The confidence term becomes $3\log(|\mathcal{R}|/\delta)/m$, yielding Theorem 6 for the same learned policy; see Appendix A.4. This additional term explains why the robust upper bound retains a $\log|\mathcal{R}|/m$ contribution as k increases.

4 Discussion

For any $k \geq 2$, our learner achieves $1/\varepsilon$ sample complexity even with suboptimal demonstrations. Under standard evaluation, larger k also reduces the reward-class dependence. Under robust evaluation, this additional gain can disappear in the high-accuracy regime. The same learner gives both guarantees. Our learner assumes a known finite reward class and exact maximization of weighted coverage, which may be expensive for large reward classes or response spaces. Whether approximate selection and weight updates retain the guarantees remains open. Our evaluation compares the best reward among k responses with the demonstrator's single-response value. Analyzing an imperfect verifier would quantify the loss from selecting one candidate.

For $\Delta > 0$, our matching robust lower bound requires ε to be small relative to Δ and $1/k$. The sharp dependence on k for larger ε remains open. We allow dependent responses; the rates under independent sampling from a single policy remain open.

Infinite reward classes require a complexity measure in place of cardinality. Pour et al. (2026) use variants of the Littlestone dimension to characterize realizable online learning with multiple correct answers and obtain sample bounds for infinite classes. Their model with one correct response but no feedback on the learner's correctness is close to our binary-reward setting with optimal demonstrations. They study single-response prediction with a different agnostic benchmark. Extending these measures to pass@k with suboptimal demonstrations remains open.

AI USE STATEMENT

We used generative AI tools to polish the writing, improve the clarity of mathematical explanations, and help identify and format relevant references. The authors take responsibility for the final content, including the mathematical claims and citations.

ETHICS STATEMENT

This work is theoretical and does not involve human participants, the collection or use of personal data, or the deployment of learning systems.

REPRODUCIBILITY STATEMENT

The learning model, assumptions, and evaluation criteria are specified in Section 2. Section 3 describes the learner and its analysis, with pseudocode in Algorithms 1 and 2. Proofs of the statistical upper and lower bounds, the online regret bound, and supporting lemmas are provided in the appendix.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. The best of n worlds: Aligning reinforcement learning with best-of-n sampling via max@ k optimisation. *arXiv preprint arXiv:2510.23393*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 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.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1713–1726, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Rianne De Heide, James Cheshire, Pierre Ménard, and Alexandra Carpentier. Bandits with many optimal arms. *Advances in Neural Information Processing Systems*, 34:22457–22469, 2021.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-majority: Minimax-optimal strategy for pass@ k inference scaling. In *International Conference on Learning Representations*, volume 2026, pp. 92146–92171, 2026.

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, 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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 13–19 Jul 2025.

Nirmit Joshi, Gene Li, Siddharth Bhandari, Shiva Kasiviswanathan, Cong Ma, and Nathan Srebro. Learning to answer from correct demonstrations. In *International Conference on Learning Representations*, volume 2026, pp. 44352–44384, 2026.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *Cambridge University*, 355, 2019.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Shay Moran, Ohad Sharon, Iska Tsubari, and Sivan Yosebashvili. List online classification. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1885–1913. PMLR, 2023.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

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.

Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day*, 1964.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

Alireza F Pour, Farnam Mansouri, and Shai Ben-David. Learning with multiple correct answers–regret bounds under different feedback models. *arXiv preprint arXiv:2602.09402*, 2026.

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.

Umar Syed and Robert Schapire. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Remi Munos. Optimizing language models for inference time objectives using reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Christian Walder and Deep Tejas Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445, 2026.

CONTENTS

1 Introduction **1**

1.1 Related work 2

2 Problem Setup and Main Results **3**

2.1 Setup and evaluation criteria 3

2.2 Standard evaluation 4

2.3 Robust evaluation 6

3 Learning algorithm and analysis **7**

3.1 The online learner 7

3.2 Analysis of the online learner 8

3.3 Online-to-batch conversion 9

4 Discussion **9**

A Omitted details and proofs for Section 2 **13**

A.1 Proof of Theorem 3 13

A.2 Proof of Theorem 4 15

A.3 Proof of Theorem 5 18

A.4 Proof of Theorem 6 21

A.5 Proof of Theorem 7 22

A.6 Proof of Theorem 8 25

B Omitted details and proofs for Section 3 **28**

B.1 Proof of Theorem 9 30

C Tables **32**

A Omitted details and proofs for Section 2

A.1 Proof of Theorem 3

Fix $k \geq 2$. The learning rule $\mathcal{A}_k$ runs Algorithm 1 with $T = L = m$ and returns the mixture $\widehat{\pi}_{\mathrm{o2b}}$ in equation 4, as specified in Algorithm 2. With fixed tie-breaking, the learned policy is determined by S. Its prediction-time randomness is already averaged in V_r^k. The rule does not depend on the evaluation reward, the confidence level, or the demonstrator's suboptimality.

If $|\mathcal{Y}| < k$, returning every response and padding with repetitions gives $V_r^k(\widehat{\pi}_{\mathrm{o2b}}) = V_r^\star \geq V_r(\pi_{\mathrm{demo}})$ for every $r \in \mathcal{R}$, so the theorem is immediate. Henceforth, assume $|\mathcal{Y}| \geq k$. We establish the auxiliary bounds for arbitrary integers $L \geq 1$ and set $L = m$ at the end.

Threshold quantities and the weight bound. For $r \in \mathcal{R}$, write

$$d_t^r := r(x_t, y_t), \qquad a_t^r := \max_{i \in [k]} r(x_t, \hat{y}_t^{(i)}).$$

With $u_j = j/L$, define

$$q_L(v) := \frac{1}{L} \sum_{j=1}^{L} \mathbb{1}\{v \geq u_j\} = \frac{\lfloor Lv \rfloor}{L}, \qquad v \in [0,1], \tag{5}$$

$$M_t^r := \frac{1}{L} \sum_{j=1}^{L} \mathbb{1}\{d_t^r \geq u_j > a_t^r\}, \tag{6}$$

$$H_t^r := \frac{1}{L} \sum_{j=1}^{L} \mathbb{1}\{a_t^r \geq u_j > d_t^r\}. \tag{7}$$

Thus, M_t^r and H_t^r are the fractions of thresholds covered only by the demonstration and only by the learner, respectively. Rounding and cancellation of shared thresholds give

$$q_L(v) \leq v \leq q_L(v) + \frac{1}{L}, \qquad v \in [0,1], \tag{8}$$

$$M_t^r - H_t^r = q_L(d_t^r) - q_L(a_t^r). \tag{9}$$

Define the corresponding average log-weight change by

$$\alpha_k := \log(1 + k/4), \qquad \beta := \log(4/3), \qquad z_t^r := \alpha_k M_t^r - \beta H_t^r. \tag{10}$$

We have $\alpha_k \geq \log(3/2) > \beta$ and

$$1 + \frac{k}{4} - \sqrt{k} = \frac{(\sqrt{k} - 2)^2}{4} \geq 0 \quad \Longrightarrow \quad \alpha_k \geq \frac{1}{2} \log k. \tag{11}$$

Lemma 10 (Cumulative log-weight bound)**.** *For every $r \in \mathcal{R}$, integer $T \geq 1$, and sequence of contexts and demonstrations, Algorithm 1 satisfies*

$$\sum_{t=1}^{T} z_t^r \leq \log |\mathcal{R}|.$$

Proof. The initialization $w^{(1)}(r, j) = 1$ and the multiplicative updates imply

$$\frac{1}{L} \sum_{j=1}^{L} \log w^{(T+1)}(r, j) = \sum_{t=1}^{T} (\alpha_k M_t^r - \beta H_t^r) = \sum_{t=1}^{T} z_t^r.$$

Writing $W_t := \sum_{r \in \mathcal{R}} \sum_{j=1}^{L} w^{(t)}(r, j)$, Jensen's inequality and Corollary 20 give

$$\exp\left\{ \sum_{t=1}^{T} z_t^r \right\} \leq \frac{1}{L} \sum_{j=1}^{L} w^{(T+1)}(r, j) \leq \frac{W_{T+1}}{L} \leq |\mathcal{R}|.$$

Taking logarithms proves the claim. In particular, averaging over thresholds removes the factor L from the total-weight bound. □

From online shortfalls to population excess. Now let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ consist of i.i.d. demonstrations with $x_t \sim \mathcal{D}$ and $y_t \sim \pi_{\mathrm{demo}}(\cdot \mid x_t)$. Let $\mathcal{F}_0$ be the trivial sigma-algebra and set

$$\mathcal{F}_t := \sigma\big((x_1, y_1), \ldots, (x_t, y_t)\big), \qquad t \in [m].$$

The snapshot $\widehat{\pi}_t$ is constructed from the first $t-1$ demonstrations, so it is $\mathcal{F}_{t-1}$-measurable. For each $r \in \mathcal{R}$, define its conditional rounded shortfall

$$g_t^r := \mathbb{E}[M_t^r - H_t^r \mid \mathcal{F}_{t-1}].$$

Lemma 11 (Online-to-batch reduction)**.** *For every $r \in \mathcal{R}$, almost surely,*

$$g_t^r = V_{q_L \circ r}(\pi_{\mathrm{demo}}) - V_{q_L \circ r}^k(\widehat{\pi}_t), \tag{12}$$

$$\mathcal{E}_r^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}}) \le \frac{1}{m}\sum_{t=1}^m g_t^r + \frac{1}{L}. \tag{13}$$

Here the value notation is also applied to the bounded reward $q_L \circ r$.

Proof. Independence of (x_t, y_t) from $\mathcal{F}_{t-1}$ gives

$$\mathbb{E}[d_t^r - a_t^r \mid \mathcal{F}_{t-1}] = V_r(\pi_{\mathrm{demo}}) - V_r^k(\widehat{\pi}_t). \tag{14}$$

Since q_L is nondecreasing, it commutes with the maximum over the learner's responses. Applying the same argument to $q_L \circ r$ and using equation 9 proves equation 12.

The mixture in equation 4 is over complete k-response tuples, so its value is the average snapshot value. Together with equation 8, this yields

$$\begin{aligned}
\mathcal{E}_r^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}}) &\le V_{q_L \circ r}(\pi_{\mathrm{demo}}) - V_{q_L \circ r}^k(\widehat{\pi}_{\mathrm{o2b}}) + \frac{1}{L} \\
&= \frac{1}{m}\sum_{t=1}^m \big[V_{q_L \circ r}(\pi_{\mathrm{demo}}) - V_{q_L \circ r}^k(\widehat{\pi}_t)\big] + \frac{1}{L} \\
&= \frac{1}{m}\sum_{t=1}^m g_t^r + \frac{1}{L}.
\end{aligned}$$

Only one rounding error is needed: the demonstration value is rounded down by at most $1/L$, while rounding down the learner's value already gives an upper bound on the excess. □

Exponential control of the conditional shortfall. The asymmetric update provides an exponential bound that turns the cumulative weight control into a high-probability guarantee. Set

$$\theta_k := \frac{\log(3/2)}{\alpha_k}.$$

Since $k \ge 2$, we have $0 < \theta_k \le 1$; moreover, equation 11 implies

$$\theta_k \le \frac{2\log(3/2)}{\log k} \le \frac{1}{\log k}. \tag{15}$$

Lemma 12 (One-step exponential bound)**.** *For every fixed $r \in \mathcal{R}$ and every $t \in [m]$,*

$$\mathbb{E}\Big[e^{-\theta_k z_t^r} \mid \mathcal{F}_{t-1}\Big] \le 1 - \frac{g_t^r}{3} \le e^{-g_t^r/3}.$$

Proof. Fix r, t and, for $j \in [L]$, write

$$I_j := \mathbb{1}\{d_t^r \ge u_j > a_t^r\}, \qquad J_j := \mathbb{1}\{a_t^r \ge u_j > d_t^r\}.$$

These indicators cannot both equal one. Because $e^{-\theta_k \alpha_k} = 2/3$ and $e^{\theta_k \beta} \le e^{\beta} = 4/3$,

$$\begin{aligned}
e^{-\theta_k(\alpha_k I_j - \beta J_j)} &= 1 + (e^{-\theta_k \alpha_k} - 1)I_j + (e^{\theta_k \beta} - 1)J_j \\
&\le 1 - \frac{1}{3}(I_j - J_j).
\end{aligned}$$

Jensen's inequality over the thresholds therefore gives

$$e^{-\theta_k z_t^r} \le \frac{1}{L}\sum_{j=1}^{L} e^{-\theta_k(\alpha_k I_j - \beta J_j)} \le 1 - \frac{1}{3}(M_t^r - H_t^r).$$

Taking conditional expectations and using $1-u \le e^{-u}$ proves the result. No nonnegativity assumption on g_t^r is needed. □

Lemma 13 (Exponential supermartingale)**.** *For each fixed $r \in \mathcal{R}$, the process*

$$K_{0,r} := 1, \qquad K_{t,r} := \exp\left\{\frac{1}{3}\sum_{s=1}^{t} g_s^r - \theta_k \sum_{s=1}^{t} z_s^r\right\}$$

is a nonnegative integrable supermartingale with respect to $(\mathcal{F}_t)_{t=0}^m$. In particular, $\mathbb{E}_S K_{m,r} \le 1$.

Proof. The process is adapted and integrable because $|g_t^r| \le 1$ and $-\beta \le z_t^r \le \alpha_k$. By the $\mathcal{F}_{t-1}$-measurability of g_t^r and Lemma 12,

$$\mathbb{E}[K_{t,r} \mid \mathcal{F}_{t-1}] = K_{t-1,r} e^{g_t^r/3} \mathbb{E}\left[e^{-\theta_k z_t^r} \mid \mathcal{F}_{t-1}\right] \le K_{t-1,r}.$$

Iterating expectations from $K_{0,r} = 1$ proves the claim. □

Proof of Theorem 3. Fix $\mathcal{D}$, π_{demo}, and $r^\star \in \mathcal{R}$ before sampling, and set $L = m$. Lemmas 10 and 13 imply

$$\begin{aligned}\mathbb{E}_S \exp\left\{\frac{1}{3}\sum_{t=1}^{m} g_t^{r^\star}\right\} &= \mathbb{E}_S\left[K_{m,r^\star} \exp\left\{\theta_k \sum_{t=1}^{m} z_t^{r^\star}\right\}\right] \\ &\le \exp\{\theta_k \log|\mathcal{R}|\}.\end{aligned}$$

Applying Lemma 11 gives

$$\mathbb{E}_S \exp\left\{\frac{m}{3}\mathcal{E}_{r^\star}^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}})\right\} \le \exp\left\{\theta_k \log|\mathcal{R}| + \frac{1}{3}\right\}.$$

Hence, Markov's inequality yields

$$\mathbb{P}_S\left\{\mathcal{E}_{r^\star}^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}}) > \frac{3\theta_k \log|\mathcal{R}| + 3\log(1/\delta) + 1}{m}\right\} \le \delta.$$

Using equation 15, we obtain

$$\mathbb{P}_S\left\{\mathcal{E}_{r^\star}^k(\widehat{\pi}_{\mathrm{o2b}}, \pi_{\mathrm{demo}}) > \frac{3\log|\mathcal{R}|}{m\log k} + \frac{3\log(1/\delta)+1}{m}\right\} \le \delta. \tag{16}$$

For $\varepsilon \in (0,1)$, the threshold in equation 16 is at most ε whenever

$$m \ge \left\lceil \frac{3\log|\mathcal{R}|/\log k + 3\log(1/\delta) + 1}{\varepsilon} \right\rceil.$$

This guarantee holds for every demonstrator and therefore for every family satisfying $\Delta_{r^\star} \le \Delta$, where $\Delta \in [0,1]$. Definition 1 now gives

$$m_{\mathrm{std}}^{(k)}(\mathcal{R}, \varepsilon, \delta, \Delta) \le \left\lceil \frac{3\log|\mathcal{R}|/\log k + 3\log(1/\delta) + 1}{\varepsilon} \right\rceil. \tag{17}$$

For $\varepsilon \ge 1$, the accuracy requirement is automatic because $\mathcal{E}_{r^\star}^k(\pi, \pi_{\mathrm{demo}}) \le 1$ for every policy. □

A.2 Proof of Theorem 4

Proof. We use one fixed reward class for two constructions: one gives the dependence on $|\mathcal{R}|$, and the other gives the confidence term. Both use demonstrators that are optimal for the selected reward. The construction is valid for $k \ge 1$, so its $k = 1$ case also applies in Appendix A.3.

The reward class and inclusion probabilities. Fix $k \geq 1$ and a prescribed class size $|\mathcal{R}| \geq 2k$. Set

$$q := 2k, \qquad d := \left\lfloor \frac{\log |\mathcal{R}|}{\log q} \right\rfloor \geq 1.$$

Start with $\mathcal{X}_0 := \{0, 1, \ldots, d\}$ and $\mathcal{Y} := \{0, 1, \ldots, q\}$. For each $\theta \in [q]^d$, define

$$f_\theta(0) := 0, \qquad f_\theta(j) := \theta_j \quad (j \in [d]), \qquad r_\theta(x, y) := \mathbb{1}\{y = f_\theta(x)\}.$$

This gives q^d distinct binary rewards.

To reach the prescribed class size, enlarge $\mathcal{X}_0$ to $\mathcal{X}$ by adding $|\mathcal{R}| - q^d$ auxiliary contexts, and extend every f_θ by zero there. For each auxiliary context, add a graph-indicator reward with target response one at that context and at every $j \in [d]$, and zero at all other contexts. These additional rewards are distinct from each other and from the original rewards. Assign every auxiliary context probability zero in all distributions below. The resulting spaces and class $\mathcal{R}$ depend only on k and the prescribed class size.

Fix $m \geq 1$, $0 < \delta < 1/16$, and an arbitrary, possibly randomized learning rule returning $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S)$. For a realized output policy, define, for $j \in \mathcal{X}$ and $a \in \mathcal{Y}$,

$$C_j(a) := \mathbb{P}_{\mathbf{y} \sim \widehat{\pi}(\cdot|j)}\{a \in \{y^{(1)}, \ldots, y^{(k)}\}\}.$$

Prediction-time randomness is averaged in $C_j(a)$; probabilities over the learned policy below include any training randomness. Since a tuple contains at most k distinct responses,

$$\sum_{a=1}^{q} C_j(a) = \mathbb{E}_{\mathbf{y} \sim \widehat{\pi}(\cdot|j)}\big|[q] \cap \{y^{(1)}, \ldots, y^{(k)}\}\big| \leq k. \tag{18}$$

This holds for arbitrarily dependent responses.

Let $\pi_{\mathrm{demo}_\theta}$ return $f_\theta(x)$ deterministically. For every distribution supported on $\mathcal{X}_0$,

$$\begin{aligned} V_{r_\theta}(\pi_{\mathrm{demo}_\theta}) &= V^\star_{r_\theta} = 1, \\ \mathcal{E}^k_{r_\theta}(\widehat{\pi}, \pi_{\mathrm{demo}_\theta}) &= \sum_{j=0}^{d} \mathcal{D}(j)\big[1 - C_j(f_\theta(j))\big]. \end{aligned} \tag{19}$$

Thus every such demonstrator has $\Delta_{r_\theta} = 0$.

The reward-class contribution. Let $b := \min\{m, d\}$ and choose

$$\mathcal{D}(j) := \frac{1}{2m} \quad (j \in [b]), \qquad \mathcal{D}(0) := 1 - \frac{b}{2m},$$

with zero mass elsewhere. Temporarily draw Θ uniformly from $[q]^d$, and then draw the training sample using $\mathcal{D}$ and $\pi_{\mathrm{demo}_\Theta}$.

For $j \in [b]$, let U_j be the event that context j is absent from the sample. Then

$$\mathbb{P}(U_j) = \left(1 - \frac{1}{2m}\right)^m \geq \frac{1}{2}.$$

On U_j, the unobserved label Θ_j remains uniform on $[q]$, even conditional on S and $\widehat{\pi}$: the learner's output depends on the hidden parameter only through the observed sample. Consequently, equation 18 gives, on U_j,

$$\mathbb{E}[1 - C_j(\Theta_j) \mid S, \widehat{\pi}] = 1 - \frac{1}{q}\sum_{a=1}^{q} C_j(a) \geq 1 - \frac{k}{q} = \frac{1}{2}.$$

Define the excess contributed by these contexts as

$$W := \frac{1}{2m}\sum_{j=1}^{b}[1 - C_j(\Theta_j)].$$

By equation 19, the total excess is at least W, and

$$0 \le W \le \frac{b}{2m}, \qquad \mathbb{E}W \ge \frac{b}{8m}.$$

Splitting the expectation according to whether $W > b/(16m)$ yields

$$\frac{b}{8m} \le \mathbb{E}W \le \frac{b}{16m} + \frac{7b}{16m}\,\mathbb{P}\left\{W > \frac{b}{16m}\right\}.$$

Hence this event has probability at least $1/7$. Since

$$d \ge \frac{\log|\mathcal{R}|}{2\log q}, \qquad \frac{b}{16m} \ge \frac{1}{32}\min\left\{1, \frac{\log|\mathcal{R}|}{m\log q}\right\},$$

averaging over Θ selects a fixed parameter θ such that, with $r^\star = r_\theta$ and $\pi_{\mathrm{demo}} = \pi_{\mathrm{demo}_\theta}$,

$$\mathbb{P}\left\{\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{1}{32}\min\left\{1, \frac{\log|\mathcal{R}|}{m\log(2k)}\right\}\right\} \ge \frac{1}{7}. \tag{20}$$

The confidence contribution. Use the same reward class, but now let

$$\eta := \min\left\{\frac{1}{2}, \frac{\log(1/(8\delta))}{2m}\right\}, \qquad \mathcal{D}(1) := \eta, \qquad \mathcal{D}(0) := 1-\eta,$$

with zero mass elsewhere. Again draw Θ uniformly from $[q]^d$ and use $\pi_{\mathrm{demo}_\Theta}$. Let E be the event that context 1 is absent from the sample. Since $0 < \eta \le 1/2$ and $\log(1-\eta) \ge -2\eta$,

$$\mathbb{P}(E) = (1-\eta)^m \ge e^{-2m\eta} \ge 8\delta.$$

On E, Θ_1 remains uniform conditional on $S, \widehat{\pi}$. Thus $H := 1 - C_1(\Theta_1)$ satisfies $0 \le H \le 1$ and

$$\frac{1}{2} \le \mathbb{E}[H \mid S, \widehat{\pi}] \le \frac{1}{4} + \frac{3}{4}\mathbb{P}\{H > 1/4 \mid S, \widehat{\pi}\}.$$

Therefore the conditional probability on the right is at least $1/3$ on E. The excess is at least ηH, so

$$\mathbb{P}\left\{\mathcal{E}^k_{r_\Theta}(\widehat{\pi}, \pi_{\mathrm{demo}_\Theta}) > \frac{\eta}{4}\right\} \ge \frac{8\delta}{3}.$$

For $\delta < 1/16$, $\log(1/(8\delta)) \ge \frac{1}{4}\log(1/\delta)$, which gives

$$\frac{\eta}{4} \ge \frac{1}{32}\min\left\{1, \frac{\log(1/\delta)}{m}\right\}.$$

Averaging over Θ therefore selects a fixed optimal-demonstrator instance satisfying

$$\mathbb{P}\left\{\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{1}{32}\min\left\{1, \frac{\log(1/\delta)}{m}\right\}\right\} \ge \frac{8\delta}{3}. \tag{21}$$

Combining the bounds and obtaining sample complexity. Choose the construction with the larger threshold in equation 20 and equation 21. Both use the same class $\mathcal{R}$, and both failure probabilities exceed 2δ. Since, for $u, v \ge 0$,

$$\max\{\min\{1,u\}, \min\{1,v\}\} \ge \frac{1}{2}\min\{1, u+v\},$$

the selected instance satisfies

$$\mathbb{P}\left\{\mathcal{E}^k_{r^\star}(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{1}{64}\min\left\{1, \frac{1}{m}\left(\frac{\log|\mathcal{R}|}{\log(2k)} + \log\frac{1}{\delta}\right)\right\}\right\} \ge 2\delta. \tag{22}$$

The distribution, demonstrator, and evaluation reward are selected using the learner's distributional performance, before the actual training sample or training randomness is drawn. Thus this is a lower bound for standard evaluation at a fixed reward.

Let $0 < \varepsilon \le 1/128$. Whenever

$$1 \le m < \frac{1}{64\varepsilon}\left(\frac{\log|\mathcal{R}|}{\log(2k)} + \log\frac{1}{\delta}\right),$$

the excess threshold in equation 22 is strictly larger than ε. Hence no learner can satisfy the PAC requirement at such a sample size. The upper limit here exceeds one, so a sample-complexity threshold of zero is also impossible: it would require success at $m = 1$.

All constructed demonstrators satisfy $\Delta_{r^\star} = 0$, so they are admissible for every $\Delta \in [0, 1]$. Definition 1 therefore gives, for every $k \ge 1$,

$$m^{(k)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \ge \frac{1}{64\varepsilon}\left(\frac{\log|\mathcal{R}|}{\log(2k)} + \log\frac{1}{\delta}\right).$$

Finally, for $k \ge 2$, we have $\log(2k) \le 2\log k$, and hence

$$m^{(k)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \ge \frac{1}{128\varepsilon}\left(\frac{\log|\mathcal{R}|}{\log k} + \log\frac{1}{\delta}\right).$$

This proves the theorem with $c = c_0 = 1/128$. □

A.3 Proof of Theorem 5

We first record the information-theoretic tools used here and in the robust lower-bound proof. For probability laws P, Q on a finite set, write

$$D_{\mathrm{KL}}(P\|Q) := \sum_z P(z)\log\frac{P(z)}{Q(z)}, \qquad \mathrm{TV}(P, Q) := \frac{1}{2}\sum_z |P(z) - Q(z)|,$$

with the usual conventions at zero, and let $\mathrm{kl}(p, q)$ denote the divergence between $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$. Testing probabilities below include any randomization used by the test.

Lemma 14 (KL identities and data processing)**.** *The following identities and inequality follow from Lattimore & Szepesvári (2019, Exercises 14.10 and 14.12). For $m \ge 1$,*

$$D_{\mathrm{KL}}(P^{\otimes m}\|Q^{\otimes m}) = mD_{\mathrm{KL}}(P\|Q).$$

If $P(x, y) = D(x)P_x(y)$ and $Q(x, y) = D(x)Q_x(y)$, then

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x:D(x)>0} D(x)D_{\mathrm{KL}}(P_x\|Q_x).$$

Adjoining independent randomization with the same law ν leaves the divergence unchanged: $D_{\mathrm{KL}}(P \otimes \nu\|Q \otimes \nu) = D_{\mathrm{KL}}(P\|Q)$. For every possibly randomized test $T \in \{0, 1\}$,

$$\mathrm{kl}\big(P\{T = 1\}, Q\{T = 1\}\big) \le D_{\mathrm{KL}}(P\|Q).$$

Lemma 15 (Pinsker's inequality and binary testing, Pinsker (1964); Lattimore & Szepesvári (2019), Eqs. (14.12)–(14.13))**.** *For probability laws P, Q on a finite set,*

$$\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)}.$$

Every possibly randomized test $T \in \{+1, -1\}$ satisfies

$$P\{T = -1\} + Q\{T = +1\} \ge 1 - \mathrm{TV}(P, Q).$$

Lemma 16 (Bretagnolle–Huber testing inequality,Lattimore & Szepesvári (2019)Theorem 14.2; Tsybakov & Zaiats (2009))**.** *For probability laws P, Q on a finite set and every possibly randomized test $T \in \{+1, -1\}$,*

$$\frac{1}{2}P\{T = -1\} + \frac{1}{2}Q\{T = +1\} \ge \frac{1}{4}\exp\{-D_{\mathrm{KL}}(P\|Q)\}.$$

Lemma 17 (Divergence between opposite Bernoulli biases,De Heide et al. (2021), Appendix A.2, inequality following (4))**.** *For* $0 < h \le 1/4$*,*

$$\begin{aligned}\mathrm{kl}(1/2+h, 1/2-h) &= \mathrm{kl}(1/2-h, 1/2+h)\\ &= 2h \log \frac{1+2h}{1-2h} \le 16h^2.\end{aligned}$$

Moreover, for $0 \le h < 1/2$*, since* $\log(1+u) \le u$ *for* $u > -1$*,*

$$\mathrm{kl}(1/2 \pm h, 1/2) = \left(\tfrac{1}{2}+h\right)\log(1+2h) + \left(\tfrac{1}{2}-h\right)\log(1-2h) \le \left(\tfrac{1}{2}+h\right)2h - \left(\tfrac{1}{2}-h\right)2h = 4h^2.$$

Proof. The equality follows by expanding Bernoulli relative entropy. Since $\log(1+u) \le u$ for $u \ge 0$,

$$\log \frac{1+2h}{1-2h} = \log\left(1 + \frac{4h}{1-2h}\right) \le \frac{4h}{1-2h} \le 8h.$$

Multiplying by $2h$ proves the bound. □

Proof of Theorem 5. We construct a class of the prescribed cardinality, denoted by $|\mathcal{R}| \ge 2$, that supports all three terms in the lower bound. Fix $\Delta \in [0,1]$ and $0 < \varepsilon, \delta \le 1/64$. All probabilities include training randomness; prediction-time randomness is averaged in the policy values.

A common reward class. Start with the class from Appendix A.2 at $k = 1$, including its auxiliary contexts. Its response space is $\mathcal{Y} = \{0,1,2\}$, all rewards select response zero at context 0, and the class contains two rewards selecting responses one and two, respectively, at context 1.

Adjoin a disjoint context set $\mathcal{Z} := \{-1,+1\}^{\mathcal{R}}$ and let μ be its uniform distribution. For each original reward r, extend it to $z \in \mathcal{Z}$ by

$$r(z,0) := 0, \qquad r(z,1) := \frac{1+z_r}{2}, \qquad r(z,2) := \frac{1-z_r}{2}.$$

Continue to denote the extended class by $\mathcal{R}$ and the enlarged context space by $\mathcal{X}$. The class retains its cardinality, every reward is binary, and every reward has optimal value one. The spaces and class depend only on the prescribed cardinality.

The fast-rate contribution. Assign zero mass to $\mathcal{Z}$. The constructions in Appendix A.2 remain valid, because the added reward values are fixed and reveal no information about the selected reward. Using equation 20 and equation 21 with $k = 1$, a learner satisfying the PAC requirement must have

$$m \ge \frac{1}{32\varepsilon} \max\left\{ \frac{\log|\mathcal{R}|}{\log 2}, \log\frac{1}{\delta} \right\}.$$

Indeed, below either threshold the corresponding excess bound is strictly larger than ε, since $\varepsilon \le 1/64 < 1/32$. The failure probability exceeds δ. These thresholds also rule out a sample-complexity threshold of zero by considering $m = 1$. Since the demonstrators are optimal for their selected rewards, for every budget Δ we obtain

$$m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \ge \frac{\log|\mathcal{R}| + \log(1/\delta)}{64\varepsilon}. \tag{23}$$

The suboptimality-dependent confidence term. Suppose $\Delta \ge 8\varepsilon$, and set

$$h := \frac{2\varepsilon}{\Delta} \le \frac{1}{4}, \qquad \mathcal{D}(1) := \Delta, \qquad \mathcal{D}(0) := 1 - \Delta,$$

with zero mass elsewhere. Choose rewards r_+, r_- whose target responses at context 1 are one and two, respectively. For $\sigma \in \{-1,+1\}$, let

$$\pi_{\mathrm{demo}_\sigma}(0 \mid 0) = 1, \qquad \pi_{\mathrm{demo}_\sigma}(1 \mid 1) = \frac{1}{2} + \sigma h, \qquad \pi_{\mathrm{demo}_\sigma}(2 \mid 1) = \frac{1}{2} - \sigma h.$$

The selected-reward suboptimality is $\Delta_{r_\sigma} = \Delta(1/2 - h) \le \Delta$. Behavior at zero-probability contexts may be fixed arbitrarily.

From the returned policy, define $\widehat{\sigma} = +1$ if $\widehat{\pi}(1 \mid 1) \geq 1/2$, and -1 otherwise. A test error implies that the learner assigns probability at most $1/2$ to the rewarded response at context 1. Its shortfall at context 0 is nonnegative, so

$$\{\widehat{\sigma} \neq \sigma\} \subseteq \left\{\mathcal{E}^1_{r_\sigma}(\widehat{\pi}, \pi_{\mathrm{demo}_\sigma}) \geq \Delta h = 2\varepsilon\right\}. \tag{24}$$

Let P_σ be the law of one demonstration. Lemmas 14 and 17 give

$$D_{\mathrm{KL}}(P_+^{\otimes m} \| P_-^{\otimes m}) \leq 16 m \Delta h^2 = \frac{64 m \varepsilon^2}{\Delta}.$$

If the learner satisfies the PAC requirement for both fixed rewards, its test-error probability is at most δ under each law. Lemma 16 implies

$$\delta \geq \frac{1}{4} \exp\left\{-\frac{64 m \varepsilon^2}{\Delta}\right\}.$$

Since $\log(1/(4\delta)) \geq \frac{1}{2}\log(1/\delta)$,

$$m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \geq \frac{\Delta \log(1/\delta)}{128 \varepsilon^2}, \qquad \Delta \geq 8\varepsilon. \tag{25}$$

The suboptimality-dependent reward-class term. Keep $h = 2\varepsilon/\Delta$ and $\Delta \geq 8\varepsilon$, but now put mass Δ on $\mathcal{Z}$ according to μ and mass $1 - \Delta$ on context 0. For each $r \in \mathcal{R}$, let

$$\pi_{\mathrm{demo}_r}(0 \mid 0) = 1, \qquad \pi_{\mathrm{demo}_r}(1 \mid z) = \frac{1}{2} + h z_r, \qquad \pi_{\mathrm{demo}_r}(2 \mid z) = \frac{1}{2} - h z_r.$$

Again, $\Delta_r = \Delta(1/2 - h) \leq \Delta$. Write P_r for the law of one demonstration, and let P_0 be the corresponding law with $h = 0$.

For a realized learned policy, define

$$\widehat{f}(z) := \widehat{\pi}(1 \mid z) - \widehat{\pi}(2 \mid z) \in [-1, 1].$$

The learner's reward at z is at most $(1 + \widehat{f}(z) z_r)/2$, whereas the demonstrator's expected reward is $1/2 + h$. Therefore

$$\mathcal{E}^1_r(\widehat{\pi}, \pi_{\mathrm{demo}_r}) \geq \Delta \left(h - \frac{1}{2}\mathbb{E}_{Z \sim \mu}[\widehat{f}(Z) Z_r]\right).$$

Since $\varepsilon = \Delta h/2$, success requires $\mathbb{E}_\mu[\widehat{f}(Z) Z_r] \geq h$. Thus define the candidate list

$$\mathcal{L}(\widehat{\pi}) := \left\{r \in \mathcal{R} : \mathbb{E}_\mu[\widehat{f}(Z) Z_r] \geq h\right\}.$$

The coordinate functions $z \mapsto z_r$ are orthonormal under μ. Bessel's inequality gives

$$h^2 |\mathcal{L}(\widehat{\pi})| \leq \sum_{r \in \mathcal{R}} \left(\mathbb{E}_\mu[\widehat{f}(Z) Z_r]\right)^2 \leq \mathbb{E}_\mu[\widehat{f}(Z)^2] \leq 1. \tag{26}$$

Hence a single policy can meet the necessary correlation condition for at most h^{-2} rewards.

To use this restriction, draw $r^\star$ uniformly from $\mathcal{R}$. Let $\mathbb{P}$ be the joint law of $(r^\star, S)$ with $S \mid r^\star = r \sim P_r^{\otimes m}$, and let $\mathbb{Q}$ instead draw $S \sim P_0^{\otimes m}$ independently of $r^\star$. Under both laws, apply the same learning rule to S. For the possibly randomized event $E := \{r^\star \in \mathcal{L}(\widehat{\pi})\}$, the PAC requirement and equation 26 imply

$$\mathbb{P}(E) \geq 1 - \delta, \qquad \mathbb{Q}(E) \leq \frac{1}{|\mathcal{R}| h^2}.$$

Meanwhile, Lemmas 14 and 17 yield

$$D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} D_{\mathrm{KL}}(P_r^{\otimes m} \| P_0^{\otimes m}) \leq 4 m \Delta h^2.$$

If $|\mathcal{R}| h^2 \geq 4$, data processing applied to E, together with $\mathrm{kl}(p, q) \geq p \log(1/q) - \log 2$, gives

$$4 m \Delta h^2 \geq (1 - \delta) \log(|\mathcal{R}| h^2) - \log 2 \geq \frac{1}{4} \log(|\mathcal{R}| h^2).$$

Consequently, using $h^2 = 4\varepsilon^2/\Delta^2$,

$$m \geq \frac{\Delta}{64\varepsilon^2} \log(|\mathcal{R}|h^2) \geq \frac{\Delta}{64\varepsilon^2} \log\left(1 + \frac{|\mathcal{R}|\varepsilon^2}{(\Delta+\varepsilon)^2}\right).$$

The last inequality follows from $1 + |\mathcal{R}|\varepsilon^2/(\Delta+\varepsilon)^2 \leq 1 + |\mathcal{R}|h^2/4 \leq |\mathcal{R}|h^2$. If $|\mathcal{R}|h^2 < 4$, the logarithm on the right is at most $\log 2 \leq \log(1/\delta)$, so the confidence bound equation 25 supplies the same conclusion with constant $1/128$. Thus, in both cases,

$$m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \geq \frac{\Delta}{128\varepsilon^2} \log\left(1 + \frac{|\mathcal{R}|\varepsilon^2}{(\Delta+\varepsilon)^2}\right), \qquad \Delta \geq 8\varepsilon. \tag{27}$$

Combining the three contributions. For $\Delta \geq 8\varepsilon$, the same class satisfies equation 23, equation 25, and equation 27. The maximum of these three lower bounds is at least one third of their sum, giving the claimed bound with constant $1/384$ in this regime.

For $0 \leq \Delta < 8\varepsilon$, the remaining terms are already controlled by the fast contribution:

$$\begin{aligned}
&\frac{\Delta}{\varepsilon^2}\left[\log\frac{1}{\delta} + \log\left(1 + \frac{|\mathcal{R}|\varepsilon^2}{(\Delta+\varepsilon)^2}\right)\right] \\
&\quad\leq \frac{8}{\varepsilon}\left[\log\frac{1}{\delta} + \log(1+|\mathcal{R}|)\right] \\
&\quad\leq \frac{16}{\varepsilon}\left[\log|\mathcal{R}| + \log\frac{1}{\delta}\right],
\end{aligned}$$

where $\log(1+|\mathcal{R}|) \leq 2\log|\mathcal{R}|$ because $|\mathcal{R}| \geq 2$. Combining this with equation 23 gives the claimed bound with constant $1/1088$ in this regime.

Therefore, for example, the universal choice $c = 1/2048$ works throughout:

$$\begin{aligned}
m^{(1)}_{\mathrm{std}}(\mathcal{R}, \varepsilon, \delta, \Delta) \geq \frac{1}{2048}\Bigg[&\frac{\log|\mathcal{R}| + \log(1/\delta)}{\varepsilon} \\
&+ \frac{\Delta}{\varepsilon^2}\left\{\log\frac{1}{\delta} + \log\left(1 + \frac{|\mathcal{R}|\varepsilon^2}{(\Delta+\varepsilon)^2}\right)\right\}\Bigg].
\end{aligned} \tag{28}$$

The class has exactly the prescribed cardinality. Each averaging or testing argument identifies a fixed evaluation reward before the actual sample is drawn, as required for standard evaluation. □

A.4 Proof of Theorem 6

Proof. Fix $k \geq 2$, $m \geq 1$, $\delta \in (0,1)$, a context distribution $\mathcal{D}$, and an arbitrary demonstrator π_{demo}. Let $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S) = \widehat{\pi}_{\mathrm{o2b}}$ be the learner constructed in Appendix A.1, with $L = m$. With fixed tie-breaking, this policy is determined by S, and the learning rule does not depend on the evaluation reward or the confidence level.

For each fixed $r \in \mathcal{R}$, apply Theorem 3 with failure probability $\delta/|\mathcal{R}|$. This gives

$$\mathbb{P}_S\left\{\mathcal{E}^k_r(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{3\log|\mathcal{R}|}{m\log k} + \frac{3\log(|\mathcal{R}|/\delta)+1}{m}\right\} \leq \frac{\delta}{|\mathcal{R}|}.$$

These bounds concern the same learned policy. Since $\mathcal{R}$ is finite, a union bound therefore yields

$$\mathbb{P}_S\left\{\sup_{r\in\mathcal{R}} \mathcal{E}^k_r(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{3\log|\mathcal{R}|}{m\log k} + \frac{3\log(|\mathcal{R}|/\delta)+1}{m}\right\} \leq \sum_{r\in\mathcal{R}} \frac{\delta}{|\mathcal{R}|} = \delta. \tag{29}$$

Thus requiring the guarantee simultaneously over $\mathcal{R}$ replaces the confidence term $\log(1/\delta)$ by $\log(|\mathcal{R}|/\delta)$.

For the sample-complexity conclusion, let $\varepsilon \in (0,1)$ and $\Delta \in [0,1]$. The threshold in equation 29 is at most ε whenever

$$m \geq \left\lceil \frac{3\log|\mathcal{R}|/\log k + 3\log(|\mathcal{R}|/\delta) + 1}{\varepsilon} \right\rceil.$$

The same learning rule, setting $L = m$ at each sample size, satisfies this guarantee for every $\mathcal{D}$ and every π_{demo}, hence also for all demonstrators with $\Delta_{\mathcal{R}} \le \Delta$. Definition 2 gives

$$m^{(k)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, \Delta) \le \left\lceil \frac{3 \log |\mathcal{R}| / \log k + 3 \log(|\mathcal{R}|/\delta) + 1}{\varepsilon} \right\rceil.$$

Finally, $k \ge 2$ and $\log |\mathcal{R}| \le \log(|\mathcal{R}|/\delta)$ imply

$$\frac{3 \log |\mathcal{R}|}{\log k} + 3 \log \frac{|\mathcal{R}|}{\delta} \le \left(\frac{3}{\log 2} + 3 \right) \log \frac{|\mathcal{R}|}{\delta} \le 8 \log \frac{|\mathcal{R}|}{\delta}.$$

Substituting this inequality proves equation 1. For $\varepsilon \ge 1$, the accuracy requirement is automatic because $\sup_{r \in \mathcal{R}} \mathcal{E}^k_r(\pi, \pi_{\mathrm{demo}}) \le 1$ for every policy. □

A.5 Proof of Theorem 7

Proof. We construct one reward class that supports both the reward-class and confidence contributions. Throughout, the demonstrators are deterministic, and probabilities include any training randomization used by the learner.

A fixed reward class. Fix $k \ge 2$ and a prescribed class size $|\mathcal{R}| \ge 2k^2$. Set

$$q := 2k, \qquad b := \left\lfloor \frac{\log |\mathcal{R}|}{\log(2k^2)} \right\rfloor \ge 1, \qquad \mathcal{Y} := \{0, 1, \dots, q\}.$$

On $\mathcal{X}_0 := \{0\} \cup ([b] \times [k])$, define, for $v = ((j_\ell, a_\ell))_{\ell=1}^b \in ([k] \times [q])^b$,

$$r_v(0, y) := 1, \qquad r_v((\ell, j), y) := \begin{cases} \mathbb{1}\{y = a_\ell\}, & j = j_\ell, \\ 1, & j \ne j_\ell. \end{cases}$$

Each reward selects one context and one preferred response in each block; it is identically one at all other core contexts. There are $(2k^2)^b$ distinct such rewards.

To obtain the prescribed cardinality, add $|\mathcal{R}| - (2k^2)^b$ auxiliary contexts and extend every core reward by $\mathbb{1}\{y = 0\}$ there. Fix one core reward. For each auxiliary context, add a reward that agrees with this fixed reward on $\mathcal{X}_0$, prefers response one at that auxiliary context, and prefers response zero at every other auxiliary context. Let $\mathcal{X}$ and $\mathcal{R}$ denote the resulting space and class. All distributions below assign zero mass to the auxiliary contexts, so the added rewards duplicate a core reward on the support and do not affect the robust excess or suboptimality. The spaces and class depend only on k and the prescribed cardinality, and $V^\star_r = 1$ for every $r \in \mathcal{R}$.

Fix $m \ge 1$, $0 < \Delta \le 1$, $0 < \delta < 1/16$, and an arbitrary learning rule returning $\widehat{\pi} = \mathcal{A}_k(\mathcal{R}, S)$. For a realized output policy, let

$$C_x(a) := \mathbb{P}_{\mathbf{y} \sim \widehat{\pi}(\cdot|x)}\{a \in \{y^{(1)}, \dots, y^{(k)}\}\}.$$

These inclusion probabilities average over prediction-time randomness. Since a tuple contains at most k distinct responses,

$$\sum_{a=1}^{q} C_x(a) = \mathbb{E}_{\mathbf{y} \sim \widehat{\pi}(\cdot|x)} \big| [q] \cap \{y^{(1)}, \dots, y^{(k)}\} \big| \le k. \tag{30}$$

This allows arbitrary dependence among the responses.

The reward-class contribution. Choose

$$p := \min\left\{ \frac{\Delta}{b}, \frac{1}{2bk}, \frac{\log k}{2m} \right\}, \qquad \mathcal{D}((\ell, j)) := p, \qquad \mathcal{D}(0) := 1 - bkp \ge \frac{1}{2}.$$

For $\theta \in [q]^{bk}$, let $\pi_{\mathrm{demo}_\theta}$ return $\theta_{\ell,j}$ at (ℓ, j) and zero elsewhere. For each core reward,

$$V^\star_{r_v} - V_{r_v}(\pi_{\mathrm{demo}_\theta}) = p \sum_{\ell=1}^{b} \mathbb{1}\{a_\ell \ne \theta_{\ell, j_\ell}\}.$$

Choosing a mismatching response in every block gives $\Delta_{\mathcal{R}} = bp \le \Delta$. The excess under r_v is

$$\mathcal{E}^k_{r_v}(\widehat{\pi}, \pi_{\mathrm{demo}_\theta}) = p \sum_{\ell=1}^{b} \left[\mathbb{1}\{a_\ell = \theta_{\ell,j_\ell}\} - C_{(\ell,j_\ell)}(a_\ell) \right].$$

At a selected context, taking the preferred response to be the demonstrated response gives $1 - C_x(\theta_x) \ge 0$; every other choice gives a nonpositive contribution. The supremum can then choose the context with the largest miss probability independently in each block. Hence

$$\begin{aligned} Z_\theta &:= \sup_{r \in \mathcal{R}} \mathcal{E}^k_r(\widehat{\pi}, \pi_{\mathrm{demo}_\theta}) \\ &= p \sum_{\ell=1}^{b} \max_{j \in [k]} \big[1 - C_{(\ell,j)}(\theta_{\ell,j})\big], \qquad 0 \le Z_\theta \le bp. \end{aligned} \tag{31}$$

Temporarily draw the coordinates of $\Theta \in [q]^{bk}$ independently and uniformly, and generate S using $\mathcal{D}$ and $\pi_{\mathrm{demo}_\Theta}$. Let U_ℓ count the contexts in block ℓ that do not appear in the sample. Since $p \le 1/4$ and $2mp \le \log k$,

$$\lambda := \mathbb{E}U_\ell = k(1-p)^m \ge ke^{-2mp} \ge 1.$$

Two distinct contexts are both absent with probability $(1-2p)^m \le (1-p)^{2m}$, so

$$\mathbb{E}U_\ell^2 = k(1-p)^m + k(k-1)(1-2p)^m \le \lambda + \lambda^2.$$

Cauchy–Schwarz therefore yields

$$\mathbb{P}\{U_\ell > 0\} \ge \frac{(\mathbb{E}U_\ell)^2}{\mathbb{E}U_\ell^2} \ge \frac{\lambda}{1+\lambda} \ge \frac{1}{2}.$$

On $\{U_\ell > 0\}$, let J_ℓ be the smallest index of an unobserved context in that block. Conditional on S and $\widehat{\pi}$, its hidden response remains uniform on $[q]$: the observed data reveal only the coordinates at observed contexts, and the learner depends on Θ only through S. Thus, on this event,

$$\mathbb{E}\big[1 - C_{(\ell,J_\ell)}(\Theta_{\ell,J_\ell}) \mid S, \widehat{\pi}\big] = 1 - \frac{1}{q}\sum_{a=1}^{q} C_{(\ell,J_\ell)}(a) \ge \frac{1}{2}.$$

Combining this with equation 31 gives

$$\mathbb{E}Z_\Theta \ge \frac{p}{2}\sum_{\ell=1}^{b} \mathbb{P}\{U_\ell > 0\} \ge \frac{bp}{4}.$$

Since $Z_\Theta \in [0, bp]$,

$$\frac{bp}{4} \le \mathbb{E}Z_\Theta \le \frac{bp}{8} + \frac{7bp}{8}\mathbb{P}\{Z_\Theta > bp/8\},$$

and therefore $\mathbb{P}\{Z_\Theta > bp/8\} \ge 1/7$. Averaging over the finite prior selects a fixed θ for which this probability bound holds over training alone.

Using $\lfloor u \rfloor \ge u/2$ for $u \ge 1$ and $\log(2k^2) \le 3\log k$, we have

$$b\log k \ge \frac{\log|\mathcal{R}|}{6}, \qquad \frac{bp}{8} = \min\left\{\frac{\Delta}{8}, \frac{1}{16k}, \frac{b\log k}{16m}\right\} \ge \frac{1}{96}\min\left\{\Delta, \frac{1}{k}, \frac{\log|\mathcal{R}|}{m}\right\}.$$

Consequently, a fixed admissible instance satisfies

$$\mathbb{P}\left\{ Z_\theta > \frac{1}{96}\min\left\{\Delta, \frac{1}{k}, \frac{\log|\mathcal{R}|}{m}\right\}\right\} \ge \frac{1}{7}. \tag{32}$$

The confidence contribution. Use the same spaces and reward class. Set

$$\eta := \min\left\{\Delta, \frac{1}{2}, \frac{\log(1/(8\delta))}{2m}\right\}, \qquad x_\bullet := (1,1),$$

and put mass η on $x_\bullet$ and mass $1-\eta$ on context 0. For $\theta \in [q]$, let $\pi_{\mathrm{demo}_\theta}$ return θ at $x_\bullet$ and zero elsewhere. A reward selecting $x_\bullet$ and a response distinct from θ has suboptimality η, so $\Delta_{\mathcal{R}} = \eta \le \Delta$. Maximizing the excess gives

$$Z_\theta := \sup_{r\in\mathcal{R}} \mathcal{E}_r^k(\widehat{\pi}, \pi_{\mathrm{demo}_\theta}) = \eta[1 - C_{x_\bullet}(\theta)].$$

Draw Θ uniformly from $[q]$, and let E be the event that $x_\bullet$ is absent from the sample. Because $0 < \eta \le 1/2$ and $\log(1-\eta) \ge -2\eta$,

$$\mathbb{P}(E) = (1-\eta)^m \ge e^{-2m\eta} \ge 8\delta.$$

On E, Θ remains uniform conditional on $S, \widehat{\pi}$. Thus $H := 1 - C_{x_\bullet}(\Theta) \in [0,1]$ satisfies, by equation 30,

$$\frac{1}{2} \le \mathbb{E}[H \mid S, \widehat{\pi}] \le \frac{1}{4} + \frac{3}{4}\mathbb{P}\{H > 1/4 \mid S, \widehat{\pi}\}.$$

The conditional probability on the right is at least $1/3$ on E, and hence

$$\mathbb{P}\{Z_\Theta > \eta/4\} \ge \frac{\mathbb{P}(E)}{3} \ge \frac{8\delta}{3}.$$

Since $\delta < 1/16$ and $\Delta \le 1$,

$$\log\frac{1}{8\delta} \ge \frac{1}{4}\log\frac{1}{\delta}, \qquad \frac{\eta}{4} \ge \frac{1}{32}\min\left\{\Delta, \frac{\log(1/\delta)}{m}\right\}.$$

Prior averaging therefore selects a fixed admissible instance such that

$$\mathbb{P}\left\{Z_\theta > \frac{1}{32}\min\left\{\Delta, \frac{\log(1/\delta)}{m}\right\}\right\} \ge \frac{8\delta}{3}. \tag{33}$$

Combining the bounds. Both failure probabilities exceed 2δ. Use the constant $1/96$ in both bounds and select the construction with the larger threshold. Then $\max\{u,v\} \ge (u+v)/2$ gives an admissible instance satisfying

$$\mathbb{P}\left\{\sup_{r\in\mathcal{R}} \mathcal{E}_r^k(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{1}{192}\left[\min\left\{\Delta, \frac{1}{k}, \frac{\log|\mathcal{R}|}{m}\right\} + \min\left\{\Delta, \frac{\log(1/\delta)}{m}\right\}\right]\right\} \ge 2\delta. \tag{34}$$

Both constructions use the same fixed class. The context distribution and demonstrator are selected before the sample and training randomness are drawn; the supremum then evaluates the returned policy over that class.

Sample complexity. Suppose

$$0 < \varepsilon \le \frac{1}{384}\min\left\{\Delta, \frac{1}{k}\right\}.$$

For every positive integer

$$m \le \frac{\log|\mathcal{R}| + \log(1/\delta)}{384\,\varepsilon},$$

the sum of the two truncated terms in equation 34 satisfies

$$\min\left\{\Delta, \frac{1}{k}, \frac{\log|\mathcal{R}|}{m}\right\} + \min\left\{\Delta, \frac{\log(1/\delta)}{m}\right\} \ge \min\left\{\Delta, \frac{1}{k}, \frac{\log|\mathcal{R}| + \log(1/\delta)}{m}\right\} \ge 384\,\varepsilon.$$

The excess threshold in equation 34 is therefore at least 2ε, and every learner has failure probability at least $2\delta > \delta$ on an admissible instance. This includes the integer sample size at the floor of the displayed upper limit. That limit exceeds one, so a sample-complexity threshold of zero is also ruled out.

Definition 2 consequently gives

$$m_{\mathrm{rob}}^{(k)}(\mathcal{R}, \varepsilon, \delta, \Delta) > \left\lfloor \frac{\log|\mathcal{R}| + \log(1/\delta)}{384\,\varepsilon} \right\rfloor,$$

which is equation 2 for the prescribed class size. □

A.6 Proof of Theorem 8

Proof. We use the testing lemmas from Appendix A.3. All probabilities include any training randomization; prediction-time randomness is averaged in the policy values.

A fixed reward class. Fix a prescribed cardinality $|\mathcal{R}| \geq 2$, and set

$$d := \left\lfloor \frac{\log |\mathcal{R}|}{\log 2} \right\rfloor, \qquad \mathcal{X}_0 := \{0, 1, \ldots, d\}, \qquad \mathcal{Y} := \{-1, +1\}.$$

For each $v \in \{-1, +1\}^d$, define

$$f_v(0) := +1, \qquad f_v(i) := v_i \quad (i \in [d]), \qquad r_v(x, y) := \mathbb{1}\{y = f_v(x)\}.$$

These give 2^d distinct binary rewards. To reach the prescribed cardinality, add $|\mathcal{R}| - 2^d$ auxiliary contexts and extend every f_v by $+1$ there. For each auxiliary context, add the graph-indicator reward of a function that equals -1 at that context and $+1$ everywhere else. Denote the resulting space and class by $\mathcal{X}$ and $\mathcal{R}$. All distributions below give zero mass to auxiliary contexts, so the added rewards duplicate a core reward on the support. The class depends only on its cardinality, and every reward has optimal value one.

The reward-class contribution. Fix $m \geq 1$, $0 < \Delta \leq 1$, $0 < \delta < 1/16$, and an arbitrary learner. Put

$$\mathcal{D}(0) = 1 - \Delta, \qquad \mathcal{D}(i) = \frac{\Delta}{d} \quad (i \in [d]).$$

For $h \in (0, 1/4]$ and $\theta \in \{-1, +1\}^d$, let the demonstrator output $+1$ at context 0 and all auxiliary contexts, and set

$$p_{\theta,i} := \pi_{\mathrm{demo}_\theta}(+1 \mid i) = \frac{1}{2} + \theta_i h \quad (i \in [d]).$$

For each core reward,

$$V^\star_{r_v} - V_{r_v}(\pi_{\mathrm{demo}_\theta}) = \frac{\Delta}{d} \sum_{i=1}^{d} \left(\frac{1}{2} - v_i \theta_i h \right).$$

The maximum is attained at $v = -\theta$, hence

$$\Delta_{\mathcal{R}} = \Delta(1/2 + h) \leq \Delta. \tag{35}$$

Thus the construction is admissible for every budget $0 < \Delta \leq 1$.

For the realized output policy, write $q_i := \widehat{\pi}(+1 \mid i)$. Maximizing over the coordinates of v independently gives

$$Z_\theta := \sup_{r \in \mathcal{R}} \mathcal{E}^1_r(\widehat{\pi}, \pi_{\mathrm{demo}_\theta}) = (1 - \Delta)(1 - q_0) + \frac{\Delta}{d} \sum_{i=1}^{d} |p_{\theta,i} - q_i|. \tag{36}$$

The supremum therefore exposes all errors in estimating the demonstrator's response probabilities. Decode $\widehat{\theta}_i = +1$ if $q_i \geq 1/2$, and $\widehat{\theta}_i = -1$ otherwise. With $H(\widehat{\theta}, \theta) := \sum_{i=1}^{d} \mathbb{1}\{\widehat{\theta}_i \neq \theta_i\}$,

$$Z_\theta \geq \frac{\Delta h}{d} H(\widehat{\theta}, \theta). \tag{37}$$

Let P_θ be the law of one demonstration under $(\mathcal{D}, \pi_{\mathrm{demo}_\theta})$, and let $\theta^{(i)}$ differ from θ only at coordinate i. By Lemmas 14 and 17,

$$D_{\mathrm{KL}}(P_\theta^{\otimes m} \| P_{\theta^{(i)}}^{\otimes m}) = \frac{m\Delta}{d} \, \mathrm{kl}(1/2 + h, 1/2 - h) \leq \frac{16 m \Delta h^2}{d}. \tag{38}$$

Choose

$$h := \min\left\{ \frac{1}{4}, \sqrt{\frac{d}{32 m \Delta}} \right\}.$$

The divergence is at most $1/2$. Lemma 15, which also applies to randomized tests, therefore gives

$$\mathbb{P}_\theta\{\widehat{\theta}_i \neq \theta_i\} + \mathbb{P}_{\theta^{(i)}}\{\widehat{\theta}_i \neq \theta_i^{(i)}\} \geq \frac{1}{2}.$$

Under the uniform prior on $\Theta \in \{-1,+1\}^d$, averaging over neighboring pairs and summing over coordinates yields $\mathbb{E}H \geq d/4$. Since $H \leq d$,

$$\frac{d}{4} \leq \mathbb{E}H \leq \frac{d}{8} + d\,\mathbb{P}\{H > d/8\}, \qquad \mathbb{P}\{H > d/8\} \geq \frac{1}{8}.$$

By equation 37 and averaging, some fixed θ satisfies $\mathbb{P}_\theta\{Z_\theta > \Delta h/8\} \geq 1/8$. Using $d \geq \log|\mathcal{R}|/4$ gives

$$\frac{\Delta h}{8} \geq \frac{1}{128}\min\left\{\Delta, \sqrt{\frac{\Delta \log|\mathcal{R}|}{m}}\right\}.$$

Consequently,

$$\mathbb{P}_\theta\left\{Z_\theta > \frac{1}{128}\min\left\{\Delta, \sqrt{\frac{\Delta \log|\mathcal{R}|}{m}}\right\}\right\} \geq \frac{1}{8} > 2\delta. \tag{39}$$

The confidence contribution. Use the same reward class, but now put $\mathcal{D}(1) = \Delta$, $\mathcal{D}(0) = 1-\Delta$, and zero mass elsewhere. Set

$$h := \min\left\{\frac{1}{4}, \sqrt{\frac{\log(1/(8\delta))}{16m\Delta}}\right\}.$$

For $\theta \in \{-1,+1\}$, let $\pi_{\mathrm{demo}_\theta}(+1 \mid 1) = 1/2 + \theta h$ and let the demonstrator output $+1$ elsewhere. Again, $\Delta_{\mathcal{R}} = \Delta(1/2+h) \leq \Delta$. The corresponding one-demonstration laws satisfy

$$D_{\mathrm{KL}}(P_+^{\otimes m} \| P_-^{\otimes m}) \leq 16m\Delta h^2 \leq \log(1/(8\delta)).$$

Apply Lemma 16 to the test $\widehat{\theta} = +1$ if $q_1 \geq 1/2$ and -1 otherwise:

$$\frac{1}{2}\mathbb{P}_+\{\widehat{\theta} = -1\} + \frac{1}{2}\mathbb{P}_-\{\widehat{\theta} = +1\} \geq \frac{1}{4}e^{-\log(1/(8\delta))} = 2\delta.$$

Here the robust excess is

$$Z_\theta = (1-\Delta)(1-q_0) + \Delta|1/2 + \theta h - q_1|,$$

so a test error implies $Z_\theta \geq \Delta h$. Since $\delta < 1/16$ implies $\log(1/(8\delta)) \geq \frac{1}{4}\log(1/\delta)$,

$$\Delta h \geq \min\left\{\frac{\Delta}{4}, \sqrt{\frac{\Delta \log(1/\delta)}{64m}}\right\} > \frac{1}{16}\min\left\{\Delta, \sqrt{\frac{\Delta \log(1/\delta)}{m}}\right\}.$$

Thus some fixed sign satisfies

$$\mathbb{P}_\theta\left\{Z_\theta > \frac{1}{16}\min\left\{\Delta, \sqrt{\frac{\Delta \log(1/\delta)}{m}}\right\}\right\} \geq 2\delta. \tag{40}$$

Combining the bounds. Use the first construction if $\log|\mathcal{R}| \geq \log(1/\delta)$, and the second otherwise. The larger logarithm is at least half their sum, so equation 39 and equation 40 imply that some admissible instance satisfies

$$\mathbb{P}\left\{\sup_{r\in\mathcal{R}} \mathcal{E}_r^1(\widehat{\pi}, \pi_{\mathrm{demo}}) > \frac{1}{256}\min\left\{\Delta, \sqrt{\frac{\Delta[\log|\mathcal{R}| + \log(1/\delta)]}{m}}\right\}\right\} \geq 2\delta. \tag{41}$$

Both constructions use the same fixed class. The hidden parameter is selected by averaging before the sample and training randomness are drawn.

Suppose $0 < \varepsilon \leq \Delta/256$. For every integer

$$1 \leq m \leq \frac{\Delta[\log|\mathcal{R}| + \log(1/\delta)]}{262144\,\varepsilon^2},$$

the threshold in equation 41 is at least

$$\frac{1}{256}\min\{256\varepsilon, 512\varepsilon\} = \varepsilon.$$

Every learner therefore fails the robust PAC requirement at each such sample size, including the upper endpoint when it is an integer.

Zero samples also cannot suffice. In the two-sign construction, take $h = 1/4$. Without data, the decoded sign is independent of a uniform hidden sign, so its average error probability is $1/2$. For one fixed sign, the robust excess is therefore at least $\Delta/4 > \varepsilon$ with probability at least $1/2$. Definition 2 now gives

$$m^{(1)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, \Delta) > \left\lfloor \frac{\Delta[\log|\mathcal{R}| + \log(1/\delta)]}{262144\,\varepsilon^2} \right\rfloor,$$

which proves equation 3. □

Upper bound and minimax rate. Theorem 6 and Appendix D.2 of Joshi et al. (2026) give a single policy whose guarantee holds simultaneously for all rewards with demonstrator suboptimality at most Δ. Thus, whenever $\Delta_{\mathcal{R}} \le \Delta$, with probability at least $1 - \delta$,

$$\sup_{r\in\mathcal{R}} \mathcal{E}^1_r(\widehat{\pi}, \pi_{\mathrm{demo}}) \le C\left(\frac{\log(|\mathcal{R}|/\delta)}{m} + \sqrt{\frac{\Delta\log(|\mathcal{R}|/\delta)}{m}}\right)$$

for a universal constant C. Making each term at most $\varepsilon/2$ gives

$$m^{(1)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, \Delta) \lesssim \frac{\log(|\mathcal{R}|/\delta)}{\varepsilon} + \frac{\Delta\log(|\mathcal{R}|/\delta)}{\varepsilon^2}.$$

For $0 < \varepsilon \le \Delta/256$, the second term dominates. The upper bound holds uniformly over reward classes, while the construction above supplies a matching worst-case lower bound of order $\Delta\log(|\mathcal{R}|/\delta)/\varepsilon^2$.

Uniformly optimal demonstrators. For completeness, suppose $\Delta_{\mathcal{R}} = 0$. For every $r \in \mathcal{R}$,

$$\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\mathrm{demo}}(\cdot|x)}\left[\sup_z r(x,z) - r(x,y)\right] = 0.$$

The integrands are nonnegative and $\mathcal{R}$ is finite. Hence, for $\mathcal{D}$-almost every x, the demonstrator assigns probability one to responses maximizing every reward simultaneously. In particular,

$$\max_{y\in\mathcal{Y}} \sum_{r\in\mathcal{R}} r(x,y) = \sum_{r\in\mathcal{R}} \sup_{y\in\mathcal{Y}} r(x,y) \qquad \text{for } \mathcal{D}\text{-almost every } x.$$

Using the same selection convention as in the algorithms, choose $y_0(x)$ maximizing the sum of the known rewards whenever the maximum exists, and an arbitrary response otherwise. At almost every context,

$$\sum_{r\in\mathcal{R}} \left[\sup_y r(x,y) - r(x,y_0(x))\right] = 0.$$

Each summand is nonnegative, so $y_0(x)$ maximizes every reward. For any $k \ge 1$, the policy π_0 returning k copies of $y_0(x)$ therefore satisfies

$$V^k_r(\pi_0) = V^\star_r = V_r(\pi_{\mathrm{demo}}) \quad (r \in \mathcal{R}), \qquad m^{(k)}_{\mathrm{rob}}(\mathcal{R}, \varepsilon, \delta, 0) = 0.$$

This policy depends only on the known reward class and uses no demonstrations.

B OMITTED DETAILS AND PROOFS FOR SECTION 3

Online learning from demonstrations under pass@k evaluation

The learner knows the reward class $\mathcal{R}$, while the evaluation reward $r^\star \in \mathcal{R}$ is unknown.

For each round $t = 1, \dots, T$:

1. **Observe the context.** A context $x_t \in \mathcal{X}$ is revealed to the learner.
2. **Produce responses.** Using x_t and the previously observed demonstrations $\{(x_s, y_s)\}_{s<t}$, the learner outputs
$$(\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(k)}) \in \mathcal{Y}^k.$$
The current demonstration y_t is not yet available.
3. **Evaluate the responses.** The tuple is assigned the pass@k reward
$$\max_{i \in [k]} r^\star(x_t, \hat{y}_t^{(i)}).$$
This reward is not revealed to the learner. For binary rewards, a mistake occurs when all k responses have reward zero; no correctness feedback is provided.
4. **Observe the demonstration and update.** The learner receives $y_t \in \mathcal{Y}$ and uses (x_t, y_t) to update its state for the next round. The demonstrated response need not be optimal.

The online protocol allows arbitrary sequences of contexts and demonstrations; independence is not required.

Figure 1: Online learning protocol under pass@k evaluation, adapted from Joshi et al. (2026). The learner observes a demonstration only after producing its responses, without receiving evaluation reward feedback.

Algorithm 2 Online-to-batch conversion for pass@k

Input: Finite reward class $\mathcal{R}$, training sample $S = ((x_1, y_1), \dots, (x_m, y_m)) \sim (\mathcal{D} \times \pi_{\mathrm{demo}})^{\otimes m}$, response budget $k \geq 2$, discretization level $L \in \mathbb{N}$.
Output: A policy $\widehat{\pi}_{\mathrm{o2b}} : \mathcal{X} \to \Delta(\mathcal{Y}^k)$.

Initialize Algorithm 1 with reward class $\mathcal{R}$ and parameters k, L.
for $t = 1, \dots, m$ **do**
  Before processing (x_t, y_t), save a fixed copy of the current online policy
$$\widehat{\pi}_t : \mathcal{X} \to \Delta(\mathcal{Y}^k).$$
This policy depends only on the demonstrations $\{(x_s, y_s)\}_{s<t}$.
  Provide x_t to the online learner and obtain its response tuple $(\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(k)})$.
  Reveal y_t and update the online learner according to Algorithm 1.
end for
Define the learned policy as the uniform mixture
$$\widehat{\pi}_{\mathrm{o2b}}(\cdot \mid x) := \frac{1}{m} \sum_{t=1}^{m} \widehat{\pi}_t(\cdot \mid x), \qquad x \in \mathcal{X}.$$
Prediction at a new context x:
Draw $\tau \sim \mathrm{Unif}([m])$, independently of the training sample and x.
Sample the entire response tuple
$$(\hat{y}^{(1)}, \dots, \hat{y}^{(k)}) \sim \widehat{\pi}_\tau(\cdot \mid x).$$
Output $(\hat{y}^{(1)}, \dots, \hat{y}^{(k)})$.

Lemma 18 (Greedy coverage)**.** *At every round of Algorithm 1,*
$$w^{(t)}(U_t \setminus D_t) \geq k\, w^{(t)}(D_t \setminus U_t).$$

Proof of Lemma 18. For a fixed round t, let $Y_i = \{\hat{y}_t^{(1)}, \ldots, \hat{y}_t^{(i)}\}$, and by convention, we set $Y_0 = \varnothing$. Then, for $i \in \{0, \ldots, k\}$, we can define the sum of weights for the i-th greedy step from the Algorithm 1 as:

$$a_i^{(t)} := w^{(t)}\left(D_t \backslash \cup_{z \in Y_i} A_t^z\right).$$

Then, trivially we get that, $a_i^{(t)}$ are non-increasing, also by definition, we get that, $a_k^{(t)} = w^{(t)}(D_t \backslash U_t)$. Then,

$$\begin{aligned}
w^{(t)}(U_t \backslash D_t) &= w^{(t)}\left(\cup_{i=1}^k A_t^{\hat{y}_t^{(i)}} \backslash D_t\right) \\
&= \sum_{i=1}^k w^{(t)}\left(A_t^{\hat{y}_t^{(i)}} \setminus \left(D_t \cup \left(\cup_{z \in Y_{i-1}} A_t^z\right)\right)\right) \\
&= \sum_{i=1}^k \left[w^{(t)}\left(A_t^{\hat{y}_t^{(i)}} \setminus \cup_{z \in Y_{i-1}} A_t^z\right) - w^{(t)}\left((A_t^{\hat{y}_t^{(i)}} \cap D_t) \setminus \cup_{z \in Y_{i-1}} A_t^z\right)\right] \\
&\overset{(a)}{\geq} \sum_{i=1}^k \left[w^{(t)}(D_t \setminus \cup_{z \in Y_{i-1}} A_t^z) - w^{(t)}\left((A_t^{\hat{y}_t^{(i)}} \cap D_t) \setminus \cup_{z \in Y_{i-1}} A_t^z\right)\right] \\
&\overset{(b)}{=} \sum_{i=1}^k \left[a_{i-1}^{(t)} - w^{(t)}\left(D_t \setminus \cup_{z \in Y_{i-1}} A_t^z\right) + w^{(t)}\left(D_t \setminus \cup_{z \in Y_i} A_t^z\right)\right] \\
&= \sum_{i=1}^k [a_{i-1}^{(t)} - a_{i-1}^{(t)} + a_i^{(t)}] \\
&= \sum_{i=1}^k a_i^{(t)} \overset{(c)}{\geq} k a_k^{(t)} = k w^{(t)}(D_t \backslash U_t)
\end{aligned}$$

Here, (a) is due to the choice of greedy $\hat{y}_t^{(i)}$ for each i-th step. If $y_t \notin Y_{i-1}$, then by the greedy choice in the Algorithm 1, we get,

$$w^{(t)}(A_t^{\hat{y}_t^{(i)}} \setminus \cup_{z \in Y_{i-1}} A_t^z) \geq w^{(t)}(D_t \setminus \cup_{z \in Y_{i-1}} A_t^z),$$

but if $y_t \in Y_{i-1}$ then trivially the following holds

$$w^{(t)}(A_t^{\hat{y}_t^{(i)}} \setminus \cup_{z \in Y_{i-1}} A_t^z) \geq 0 = w^{(t)}(D_t \setminus \cup_{z \in Y_{i-1}} A_t^z).$$

Here (b) follows trivially from the set theoretic argument. Here, (c) is due to the non-increasing behavior of the $(a_i^{(t)})_{i \geq 0}$ sequence. □

Lemma 19. *Let*

$$W_t := \sum_{(r,j) \in \mathcal{S}_L} w^{(t)}(r, j).$$

Then the sequence $(W_t)_{t \geq 1}$ is non-increasing.

Proof of Lemma 19. The update rule given by the Algorithm 1, yields the following,

$$W_{t+1} - W_t = \frac{k}{4} w^{(t)}\left(D_t \backslash U_t\right) - \frac{1}{4} w^{(t)}\left(U_t \backslash D_t\right).$$

Using the Lemma 18, the above equation boils down to the following inequality

$$W_{t+1} - W_t = \frac{1}{4}\left(k\, w^{(t)}\left(D_t \backslash U_t\right) - w^{(t)}\left(U_t \backslash D_t\right)\right) \leq 0.$$

Therefore, we get, $W_{T+1} \leq W_T$, $\forall\, T \geq 1$. □

Corollary 20. *For all $T \geq 1$*

$$W_{T+1} \leq W_1 = L|\mathcal{R}|.$$

Proof of Corollary 20. This follows immediately from Lemma 19. One can see that

$$W_{T+1} \le W_1 = \sum_{(r,j)\in\mathcal{S}_L} w^{(1)}(r,j) \overset{(a)}{=} \sum_{(r,j)\in\mathcal{S}_L} 1 = L|\mathcal{R}|.$$

In the above, (a) is due to the initialization of the weights in the Algorithm 1. □

B.1 Proof of Theorem 9

Proof. Fix $k \ge 2$, integers $L, T \ge 1$, a reward $r \in \mathcal{R}$, and any realization of the contexts, and demonstrations. If $|\mathcal{Y}| < k$, outputting every available response and padding with repetitions gives

$$\max_{i\in[k]} r(x_t, \hat{y}_t^{(i)}) \ge r(x_t, y_t)$$

at every round. The cumulative shortfall is then nonpositive, so the claimed bound holds. Assume henceforth that $|\mathcal{Y}| \ge k$.

Recall the threshold quantities from Appendix A.1. For $t \in [T]$ and $j \in [L]$, write

$$d_t^r := r(x_t, y_t), \qquad a_t^r := \max_{i\in[k]} r(x_t, \hat{y}_t^{(i)}), \qquad u_j := \frac{j}{L},$$

and let

$$q_L(v) := \frac{1}{L}\sum_{j=1}^{L} \mathbb{1}\{v \ge u_j\} = \frac{\lfloor Lv \rfloor}{L}, \qquad v \in [0,1],$$

$$M_t^r := \frac{1}{L}\sum_{j=1}^{L} \mathbb{1}\{d_t^r \ge u_j > a_t^r\},$$

$$H_t^r := \frac{1}{L}\sum_{j=1}^{L} \mathbb{1}\{a_t^r \ge u_j > d_t^r\}.$$

By equation 8 and equation 9,

$$\begin{aligned} d_t^r - a_t^r &\le q_L(d_t^r) - q_L(a_t^r) + \frac{1}{L} \\ &= M_t^r - H_t^r + \frac{1}{L}. \end{aligned}$$

Summing over the rounds yields

$$\sum_{t=1}^{T}\left[r(x_t, y_t) - \max_{i\in[k]} r(x_t, \hat{y}_t^{(i)})\right] \le \frac{T}{L} + \sum_{t=1}^{T}(M_t^r - H_t^r). \tag{42}$$

Recall also that

$$\alpha_k := \log(1 + k/4), \qquad \beta := \log(4/3), \qquad z_t^r := \alpha_k M_t^r - \beta H_t^r.$$

Since $k \ge 2$, we have $\alpha_k \ge \log(3/2) > \beta > 0$. Together with $H_t^r \ge 0$, this gives

$$z_t^r = \alpha_k(M_t^r - H_t^r) + (\alpha_k - \beta)H_t^r \ge \alpha_k(M_t^r - H_t^r).$$

Lemma 10, which holds for every realized sequence and every discretization level L, therefore implies

$$\sum_{t=1}^{T}(M_t^r - H_t^r) \le \frac{1}{\alpha_k}\sum_{t=1}^{T} z_t^r \le \frac{\log|\mathcal{R}|}{\alpha_k}.$$

Combining this with equation 42 gives

$$\sum_{t=1}^{T}\left[r(x_t, y_t) - \max_{i\in[k]} r(x_t, \hat{y}_t^{(i)})\right] \le \frac{\log|\mathcal{R}|}{\log(1 + k/4)} + \frac{T}{L}.$$

Finally, the identity

$$1 + \frac{k}{4} - \sqrt{k} = \frac{(\sqrt{k} - 2)^2}{4} \geq 0$$

implies $\alpha_k \geq \frac{1}{2} \log k$. Consequently,

$$\sum_{t=1}^{T} \left[r(x_t, y_t) - \max_{i \in [k]} r(x_t, \hat{y}_t^{(i)}) \right] \leq \frac{2 \log |\mathcal{R}|}{\log k} + \frac{T}{L},$$

as claimed. Every step holds for each realization of the sequence, without any independence or demonstrator-optimality assumption. □

C TABLES

Table 2: Minimax PAC sample complexities for arbitrary demonstrators ($\Delta = 1$), with universal constants implicit in Θ. Upper bounds hold for every finite reward class, and matching lower bounds hold for suitable classes of the indicated cardinality. Assume $|\mathcal{R}| \geq 2$, $0 < \delta \leq 1/64$, and $0 < \varepsilon \leq c$, where $c > 0$ is a sufficiently small universal constant. For standard pass@1, assume additionally $|\mathcal{R}| \geq (1 + 1/\varepsilon)^4$ or $\delta \leq 1/|\mathcal{R}|$. For pass@$k$, standard evaluation requires $|\mathcal{R}| \geq 2k$; robust evaluation requires $|\mathcal{R}| \geq 2k^2$ and $\varepsilon \leq c/k$.

	Standard evaluation	**Robust evaluation**
pass@1	$\Theta\left(\frac{\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon^2}\right)$	$\Theta\left(\frac{\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon^2}\right)$
pass@k, $k \geq 2$	$\Theta\left[\frac{1}{\varepsilon}\left(\frac{\log\lvert\mathcal{R}\rvert}{\log k} + \log\frac{1}{\delta}\right)\right]$	$\Theta\left(\frac{\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon}\right)$

Table 3: Minimax PAC sample complexities for $0 < \Delta \leq 1$. The budget constrains $\Delta_{r^\star} \leq \Delta$ under standard evaluation and $\Delta_{\mathcal{R}} \leq \Delta$ under robust evaluation. Assume $|\mathcal{R}| \geq 2$, $0 < \delta \leq 1/64$, and $0 < \varepsilon \leq c$, where $c > 0$ is a sufficiently small universal constant. For standard pass@1, assume additionally $|\mathcal{R}| \geq (1+\Delta/\varepsilon)^4$ or $\delta \leq 1/|\mathcal{R}|$; robust pass@1 requires $\varepsilon \leq c\Delta$. For pass@k, standard evaluation requires $|\mathcal{R}| \geq 2k$; robust evaluation requires $|\mathcal{R}| \geq 2k^2$ and $\varepsilon \leq c\min\{\Delta, 1/k\}$. The Θ entries describe matching worst-case bounds over reward classes, with universal constants.

	Standard evaluation	**Robust evaluation**
pass@1	$\Theta\left(\frac{(\Delta+\varepsilon)\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon^2}\right)$	$\Theta\left(\frac{\Delta\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon^2}\right)$
pass@k, $k \geq 2$	$\Theta\left[\frac{1}{\varepsilon}\left(\frac{\log\lvert\mathcal{R}\rvert}{\log k} + \log\frac{1}{\delta}\right)\right]$	$\Theta\left(\frac{\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon}\right)$

Table 4: Minimax PAC sample complexities for optimal demonstrators ($\Delta = 0$). Under standard evaluation, optimality means $\Delta_{r^\star} = 0$ for the fixed evaluation reward; under robust evaluation, it means $\Delta_{\mathcal{R}} = 0$, that is, optimality for every reward in the class. The standard entries assume $|\mathcal{R}| \geq 2$, $0 < \delta \leq 1/64$, and $0 < \varepsilon \leq c$ for a sufficiently small universal constant $c > 0$, with $|\mathcal{R}| \geq 2k$ additionally required for pass@k. Their Θ rates are worst-case rates over reward classes, with universal constants. The robust entries are exactly zero for every finite non-empty reward class, every $k \geq 1$, $\varepsilon > 0$, and $\delta \in (0, 1)$ under class-wide optimality.

	Standard evaluation	**Robust evaluation**
pass@1	$\Theta\left(\frac{\log(\lvert\mathcal{R}\rvert/\delta)}{\varepsilon}\right)$	0
pass@k, $k \geq 2$	$\Theta\left[\frac{1}{\varepsilon}\left(\frac{\log\lvert\mathcal{R}\rvert}{\log k} + \log\frac{1}{\delta}\right)\right]$	0