REVERSE-KL POLICY DISTILLATION: STATISTICAL GUARANTEES AND LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reverse Kullback–Leibler (KL) distillation has shown empirical promise for training language models, but its statistical guarantees and limits remain poorly understood. In particular, the effects of two design choices remain unclear: Does access to the teacher's full action distribution provide an advantage over observing only the probability it assigns to the sampled action? Must training trajectories come from the student (online distillation), or can teacher-generated trajectories (offline distillation) suffice? Our work takes a first step toward understanding the theory underlying these questions.

We establish matching upper and lower bounds for online reverse-KL distillation when the teacher reveals its full action distribution, and show that substantially more data may be required when the teacher reveals only the probability of the *student*-generated action. In contrast, offline reverse-KL distillation can attain the same optimal estimation rate whether the teacher reveals its full distribution or only its probability of the *teacher*-generated action. The distinction is that teacher-generated actions themselves reveal the teacher's preferences, while queries at student-selected actions can leave those preferences unresolved. This highlights an additional statistical challenge for online distillation compared to offline distillation.

These statistical guarantees do not follow simply from optimizing standard distillation objectives: our lower bounds show that common online objectives can fail when candidates are evaluated only on past student-generated trajectories. We develop *trajectory-loss* and *importance-ratio* truncation techniques together with policy-space learning to achieve optimal online distillation guarantees.

1 INTRODUCTION

Policy distillation compresses a teacher into a student that may lack the capacity to reproduce the teacher's behavior in full. The choice of divergence therefore matters. Forward Kullback–Leibler (KL) divergence penalizes the student for assigning too little probability to teacher-supported responses, whereas reverse KL penalizes it for assigning probability mass to responses that are unlikely under the teacher. The latter allows a limited-capacity student to concentrate on teacher behavior it can represent well. Reverse KL has shown empirical promise in on-policy distillation for large language models (LLMs), where the student generates responses and learns from the teacher's action probability at the observed state (Gu et al., 2024; Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025).

Although these empirical results are promising, the optimal statistical guarantees for reverse-KL learning remain unclear. The statistical question depends first on what information the teacher provides. Some distillation methods only use the teacher's probability on student-generated actions (Lu & Thinking Machines Lab, 2025), whereas recent studies report benefits from access to the teacher's full action distribution (Li et al., 2026b; Yu et al., 2026; Fu et al., 2026). Whether this richer form of teacher feedback improves statistical efficiency, and by how much, remains unclear. Moreover, reverse KL is evaluated under the student distribution, so training typically uses student-generated responses. It is therefore unclear whether learning from offline teacher-generated data can match or even improve on the statistical guarantees attainable from online student rollouts.

To answer these questions, we study sequence-level reverse-KL learning with a student policy class that may not contain the teacher, allowing for *misspecification*. Our framework distinguishes two aspects of the learning process: who generates the training trajectories and what information the

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.

teacher provides along them. The trajectories come either from student policies chosen adaptively by the learner in the *online* setting, or from the teacher in the *offline* setting. With either data source, the teacher may reveal its entire action distribution at each visited state (*full feedback*) or only the probability of the observed action (*action feedback*). Across the resulting four settings, our goal is to compete as closely as possible with the best student in the policy class while keeping the estimation error at its minimax optimal rate.

We give a complete picture for this problem. With full feedback, we establish the minimax-optimal guarantees for reverse-KL learning under mild boundedness assumptions. For the online setting, we first show that exact refitting with common local distillation losses can be suboptimal, even when the learner selects which past trajectories to reuse. These objectives, motivated by DAgger and on-policy distillation (Ross et al., 2011; Lu & Thinking Machines Lab, 2025), evaluate candidates on prefixes generated by the rollout policies, including prefixes that a candidate itself would rarely visit, which can distort policy rankings. Our algorithm (Algorithm 1) corrects this mismatch by reweighting the observed prefixes to account for each policy's own generation behavior, attaining the optimal guarantee (Section 3). For the offline setting, we develop a new algorithm (Equation (12)) that attains the same optimal guarantee from teacher-generated trajectories.

With action feedback only, online learning can require substantially more samples than with full feedback (Section 3.4). Yet offline learning with action feedback can still obtain a comparable guarantee to full feedback learning (Sections 3.4 and 4). These results show that richer feedback can improve online sample efficiency, while surprisingly optimal reverse-KL learning need not require online interaction.

Formally, let $\pi^\star$ be the teacher policy, Π be the student policy class, and P^π the sequence distribution generated by π. Our guarantees take the form

$$\mathrm{KL}(P^{\widehat{\pi}} \| P^{\pi^\star}) \le C_{\mathrm{app}} \min_{\pi \in \Pi} \mathrm{KL}(P^{\pi} \| P^{\pi^\star}) + C_{\mathrm{est}} \frac{\log(|\Pi|/\delta)}{T}$$

with probability at least $1-\delta$, using T complete episodes and returning $\widehat{\pi} \in \Pi$. Here C_{app} (approximation factor) measures the approximation cost of competing with the best available student, and C_{est} (estimation factor) controls the estimation error. At the minimax scale of C_{est}, we characterize the smallest attainable C_{app}. Table 1 summarizes the guarantees and separations.

Table 1: Reverse-KL learning by data source and teacher feedback. Approximation coefficients are compared at the optimal estimation scale.

Data source	Feedback	Boundedness assumption	C_{app}	C_{est}
Online / Offline	Full	$\sum_{h=1}^{H} \mathrm{KL}(\pi_h(\cdot \mid s_h) \| \pi^\star_h(\cdot \mid s_h)) \le B$	$\Theta(B)$	$\Theta(B)$
Online / Offline	Full	$P^\pi(\tau) \le e^{B_\infty} P^{\pi^\star}(\tau)$	$\Theta(B_\infty)$	$\Theta(B_\infty)$
Offline	Action	$P^\pi(\tau) \le e^{B_\infty} P^{\pi^\star}(\tau)$	$\Theta(B_\infty)$	$\Theta(B_\infty)$
Online	Action	$P^\pi(\tau) \le e^{B_\infty} P^{\pi^\star}(\tau)$	—	$\Omega\left(\frac{B_\infty m}{\log(em)}\right)$

The assumptions hold for every π and feasible trajectory $\tau = (s_{1:H}, a_{1:H})$. Here $m = \min\{|\mathcal{A}|, |\Pi|, e^{B_\infty}\}$.

Related work. Our closest theoretical predecessors Rohatgi et al. (2025) and Xu et al. (2026b) also study imitation learning under misspecification. Rohatgi et al. (2025) obtain constant approximation and estimation coefficients for learning trajectory Hellinger distance with computationally inefficient algorithms, and show that standard next-token prediction approaches (e.g. cross-entropy loss) must suffer an $\Omega(H)$ approximation factor where H is the learning horizon. Xu et al. (2026b) establish a constant approximation factor for learning forward KL divergence under standard cross-entropy loss, but the estimation error scales with a boundedness factor. Both papers only study offline learning from teacher trajectories with action feedback. We study reverse-KL learning in both online and offline settings and characterize the statistical advantage of full feedback over action feedback. Appendix A provides further discussion of related work.

2 PROBLEM FORMULATION

Notation. For a positive integer n, write $[n] = \{1, \dots, n\}$. All logarithms are natural unless a base is displayed. We use $x \wedge y = \min\{x, y\}$, $x_+ = \max\{x, 0\}$, and $\mathbf{1}\{E\}$ for the indicator of an event E. The notation O, Ω, and Θ denotes upper, lower, and two-sided bounds up to positive numerical factors, and $\lesssim$ suppresses such an upper-bound factor. We write $x \asymp y$ when both $x \lesssim y$ and $y \lesssim x$. For distributions p, q, write $\mathrm{KL}(p\|q) = \sum_z p(z) \log(p(z)/q(z))$, where terms with $p(z) = 0$ equal zero and $\mathrm{KL}(p\|q) = +\infty$ if $p(z) > 0$ and $q(z) = 0$ for some z.

Setting. We consider a reward-free Markov decision process with horizon $H \geq 1$, discrete state space $\mathcal{S}$, action set $\mathcal{A}$, initial state distribution μ, and transition kernel $\mathsf{P}_h(\cdot \mid s, a)$. A Markov policy $\pi = (\pi_h)_{h=1}^H$ consists of maps $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A}) = \{p \in [0,1]^{|\mathcal{A}|} : \sum_{a \in \mathcal{A}} p(a) = 1\}$ is the probability simplex over actions. We write $\pi_h(a \mid s)$ for the probability assigned to action a at state s. Let $s_h \in \mathcal{S}$ and $a_h \in \mathcal{A}$ denote the state and action at step h. Under π, an episode generates a trajectory $\tau = (s_1, a_1, \dots, s_H, a_H) \in \mathcal{T} = (\mathcal{S} \times \mathcal{A})^H$ by sampling $s_1 \sim \mu$, $a_h \sim \pi_h(\cdot \mid s_h)$, and, for $h < H$, $s_{h+1} \sim \mathsf{P}_h(\cdot \mid s_h, a_h)$. We write P^π for the trajectory distribution under policy π, so $P^\pi(\tau) = \mu(s_1) \prod_{h=1}^H \pi_h(a_h \mid s_h) \prod_{h=1}^{H-1} \mathsf{P}_h(s_{h+1} \mid s_h, a_h)$. Let $\xi_h = (s_1, a_1, \dots, s_{h-1}, a_{h-1}, s_h)$ denote the prefix before action a_h. Its probability under π is $P^\pi(\xi_h) = \mu(s_1) \prod_{j=1}^{h-1} \pi_j(a_j \mid s_j) \mathsf{P}_j(s_{j+1} \mid s_j, a_j)$. The empty product equals one when $h = 1$.

The learner is given a finite policy class Π and seeks to imitate a fixed teacher $\pi^\star$. Define the local reverse KL between policy π and $\pi^\star$ at state s and step h by $k_h(s; \pi) = \mathrm{KL}(\pi_h(\cdot \mid s) \| \pi_h^\star(\cdot \mid s))$. Our target is the *trajectory reverse-KL risk*, whose chain-rule decomposition is $L(\pi) := \mathrm{KL}(P^\pi \| P^{\pi^\star}) = \mathbb{E}_{P^\pi}\big[\sum_{h=1}^H k_h(s_h; \pi)\big]$. Let $L^\star = \min_{\pi \in \Pi} L(\pi)$. The setting is *realizable* when $\pi^\star \in \Pi$, giving $L^\star = 0$, and *agnostic* when the teacher need not belong to the class.

Observation models. With a budget of T complete episodes, we consider four observation models that differ in their *trajectory source* and *teacher feedback*:

- Trajectory source: In the *online* setting, at each round $t \in [T]$, the learner selects $\pi_t \in \Pi$ based on past observations and executes it for a complete episode, obtaining $\tau_t \sim P^{\pi_t}$. In the *offline* setting, the learner receives independent teacher trajectories $\tau_t^\star \sim P^{\pi^\star}$ for $t \in [T]$.
- Teacher feedback: At each visited state s_h, the learner receives one of two forms of teacher feedback. *Full feedback* reveals the teacher's entire action distribution $\pi_h^\star(\cdot \mid s_h)$, whereas *action feedback* reveals only the probability $\pi_h^\star(a_h \mid s_h)$ assigned to the observed action a_h. Either feedback model can accompany online or offline trajectories, and no feedback is available at unvisited states.

Boundedness assumptions. We consider two alternative trajectory-level boundedness assumptions. Let $\mathcal{T}_{\mathrm{env}}$ denote the set of environment-feasible trajectories $\tau = (s_{1:H}, a_{1:H})$ satisfying $\mu(s_1) \prod_{h=1}^{H-1} \mathsf{P}_h(s_{h+1} \mid s_h, a_h) > 0$.

Assumption 2.1 (Pathwise accumulated KL bound)**.** There is $B \geq 1$ such that $\sum_{h=1}^H k_h(s_h; \pi) \leq B$ for every $\pi \in \Pi$ and $\tau \in \mathcal{T}_{\mathrm{env}}$.

Assumption 2.2 (Trajectory likelihood-ratio bound)**.** There is $B_\infty > 1$ such that $P^\pi(\tau) \leq e^{B_\infty} P^{\pi^\star}(\tau)$ for every $\pi \in \Pi$ and $\tau \in \mathcal{T}_{\mathrm{env}}$.

The first assumption bounds the accumulated local reverse KL along every feasible path. The second bounds the trajectory likelihood ratio between the student and teacher. Neither assumption implies the other. A uniform per-step KL bound $k_h(s; \pi) \leq G$ gives $\sum_{h=1}^H k_h(s_h; \pi) \leq HG$. Similarly, the uniform bound $\pi_h(a \mid s) \leq e^G \pi_h^\star(a \mid s)$ implies $P^\pi(\tau) \leq e^{HG} P^{\pi^\star}(\tau)$. Thus, B and B_∞ may themselves grow with H, so the absence of an explicit horizon factor in our guarantees does not imply horizon independence.

Like our Assumption 2.2, Assumption 1 of Xu et al. (2026b) also bounds likelihood ratios, but in the opposite direction $P^{\pi^\star}(\tau)/P^\pi(\tau)$ rather than our $P^\pi(\tau)/P^{\pi^\star}(\tau)$. Their assumption requires every student to cover the teacher for a forward-KL objective, whereas our Assumption 2.2 requires the teacher to cover every student for a reverse KL objective. Teacher coverage is more natural

for distillation because a limited-capacity student may concentrate on teacher-supported behaviors without representing the teacher's entire range of behaviors.

Approximation error and estimation error. Following Rohatgi et al. (2025) and Xu et al. (2026b), we seek a learner that returns $\widehat{\pi} \in \Pi$ such that, with probability at least $1-\delta$,

$$L(\widehat{\pi}) \leq \underbrace{C_{\mathrm{app}} L^\star}_{\text{approximation error}} + \underbrace{C_{\mathrm{est}} \frac{\log(|\Pi|/\delta)}{T}}_{\text{estimation error}}.$$

The coefficients may depend on $B, B_\infty, H, |\Pi|, |\mathcal{A}|$ and the observation model. Here $L^\star$ is the smallest risk achievable by any student in Π. The coefficient C_{app} measures how closely the learner competes with this best-in-class risk. The coefficient C_{est} scales the estimation term $\log(|\Pi|/\delta)/T$, determining how many episodes are needed for accurate learning. We fix C_{est} at its minimax-optimal order and characterize the smallest attainable C_{app}.

3 Online learning from student trajectories

In this section, we study online reverse-KL learning, where student policies generate trajectories and the teacher reveals its full action distribution at each visited state. We first show that exact refitting with common distillation losses can fail because it evaluates candidates under mismatched visitation distributions. We then introduce an algorithm that corrects this mismatch and attains the optimal guarantee.

3.1 Why common objectives can fail

A natural approach fits each new student to the teacher's predictions on states visited during training, as in DAgger-style refitting and the local objectives used in on-policy distillation (Ross et al., 2011; Agarwal et al., 2024). At round t, the current student π_t generates a batch of complete trajectories with full teacher feedback. Let $\mathcal{D}_t$ contain all trajectories observed so far. The learner chooses the next student by

$$\pi_{t+1} \in \arg\min_{\pi \in \Pi} Q_t(\pi; \mathcal{D}_t). \tag{1}$$

The objective may use any nonempty subset of $\mathcal{D}_t$, including only the latest batch, or assign arbitrary nonnegative weights to complete trajectories. These choices may depend on the observed data but must be the same for all candidates. If π_t is already an empirical minimizer, the learner retains it.

For example, reverse-KL refitting minimizes a weighted sum of $\sum_{h=1}^{H} k_h(s_h; \pi)$ over the selected trajectories. Such local reverse-KL losses appear in generalized knowledge distillation and recent on-policy distillation studies (Agarwal et al., 2024; Li et al., 2026a). Other choices include forward KL, mixtures of the two KL directions, and clipped KL losses (Lin et al., 2020; Wu et al., 2025; Zhao et al., 2026). Appendix B.1 specifies the allowed losses and transformations. The result concerns exact refitting with these objectives, not every gradient update used in the cited methods.

Changing which trajectories are retained does not necessarily correct the visitation mismatch. A candidate π is still evaluated at states visited by the rollout policies, rather than according to how often π itself would visit those states. In the construction below, a constant-probability event makes every observed trajectory favor a high-risk policy, so selecting a different subset cannot improve the empirical ranking.

Proposition 3.1. *For every sufficiently large B, there is an $H = 2$, $|\Pi| = 2$ instance satisfying Assumptions 2.1 and 2.2 with $B_\infty = 2B$ and $L^\star \leq e^{-\Omega(B)}$ on which the refitting rules above fail. Under uniform random initialization, after $T = \Theta(B^3)$ episodes, the policy returned by the final refit has risk $\Omega(B)$ with constant probability, even with adaptive selection and weighting of past trajectories.*

The construction has a high-risk policy π_{bad} that frequently visits a state where the near-optimal policy π_{good} has a large local loss. Although π_{good} rarely reaches this state itself, its loss on trajectories from π_{bad} makes it look worse. With constant probability, random initialization selects π_{bad} and every subsequent refit retains it, even if past data are discarded. Our algorithm reduces the influence of prefixes that the evaluated policy would rarely reach.

3.2 CORRECTING VISITATION MISMATCH

Suppose π_t generates a trajectory $(s_{1:H}, a_{1:H})$ at round $t \in [T]$. Full teacher feedback reveals the local loss $k_h(s_h; \pi)$ for every candidate π at each visited state, but the visited prefixes are sampled according to π_t rather than π. Writing $\xi_h = (s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h)$ for the prefix before action a_h, standard importance weighting multiplies a loss at ξ_h by $P^{\pi}(\xi_h)/P^{\pi_t}(\xi_h)$. Under rollouts from π_t, this changes its expected coefficient from $P^{\pi_t}(\xi_h)$ to $P^{\pi}(\xi_h)$. However, the importance ratio can be arbitrarily large when a prefix is rare under π_t, making the resulting estimate highly sensitive to rare observations.

We therefore clip the weight at one, giving $\min\{1, P^{\pi}(\xi_h)/P^{\pi_t}(\xi_h)\}$. Any loss attached to ξ_h then enters the expectation with coefficient $P^{\pi_t}(\xi_h)\min\{1, P^{\pi}(\xi_h)/P^{\pi_t}(\xi_h)\} = \min\{P^{\pi}(\xi_h), P^{\pi_t}(\xi_h)\}$. Thus, a prefix ξ_h that π visits less often than π_t is downweighted to $P^{\pi}(\xi_h)$, whereas a prefix that π visits more often is not amplified beyond $P^{\pi_t}(\xi_h)$. The ratio can be computed as $P^{\pi}(\xi_h)/P^{\pi_t}(\xi_h) = \prod_{u<h} \pi_u(a_u|s_u)/\pi_{t,u}(a_u|s_u)$ because the transition probabilities cancel.

Although each importance ratio is clipped, the local KL losses along a trajectory can still sum to a large value. We therefore consider the clipped cumulative local loss $\min\{1, \sum_{h=1}^{H} k_h(s_h; \pi)\}$. To combine this trajectory-level clipping with the step-dependent prefix weights, we decompose the clipped total loss into increments. Specifically, let $e_h(\xi_h; \pi)$ be the increase caused by adding the local loss at step h,

$$e_h(\xi_h; \pi) = \min\left\{1, \sum_{u=1}^{h} k_u(s_u; \pi)\right\} - \min\left\{1, \sum_{u=1}^{h-1} k_u(s_u; \pi)\right\}. \tag{2}$$

The increments $e_h(\xi_h; \pi)$ are nonnegative and $\sum_{h=1}^{H} e_h(\xi_h; \pi) = \min\{1, \sum_{h=1}^{H} k_h(s_h; \pi)\} \le 1$.

The two clipping operations reduce the influence of rare, high-loss observations, but they also discard information about the original reverse-KL risk. To quantify what remains, let $\overline{L}(\pi) := \mathbb{E}_{P^{\pi}}\left[\sum_{h=1}^{H} e_h(\xi_h; \pi)\right]$ denote policy π's expected clipped loss.

First, the boundedness assumptions control the cost of clipping,

$$\overline{L}(\pi) \le L(\pi) \lesssim \begin{cases} B\,\overline{L}(\pi), & \text{under Assumption 2.1,} \\ B_{\infty}\,\overline{L}(\pi), & \text{under Assumption 2.2.} \end{cases} \tag{3}$$

Thus, controlling the clipped loss $\overline{L}(\pi)$ suffices to control sequence risk $L(\pi)$, at a cost of B or B_{∞} under the respective assumption (Lemmas B.1 and D.3).

Second, the rollout policy π_t may rarely visit prefixes that are important to π. Since the importance weight is clipped, trajectories from π_t cannot fully recover π's loss on prefixes with $P^{\pi_t}(\xi_h) < P^{\pi}(\xi_h)$. The clipped importance sampling ratios retain each increment $e_h(\xi_h; \pi)$ with probability mass $\min\{P^{\pi}(\xi_h), P^{\pi_t}(\xi_h)\}$ rather than its full mass $P^{\pi}(\xi_h)$ in $\overline{L}(\pi)$. More generally, for any $\pi, \pi' \in \Pi$, we define

$$\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] := \mathbb{E}_{P^{\pi'}}\left[\sum_{h=1}^{H} \min\left\{1, \frac{P^{\pi}(\xi_h)}{P^{\pi'}(\xi_h)}\right\} e_h(\xi_h; \pi)\right] = \sum_{h, \xi_h} \min\{P^{\pi}(\xi_h), P^{\pi'}(\xi_h)\}\, e_h(\xi_h; \pi). \tag{4}$$

Here, π' in the subscript is the rollout policy, while π inside $\overline{L}(\pi)$ is the policy being evaluated. In episode t, we take $\pi' = \pi_t$. Since $\mathrm{Trun}_{\pi_t}[\overline{L}(\pi)]$ retains only the prefix mass shared by π and π_t, it may underestimate $\overline{L}(\pi)$. Indeed, $0 \le \mathrm{Trun}_{\pi_t}[\overline{L}(\pi)] \le \overline{L}(\pi)$. The key fact is that when $\overline{L}(\pi_t)$ is small, the truncated quantity still controls $\overline{L}(\pi)$. Lemma B.1 gives

$$\overline{L}(\pi) \le 64\left(\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi)\right] + \overline{L}(\pi_t)\right). \tag{5}$$

Equation (5) shows why the quality of the rollout policy matters. When $\overline{L}(\pi_t)$ is small, a candidate π with substantially larger $\overline{L}(\pi)$ must also have a large $\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi)\right]$. Thus, even after truncation, losses observed on rollouts from π_t provide evidence against π. Since the learner does not know in advance how to choose a rollout policy π_t with small $\overline{L}(\pi_t)$, the observed losses must guide both candidate evaluation and the choice of future rollout policies.

3.3 ONLINE LEARNING FOR REVERSE KL MINIMIZATION

To let the observed losses guide future rollouts, our algorithm assigns each candidate π a weight $w_t(\pi)$ and samples $\pi_t \sim w_t$ in episode t. After executing π_t, the learner observes the trajectory $\tau_t = (s_{t,1:H}, a_{t,1:H})$ and full teacher feedback $\pi^\star(\cdot|s_{t,h})$ for every $h \in [H]$. These observations are used to compare every $\pi \in \Pi$ with π_t. At the observed prefix $\xi_{t,h} = (s_{t,1}, a_{t,1}, \cdots, s_{t,h})$, the loss $\ell_t(\pi)$ subtracts $e_h(\xi_{t,h}; \pi_t)$ from $e_h(\xi_{t,h}; \pi)$ and applies the same clipped prefix importance sampling weight to both terms, yielding

$$\ell_t(\pi) = \sum_{h=1}^{H} \min\left\{1, \frac{P^\pi(\xi_{t,h})}{P^{\pi_t}(\xi_{t,h})}\right\} \big(e_h(\xi_{t,h}; \pi) - e_h(\xi_{t,h}; \pi_t)\big). \tag{6}$$

Let $\mathbb{E}_t$ denote expectation conditioned on the history before round t, averaging over both π_t and its trajectory. We have

$$\mathbb{E}_t\left[\ell_t(\pi)\right] = \mathbb{E}_{\pi_t \sim w_t}\left[\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi)\right] - \mathrm{Trun}_{\pi}\left[\overline{L}(\pi_t)\right]\right]. \tag{7}$$

Thus in expectation, $\ell_t(\pi)$ compares the losses of π and π_t under prefix probability mass $\min\{P^\pi(\xi_{t,h}), P^{\pi_t}(\xi_{t,h})\}$. From clipping, each of the two weighted sums in (6) lies in $[0, 1]$, so $\ell_t(\pi) \in [-1, 1]$.

Our algorithm is Algorithm 1. The exponential-weights update turns $\ell_t(\pi)$ into a change in sampling probability. Since $\ell_t(\pi_t) = 0$, a positive $\ell_t(\pi)$ decreases the weight of π relative to π_t, while a negative value increases the weight of π. Candidates with smaller accumulated losses $\sum_{u=1}^{t} \ell_u(\pi)$ receive larger $w_{t+1}(\pi)$ and are therefore more likely to generate subsequent trajectories. The algorithm outputs the policy with the largest final weight w_{T+1}.

Algorithm 1 Online learning by bounded prefix comparisons

Require: Policy class Π, episode budget T.

1: Set $w_1(\pi) = 1/|\Pi|$ for all $\pi \in \Pi$ and $\eta = 1/256$.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $\pi_t \sim w_t$ and execute it for the entire episode.
4: Observe the trajectory $\tau_t = (s_{t,1:H}, a_{t,1:H})$ and the teacher distribution $\pi^\star(\cdot|s_{t,h})$ for every $h \in [H]$.
5: For every $\pi \in \Pi$, compute the loss $\ell_t(\pi)$ using (6).
6: $w_{t+1}(\pi) = \dfrac{w_t(\pi)\exp(-\eta\ell_t(\pi))}{\sum_{\pi' \in \Pi} w_t(\pi')\exp(-\eta\ell_t(\pi'))}$ for all $\pi \in \Pi$.
7: **end for**
8: Return $\widehat{\pi} \in \arg\max_{\pi \in \Pi} w_{T+1}(\pi)$, breaking ties by the smallest candidate index.

The following theorem shows that Algorithm 1 achieves the optimal estimation scale while competing with $L^\star$.

Theorem 3.2 (Online full-feedback guarantee)**.** *For every $\delta \in (0, 1)$, Algorithm 1 returns a candidate satisfying, with probability at least $1 - \delta$,*

$$L(\widehat{\pi}) \le \begin{cases} O\left(B\left[L^\star + \frac{\log(|\Pi|/\delta)}{T}\right]\right), & \textit{under Assumption 2.1}, \\ O\left(B_\infty\left[L^\star + \frac{\log(|\Pi|/\delta)}{T}\right]\right), & \textit{under Assumption 2.2}. \end{cases} \tag{8}$$

The $O(\cdot)$ notation hides only universal numerical constants.

Theorem 3.3 establishes the tightness of both coefficients in (8).

Theorem 3.3 (Optimality of full-feedback learning)**.** *For both online and offline learning with full feedback, the minimax estimation coefficient is $C_{\mathrm{est}} = \Theta(B)$ under Assumption 2.1 in the realizable setting. Any learner achieving $C_{\mathrm{est}} = O(B)$ on all instances must incur $C_{\mathrm{app}} = \Omega(B)$ in the worst case. Under Assumption 2.2, the same conclusions hold with B_∞ replacing B.*

Together, Theorems 3.2 and 3.3 show that Algorithm 1 attains the optimal estimation rate and the smallest possible order of approximation coefficient at that rate. The estimation and approximation lower bounds are proved in Appendices B.4.1 and B.4.2, respectively. Below, we give a proof sketch for the upper bound.

Proof sketch of Theorem 3.2. By the standard second-order exponential-weights analysis (Lattimore & Szepesvári, 2020, Eq. (27.2)–(27.3)), for a fixed minimizer $\pi^\dagger \in \arg\min_{\pi\in\Pi} \overline{L}(\pi)$, $\ell_t(\pi) \in [-1,1]$, and $\eta \le 1$, we have

$$-\underbrace{\sum_{t=1}^{T} \ell_t(\pi^\dagger)}_{\textbf{benchmark loss}} \le \frac{\log|\Pi|}{\eta} - \underbrace{\sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)}_{\textbf{weighted loss}} + \underbrace{\eta\sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2}_{\textbf{quadratic term}}. \tag{9}$$

We further bound the conditional expectation of the three terms above. For the **benchmark loss**,

$$\begin{aligned}
\mathbb{E}_t\big[\ell_t(\pi^\dagger)\big] &= \mathbb{E}_{\pi_t\sim w_t}\left[\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi^\dagger)\right] - \mathrm{Trun}_{\pi^\dagger}\left[\overline{L}(\pi_t)\right]\right] && \text{(Equation 7)}\\
&\le L^\star + \overline{L}(\pi^\dagger) - \frac{1}{64}\mathbb{E}_{\pi_t\sim w_t}\left[\overline{L}(\pi_t)\right] \\
&&& \left(\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi^\dagger)\right] \le \overline{L}(\pi^\dagger) \le L^\star \text{ and Equation 5}\right)\\
&\le 2L^\star - \frac{1}{64}\sum_{\pi\in\Pi} w_t(\pi)\overline{L}(\pi). && (\overline{L}(\pi^\dagger) \le L^\star)
\end{aligned}$$

The **weighted loss** has zero conditional expectation:

$$\mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)\right] = \sum_{\pi,\pi_t\in\Pi} w_t(\pi)w_t(\pi_t)\Big(\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi)\right] - \mathrm{Trun}_{\pi}\left[\overline{L}(\pi_t)\right]\Big) = 0. \tag{10}$$

For the **quadratic term**, since $e_h(\xi_{t,h};\pi) \in [0,1]$, we have

$$\begin{aligned}
\mathbb{E}_t\big[\ell_t(\pi)^2\big] &\le \mathbb{E}_t\left[\sum_{h=1}^{H} \min\left\{1, \frac{P^\pi(\xi_{t,h})}{P^{\pi_t}(\xi_{t,h})}\right\}\big(e_h(\xi_{t,h};\pi) + e_h(\xi_{t,h};\pi_t)\big)\right]\\
&= \mathbb{E}_{\pi_t\sim w_t}\left[\mathrm{Trun}_{\pi_t}\left[\overline{L}(\pi)\right] + \mathrm{Trun}_{\pi}\left[\overline{L}(\pi_t)\right]\right] \le \overline{L}(\pi) + \mathbb{E}_{\pi_t\sim w_t}\left[\overline{L}(\pi_t)\right].
\end{aligned}$$

Weighting by $w_t(\pi)$ yields $\mathbb{E}_t\big[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2\big] \le 2\sum_{\pi\in\Pi} w_t(\pi)\overline{L}(\pi)$.

Combining everything above with Equation (9) gives

$$\frac{1}{64}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\overline{L}(\pi)\right] \le 2TL^\star + \frac{\log|\Pi|}{\eta} + 2\eta\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\overline{L}(\pi)\right].$$

With $\eta = 1/256$ as in Algorithm 1, self-bounding yields $\mathbb{E}\big[\frac{1}{T}\sum_t\sum_\pi w_t(\pi)\overline{L}(\pi)\big] \le O\big(L^\star + \frac{\log|\Pi|}{T}\big)$, which further implies $\mathbb{E}\big[\frac{1}{T}\sum_t\sum_\pi w_t(\pi)L(\pi)\big] \le O\big(B\big[L^\star + \frac{\log|\Pi|}{T}\big]\big)$ by (3) under Assumption 2.1, and with B_∞ replacing B under Assumption 2.2.

Since $\pi_t \sim w_t$, this gives an expected-risk bound when the output policy $\widehat{\pi}$ is uniformly sampled from $\{\pi_1,\cdots,\pi_T\}$. Our Algorithm 1 instead returns $\widehat{\pi} \in \arg\max_{\pi\in\Pi} w_{T+1}(\pi)$ for a high-probability bound. Its high-probability analysis uses the same second-moment bounds that control the quadratic term to bound accumulated sampling fluctuations. The full high-probability analysis appears in Appendix B.2. □

Extension to Hellinger distance. The same algorithmic approach also applies to Hellinger distance $D^2_{\mathrm{H}}(P,Q) = \sum_z(\sqrt{P(z)} - \sqrt{Q(z)})^2$. We define policy π's local loss on prefix ξ_h as $g_h(\xi_h;\pi) = \sqrt{\frac{P^{\pi^\star}(\xi_h)}{P^{\pi}(\xi_h)}}D^2_{\mathrm{H}}(\pi_h(\cdot|s_h),\pi^\star_h(\cdot|s_h))$ with $\mathbb{E}_{P^\pi}\big[\sum_{h=1}^H g_h(\xi_h;\pi)\big] = D^2_{\mathrm{H}}(P^\pi,P^{\pi^\star})$. We also clip the cumulative local loss by letting $e^{\mathrm{H}}_h(\xi_h;\pi) = \min\{1,\sum_{u=1}^h g_u(\xi_u;\pi)\} - \min\{1,\sum_{u=1}^{h-1} g_u(\xi_u;\pi)\}$. The algorithm then uses losses $\ell_t(\pi) = \sum_{h=1}^H \min\big\{1, \frac{P^\pi(\xi_{t,h})}{P^{\pi_t}(\xi_{t,h})}\big\}\big(e^{\mathrm{H}}_h(\xi_{t,h};\pi) - e^{\mathrm{H}}_h(\xi_{t,h};\pi_t)\big)$ in the same exponential-weights update as Algorithm 1. The full algorithm for online Hellinger distance learning is provided in Algorithm 2. Theorem 3.4 establishes constant approximation and estimation coefficients for online learning with full feedback, matching the guarantee of Rohatgi et al. (2025) for offline learning from teacher trajectories.

Theorem 3.4. *Let $\widehat{\pi} \in \Pi$ be the output policy of Algorithm 2. With probability at least $1-\delta$, $D^2_{\mathrm{H}}(P^{\widehat{\pi}},P^{\pi^\star}) \lesssim \min_{\pi\in\Pi} D^2_{\mathrm{H}}(P^\pi,P^{\pi^\star}) + \log(|\Pi|/\delta)/T$.*

Under Assumption 2.2, every policy $\pi \in \Pi$ satisfies $D_{\mathrm{H}}^2(P^\pi, P^{\pi^\star}) \le \mathrm{KL}\left(P^\pi \| P^{\pi^\star}\right) \lesssim B_\infty D_{\mathrm{H}}^2(P^\pi, P^{\pi^\star})$ because $(\sqrt{x}-1)^2 \le x\log x - x + 1 \lesssim B_\infty(\sqrt{x}-1)^2$ for $0 \le x \le e^{B_\infty}$. Thus, the guarantee of Hellinger distance can be transferred to reverse KL divergence at the cost of B_∞. Thus, Theorem 3.4 directly implies $L(\widehat{\pi}) \le O(B_\infty\,[L^\star + \log(|\Pi|/\delta)/T])$. However, a similar reduction does not hold under Assumption 2.1, so Algorithm 1 provides more general guarantees.

3.4 The Limitation of Action Feedback in the Online Setting

Algorithm 1 uses the teacher's full action distribution to evaluate the local KL losses of every $\pi \in \Pi$. With action feedback, the learner only observes the teacher's probability on its chosen action, while the probabilities of other actions remain unobserved. The following lower bound shows that this missing information can significantly increase the sample requirement of any learner, even in the realizable setting.

Theorem 3.5 (Online action-feedback lower bound)**.** *For a sufficiently large T, every online learner returning a policy $\widehat{\pi} \in \Pi$ with action feedback from the teacher has a realizable instance satisfying Assumption 2.2, on which $L(\widehat{\pi}) = \widetilde{\Omega}\Big(\min\{|\mathcal{A}|, |\Pi|, e^{B_\infty}\}\frac{B_\infty \log|\Pi|}{T}\Big)$ with probability at least $1/4$.*

Compared with the full-feedback guarantee in Theorem 3.2, this lower bound requires an additional factor of $\min\{|\mathcal{A}|, |\Pi|, e^{B_\infty}\}$ in the estimation error even in the realizable setting. Full feedback therefore provides a statistical advantage over action feedback in the online learning setting.

4 Offline learning from teacher trajectories

In Section 3, the learner chooses a rollout policy in each round and uses the observed losses to guide future rollouts. In the offline setting, the learner instead receives T independent teacher trajectories $\tau_t^\star \sim P^{\pi^\star}$ with $\tau_t^\star = (s_{t,1}^\star, a_{t,1}^\star, \cdots, s_{t,H}^\star, a_{t,H}^\star)$. We will discuss algorithms to deal with both full feedback and action feedback for the offline setting in this section.

4.1 Offline learning with full feedback

Full feedback reveals $k_h(s_{t,h}^\star; \pi)$ for every candidate π at each teacher-visited state. A natural estimator minimizes the average accumulated local KL over the teacher trajectories,

$$\widehat{\pi}_{\mathrm{raw}} \in \arg\min_{\pi\in\Pi} \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H} k_h(s_{t,h}^\star; \pi). \tag{11}$$

However, Proposition 4.1 shows that although this algorithm can obtain the optimal estimation rate in the realizable setting, it suffers $C_{\mathrm{app}} = \widetilde{\Omega}(B^2)$ approximation error rate under misspecification, which is suboptimal compared with the lower bound in Theorem 3.3.

Proposition 4.1 (Unclipped offline ERM)**.** *Under Assumption 2.1, if $\pi^\star \in \Pi$, then for every $\delta \in (0,1)$, with probability at least $1-\delta$, $L(\widehat{\pi}_{\mathrm{raw}}) \le O(B\log(|\Pi|/\delta)/T)$. Under misspecification, however, a uniform guarantee for* (11) *with an estimation term vanishing as $T \to \infty$ requires $C_{\mathrm{app}} = \widetilde{\Omega}(B^2)$.*

The failure under misspecification arises because teacher trajectories may visit states that a candidate π never reaches. Large local KL losses at these states do not contribute to $L(\pi)$, but can dominate the objective 11 and cause it to reject π even when $L(\pi)$ is small. Similar to the online setting, our estimator limits each trajectory's contribution by clipping the accumulated loss, yielding

$$\widehat{\pi} \in \arg\min_{\pi\in\Pi} \frac{1}{T}\sum_{t=1}^{T} \min\left\{1, \sum_{h=1}^{H} k_h(s_{t,h}^\star; \pi)\right\}. \tag{12}$$

Theorem 4.2 (Offline full-feedback guarantee)**.** *The estimator in* (12) *satisfies the high-probability guarantee of Theorem 3.2 under either Assumption 2.1 or Assumption 2.2.*

Under Assumption 2.1, clipping preserves the optimal estimation rate in the realizable setting while improving the approximation coefficient from the $\widetilde{\Omega}(B^2)$ required by the unclipped estimator to $O(B)$, matching the lower bound in Theorem 3.3.

We now provide a proof sketch for Theorem 4.2. Let $\widehat{\mu}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \min\left\{1, \sum_{h=1}^{H} k_h(s^\star_{t,h};\pi)\right\}$ and $\mu(\pi) = \mathbb{E}_{P^{\pi^\star}}\left[\min\left\{1, \sum_{h=1}^{H} k_h(s^\star_{t,h};\pi)\right\}\right]$. With probability at least $1-\delta$, standard Bernstein concentration for bounded nonnegative losses gives for all $\pi \in \Pi$,

$$\widehat{\mu}(\pi) \le 3\mu(\pi) + O\left(\frac{\log(|\Pi|/\delta)}{T}\right), \qquad \mu(\pi) \le 3\widehat{\mu}(\pi) + O\left(\frac{\log(|\Pi|/\delta)}{T}\right).$$

Since $\widehat{\pi}$ minimizes $\widehat{\mu}$ in Equation 12, these inequalities give

$$\mu(\widehat{\pi}) \le 3\widehat{\mu}(\widehat{\pi}) + O\left(\frac{\log(|\Pi|/\delta)}{T}\right) = 3\min_{\pi\in\Pi}\widehat{\mu}(\pi) + O\left(\frac{\log(|\Pi|/\delta)}{T}\right) \le 9\min_{\pi\in\Pi}\mu(\pi) + O\left(\frac{\log(|\Pi|/\delta)}{T}\right).$$

To get the target risk bound on $L(\pi)$, we need to connect $\mu(\pi)$ with $\overline{L}(\pi)$, and bound $L(\pi)$ through Equation (3). The challenge is that $\mu(\pi)$ averages losses under $P^{\pi^\star}$, whereas $\overline{L}(\pi)$ averages under P^{π}. After clipping, this change of measure costs only a constant factor. Lemma B.1 establishes $\mu(\pi) \asymp \overline{L}(\pi)$. Thus, with probability at least $1-\delta$,

$$\overline{L}(\widehat{\pi}) \lesssim \mu(\widehat{\pi}) \lesssim \min_{\pi\in\Pi}\mu(\pi) + \frac{\log(|\Pi|/\delta)}{T} \lesssim \min_{\pi\in\Pi}\overline{L}(\pi) + \frac{\log(|\Pi|/\delta)}{T} \le L^\star + \frac{\log(|\Pi|/\delta)}{T}.$$

Equation (3) converts this bound to $L(\widehat{\pi})$, introducing a factor B under Assumption 2.1 or B_∞ under Assumption 2.2. Appendices B.2.1 and C give the full change-of-measure and concentration arguments.

4.2 Offline learning with action feedback

Action feedback reveals the teacher's probability only for the observed action, so the local KL losses in (12) are generally unavailable. Even under realizability, Assumption 2.1 alone does not permit a uniformly vanishing risk bound for offline learners with action feedback (Proposition C.1). We instead only consider Assumption 2.2 and use the offline Hellinger estimator of Rohatgi et al. (2025), which returns $\widehat{\pi} \in \Pi$ satisfying, with probability at least $1-\delta$, $D^2_{\mathrm{H}}(P^{\widehat{\pi}}, P^{\pi^\star}) \lesssim \min_{\pi\in\Pi} D^2_{\mathrm{H}}(P^{\pi}, P^{\pi^\star}) + \log(|\Pi|/\delta)/T$. Under Assumption 2.2, the Hellinger-to-KL conversion in Section 3.3 gives $L(\widehat{\pi}) \lesssim B_\infty D^2_{\mathrm{H}}(P^{\widehat{\pi}}, P^{\pi^\star})$ and $\min_{\pi\in\Pi} D^2_{\mathrm{H}}(P^{\pi}, P^{\pi^\star}) \le L^\star$. Substituting these inequalities gives the following guarantee.

Proposition 4.3 (Offline action-feedback guarantee)**.** *Under Assumption 2.2, let $\widehat{\pi} \in \Pi$ be the output of the estimator in Equation (11) of Rohatgi et al. (2025). For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$L(\widehat{\pi}) \le O\left(B_\infty\left[L^\star + \frac{\log(|\Pi|/\delta)}{T}\right]\right). \tag{13}$$

The lower bounds in Theorem 3.3 also hold for offline learning with action feedback under Assumption 2.2. Thus, the reduction from Hellinger to reverse KL learning is minimax optimal in this setting. Proposition 4.3 and Theorem 3.5 show that, under Assumption 2.2, offline action-feedback learning can require exponentially fewer episodes than online action-feedback learning in the worst case.

5 Conclusion

We characterize how data source and teacher feedback shape reverse-KL learning under misspecification. For the online learning setting, we show the failure of common empirical algorithms and provide the first algorithm achieving the optimal guarantee under full feedback. We further show that full feedback can significantly improve online sample efficiency over action feedback. For the offline setting, we also establish algorithms with tight bounds, demonstrating that teacher trajectories can avoid the additional sample cost of online learning with action feedback. Future work includes developing and empirically evaluating practical distillation algorithms based on our prefix reweighting and trajectory clipping mechanism, and determining the optimal statistical guarantees achievable by computationally efficient learners.

AI USE STATEMENT

An OpenAI assistant was used for manuscript editing, lemma development and consistency checks. The authors independently verified the mathematics, references, and final submission.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/07d5938693cc3903b261e1a3844590ed-Abstract.html.

Nicolas Espinosa-Dice, Sanjiban Choudhury, Wen Sun, and Gokul Swamy. Efficient imitation under misspecification. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ca96d9b87059a0b408ec1a1cd0ab46ec-Abstract-Conference.html.

Dylan J. Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/da84e39ae51fd26bb5110d9659c06e13-Abstract-Conference.html.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562v2*, 2026. URL https://arxiv.org/abs/2603.25562v2.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401. URL https://tor-lattimore.com/downloads/book/book.pdf.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016v2*, 2026a. URL https://arxiv.org/abs/2604.13016v2.

Yichen Li and Chicheng Zhang. Agnostic interactive imitation learning: New theory and practical algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24ck.html.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. *arXiv preprint arXiv:2608.06296*, 2026b. URL https://arxiv.org/abs/2608.06296.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Kevin Lu and Thinking Machines Lab. On-policy distillation. Thinking Machines Lab: Connectionism, October 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Technical blog, October 27, 2025.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2024. URL https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf. Author manuscript, August 16, 2024.

Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the Fundamental Limits of Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1e7875cf32d306989d80c14308f3a099-Abstract.html.

Nived Rajaraman, Yanjun Han, Lin F. Yang, Kannan Ramchandran, and Jiantao Jiao. Provably breaking the quadratic error compounding barrier in imitation learning, optimally. *arXiv preprint arXiv:2102.12948*, 2021. URL https://arxiv.org/abs/2102.12948.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv:2502.12465v1*, 2025. URL https://arxiv.org/abs/2502.12465v1. Full version. Extended abstract in Proceedings of Thirty Eighth Conference on Learning Theory, PMLR 291:4831–4837, 2025.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pp. 661–668. PMLR, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy Distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

Igal Sason and Sergio Verdú. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151. URL https://arxiv.org/abs/1508.00335v7.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback. *arXiv preprint arXiv:2606.30923*, 2026. URL https://arxiv.org/abs/2606.30923.

Luca Viano, Antoine Moulin, Audrey Huang, Volkan Cevher, Philip Amortila, and Dylan J. Foster. When does on-policy interaction help? representational tradeoffs in value-based imitation learning. *arXiv preprint arXiv:2607.29617*, 2026. URL https://arxiv.org/abs/2607.29617.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback–Leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025. URL https://aclanthology.org/2025.coling-main.383/.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Yunbei Xu, Yuzhe Yuan, and Ruohan Zhan. Autoregressive learning in joint KL: Sharp oracle bounds and lower bounds. *arXiv preprint arXiv:2605.12316*, 2026b. URL https://arxiv.org/abs/2605.12316.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv preprint arXiv:2608.09836*, 2026. URL https://arxiv.org/abs/2608.09836.

Huaqing Zhang, Jingchu Gai, Juno Kim, Bingbin Liu, and Andrej Risteski. When Does Online Imitation Learning Help in LLM Post-Training? The Role of (Non-)Realizability Beyond Horizon. *arXiv preprint arXiv:2606.30445*, 2026. URL https://arxiv.org/abs/2606.30445.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734v3*, 2026. URL https://arxiv.org/abs/2601.18734v3.

APPENDIX CONTENTS

A Related work **14**

B Online learning **15**

B.1 Failure of exact empirical fitting 15

B.2 Online reverse-KL learning 17

B.2.1 Clipping and prefix truncation 17

B.2.2 Exponential-weights analysis 21

B.3 Online Hellinger learning from full feedback 24

B.4 Estimation and approximation lower bounds 29

B.4.1 Estimation lower bound 29

B.4.2 Approximation lower bound 30

B.5 The additional cost of online action feedback 32

C Offline learning **33**

C.1 Full feedback 33

C.2 Action feedback 35

D Auxiliary lemmas **36**

D.1 KL and Hellinger inequalities 36

D.2 Concentration and testing bounds 37

A Related work

Offline imitation learning theory Offline imitation learning asks what can be learned from a fixed set of expert demonstrations without further interaction. In the realizable setting, classical analyses of zero-one-loss behavior cloning relate action-prediction errors to compounding errors in task performance (Ross & Bagnell, 2010). Rajaraman et al. (2020) characterize minimax task-performance guarantees in finite-state MDPs, showing that error compounding can be unavoidable when dynamics are unknown. With known dynamics and a deterministic expert, improved dependence on the horizon is possible (Rajaraman et al., 2021). For general policy classes under realizability, Foster et al. (2024) show that log-loss behavior cloning admits horizon-free guarantees when the cumulative reward range and policy-class complexity are controlled. Additional offline data provide another source of information: MILO learns a dynamics model from auxiliary transition data and mitigates covariate shift when those data sufficiently cover the expert's behavior (Chang et al., 2021). Our offline learners instead use teacher trajectories and the specified probability feedback, without auxiliary transition data.

When the student policy class is not expressive enough to cover the teacher, the question becomes how closely a learner can compete with the best available policy. A common target is an oracle bound of the form $D(\widehat{\pi};\pi^\star) \le C_{\mathrm{app}} \min_{\pi\in\Pi} D(\pi;\pi^\star) + C_{\mathrm{est}} \log(|\Pi|/\delta)/T$ where $D(\cdot;\cdot)$ can be any distance measure. For trajectory-level Hellinger distance, Rohatgi et al. (2025) give an offline estimator with constant approximation and estimation coefficients, while establishing an $\Omega(H)$ approximation barrier for a broad class of next-token methods. Xu et al. (2026b) study trajectory-level forward KL divergence, obtaining constant-factor approximation error with $\Theta(HG)$ estimation error where $\sup_{\pi\in\Pi}\sup_{h\in[H]}\sup_{s_h} \log \frac{\pi_h^\star(x_h|s_h)}{\pi_h(x_h|s_h)} \le G$. Both Rohatgi et al. (2025) and Xu et al. (2026b) study offline learning with action feedback. In the same setting, we study reverse KL learning and show that $C_{\mathrm{app}}, C_{\mathrm{est}} = \Theta(B_\infty)$ under Assumption 2.2. Due to the different directions in forward and reverse KL, we assume the teacher can cover any student, whereas Xu et al. (2026b) assumes any student can cover the teacher. Teacher coverage is more natural for distillation because a limited-capacity student may concentrate on teacher-supported behaviors without representing the teacher's entire range of behaviors.

Interactive imitation learning A student may visit states rarely represented in teacher demonstrations, especially after making a mistake (Ross & Bagnell, 2010). DAgger obtains expert supervision on these states and reduces imitation to no-regret online learning (Ross et al., 2011). Li & Zhang (2024) give an oracle-efficient agnostic algorithm with access to exploratory state samples. Its comparator measures local imitation error on states visited during training; ours evaluates each candidate under its own trajectory law. Thus a small online regret alone does not imply our reverse-KL oracle bound.

Several works identify conditions under which interaction improves imitation-learning performance. Espinosa-Dice et al. (2025) avoid quadratic error compounding under a completeness condition on the policy class. Zhang et al. (2026) identify an offline information bottleneck under misspecification even at horizon one. Sriraman et al. (2026) establish an exponential benefit from interaction with a noisy expert when performance is measured against a clean expert. Viano et al. (2026) obtain efficient on-policy imitation when the expert's Q-function, rather than its policy, is realizable. These guarantees concern different targets and assumptions from our comparison of online and offline reverse-KL learning rates.

Language-model distillation Building on knowledge and policy distillation (Hinton et al., 2015; Rusu et al., 2016), language-model distillation can use teacher-generated responses (Kim & Rush, 2016) or student-generated responses (Lin et al., 2020; Agarwal et al., 2024; Li et al., 2026a), with forward KL and reverse KL as common losses. Beyond the standard losses, Wu et al. (2025) propose an adaptive combination of these two losses. Ko et al. (2024) propose skew KL losses that compare a distribution with a mixture of teacher and student distributions, while Ko et al. (2025) use a contrastive objective to increase the likelihood of teacher responses and decrease that of student responses. Gu et al. (2024) develop policy-gradient updates for sequence-level reverse KL, using returns based on accumulated future teacher-to-student log likelihood ratios. These returns allow later predictions to influence updates to earlier actions.

The feedback available at each training recipe is a separate choice. Lu & Thinking Machines Lab (2025) use teacher probabilities for sampled tokens, Xu et al. (2026a) use full next-token distributions, and Fu et al. (2026); Yu et al. (2026) incorporate information about teacher-preferred alternatives. Our full-versus-action-feedback comparison isolates the statistical value of the available information by comparing the best attainable rates under each feedback model.

B Online learning

B.1 Failure of exact empirical fitting

At round t, the current policy π_t generates a nonempty batch of H-step trajectories with full teacher feedback. Let $\mathcal{D}_t$ contain all trajectories collected up to and including round t. The learner assigns weights $\omega_{t,\tau} \geq 0$ with $\sum_{\tau \in \mathcal{D}_t} \omega_{t,\tau} > 0$ and refits by

$$\begin{aligned} \pi_{t+1} &\in \arg\min_{\pi \in \Pi} Q_t(\pi; \mathcal{D}_t), \\ Q_t(\pi; \mathcal{D}_t) &= \sum_{\tau \in \mathcal{D}_t} \omega_{t,\tau} \phi_t \left(\sum_{h=1}^{H} d_t(\pi_h(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h)) \right). \end{aligned} \tag{14}$$

The weights may depend on the observed history but are the same for all candidates. Assigning weight zero discards a trajectory, so this rule includes fitting any nonempty subset of the collected trajectories, using only the latest batch, or reusing all past data. If the current policy is already an empirical minimizer, it is retained: $\pi_t \in \arg\min_{\pi \in \Pi} Q_t(\pi; \mathcal{D}_t) \Longrightarrow \pi_{t+1} = \pi_t$. Thus a change of policy requires a strict decrease in the empirical objective. The output $\widehat{\pi}$ is the policy returned by the final step.

For student and teacher distributions p, q, the local loss $d_t(p, q)$ is finite, nonnegative, zero exactly when $p = q$, and convex in q. The function ϕ_t is nondecreasing, with $\phi_t(0) = 0$ and $\phi_t(u) > 0$ for $u > 0$. With $\phi_t(u) = u$, forward KL uses $d_t(p, q) = \mathrm{KL}(q \| p)$ and has the same minimizers as the teacher-weighted cross-entropy $-\sum_a q(a) \log p(a)$ used by Lin et al. (2020). Reverse KL uses $d_t(p, q) = \mathrm{KL}(p \| q)$, as in the local distillation objective of Agarwal et al. (2024). Skew reverse KL uses $d_t(p, q) = \mathrm{KL}(p \| \alpha_t p + (1 - \alpha_t) q)$ for $\alpha_t \in [0, 1)$ (Ko et al., 2024). These losses are convex in q by joint convexity of KL. The class also includes mixtures of the two KL directions and clipping with $\phi_t(u) = \min\{c_t, u\}$ for $c_t > 0$.

Proposition 3.1 (detailed statement). *For every sufficiently large λ, there is an $H = 2$, $|\Pi| = 2$ instance satisfying Assumptions 2.1 and 2.2 with $B = 4\lambda$, $B_\infty = 8\lambda$, and $L^\star \lesssim \lambda e^{-\lambda}$. Every update rule defined in* (14)*, initialized uniformly over Π and using $T = \lceil \lambda^3 \rceil$ episodes, returns $\widehat{\pi}$ satisfying*

$$\mathbb{P}\{L(\widehat{\pi}) \geq \lambda/4\} \geq \frac{1}{4}.$$

Proof of Proposition 3.1. MDP and policy class. The two-step MDP has

$$\mathcal{A} = \{a_1, a_2\}, \qquad \mathcal{S}_1 = \{s_{1,1}, s_{1,2}\}, \qquad \mathcal{S}_2 = \{s_{2,1}, s_{2,2}\}, \qquad \mu(s_{1,1}) = \mu(s_{1,2}) = \tfrac{1}{2}.$$

Its transitions are

$$\begin{aligned} &\mathsf{P}_1(s_{2,1} \mid s_{1,1}, a_1) = 1, \\ &\quad \mathsf{P}_1(s_{2,2} \mid s, a) = 1 \quad \big((s, a) \in (\mathcal{S}_1 \times \mathcal{A}) \setminus \{(s_{1,1}, a_1)\}\big). \end{aligned}$$

The episode ends after the second action. Set $\epsilon = e^{-\lambda}$, $\bar{p} = 1 - \epsilon^2$, and $v = e^{-4\lambda}$, and let $\Pi = \{\pi_{\mathrm{good}}, \pi_{\mathrm{bad}}\}$. The table gives the probability of a_1 at each state; the probability of a_2 is its complement.

Policy	$s_{1,1}$	$s_{1,2}$	$s_{2,1}$	$s_{2,2}$
$\pi^\star$	ϵ	ϵ	v	$1/2$
π_{good}	ϵ	ϵ^2	$\bar{p}$	$1/2$
π_{bad}	$\bar{p}$	ϵ	v	$1/2$

Write $\mathrm{Ber}(p) = (p, 1-p)$ on (a_1, a_2). All action probabilities are positive.

Boundedness and trajectory risks. The only candidate–state pairs at which the policies differ are listed below. For $\lambda \geq 1$,

π	s_h	$\max_a \frac{\pi_h(a \mid s_h)}{\pi_h^\star(a \mid s_h)}$	$k_h(s_h;\pi)$
π_{bad}	$s_{1,1}$	$\bar{p}/\epsilon \leq e^{\lambda}$	$\leq \lambda$
π_{good}	$s_{1,2}$	$(1-\epsilon^2)/(1-\epsilon) = 1+\epsilon$	$\leq \log(1+\epsilon) \leq \epsilon$
π_{good}	$s_{2,1}$	$\bar{p}/v \leq e^{4\lambda}$	$\leq 4\lambda$

A feasible trajectory cannot contain both $s_{1,2}$ and $s_{2,1}$. For each $\pi \in \Pi$ and $\tau \in \mathcal{T}_{\mathrm{env}}$,

$$\sum_{h=1}^{2} k_h(s_h;\pi) \leq 4\lambda$$

$$\log \frac{P^{\pi}(\tau)}{P^{\pi^\star}(\tau)} = \sum_{h=1}^{2} \log \frac{\pi_h(a_h \mid s_h)}{\pi_h^\star(a_h \mid s_h)} \leq 4\lambda \leq 8\lambda.$$

Thus Assumptions 2.1 and 2.2 hold with the stated budgets. The trajectory risks satisfy

$$\begin{aligned} L(\pi_{\mathrm{good}}) &= \tfrac{1}{2}\left[k_1(s_{1,2};\pi_{\mathrm{good}}) + \epsilon k_2(s_{2,1};\pi_{\mathrm{good}})\right] \leq \tfrac{1}{2}(1+4\lambda)\epsilon, \\ L(\pi_{\mathrm{bad}}) &= \tfrac{1}{2}\,\mathrm{KL}(\mathrm{Ber}(\bar{p}) \| \,\mathrm{Ber}(\epsilon)) \geq \tfrac{1}{2}(\bar{p}\lambda - \log 2) \geq \lambda/4. \end{aligned} \tag{15}$$

The first equality in each line is the KL chain rule; the second line uses binary entropy at most $\log 2$ and sufficiently large λ. Hence $L^\star = L(\pi_{\mathrm{good}}) \lesssim \lambda e^{-\lambda}$.

Empirical loss on each observed trajectory. For $\theta = (\bar{p}-\epsilon)/(\bar{p}-v) \in (0,1)$, $\mathrm{Ber}(\epsilon) = \theta\,\mathrm{Ber}(v) + (1-\theta)\,\mathrm{Ber}(\bar{p})$.

On a trajectory with $(s_1, s_2) = (s_{1,1}, s_{2,1})$,

$$\begin{aligned} &\sum_{h=1}^{2} d_t(\pi_{\mathrm{bad},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h)) \\ &\quad = d_t(\mathrm{Ber}(\bar{p}), \mathrm{Ber}(\epsilon)) && (d_t(\mathrm{Ber}(v), \mathrm{Ber}(v)) = 0) \\ &\quad \leq \theta d_t(\mathrm{Ber}(\bar{p}), \mathrm{Ber}(v)) + (1-\theta) d_t(\mathrm{Ber}(\bar{p}), \mathrm{Ber}(\bar{p})) && \text{(Convexity in the second parameter)} \\ &\quad = \theta d_t(\mathrm{Ber}(\bar{p}), \mathrm{Ber}(v)) < d_t(\mathrm{Ber}(\bar{p}), \mathrm{Ber}(v)) && (d_t(p,p) = 0,\ 0 < \theta < 1,\ v \neq \bar{p}) \\ &\quad = \sum_{h=1}^{2} d_t(\pi_{\mathrm{good},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h))\,. && (d_t(\mathrm{Ber}(\epsilon), \mathrm{Ber}(\epsilon)) = 0) \end{aligned} \tag{16}$$

On a trajectory with $(s_1, s_2) = (s_{1,2}, s_{2,2})$,

$$\begin{aligned} &\sum_{h=1}^{2} d_t(\pi_{\mathrm{bad},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h)) = 0 && (\pi_{\mathrm{bad}} = \pi^\star \text{ at both states}) \\ &\quad < d_t(\mathrm{Ber}(\epsilon^2), \mathrm{Ber}(\epsilon)) && (d_t(p,q) > 0 \text{ for } p \neq q,\ \epsilon^2 \neq \epsilon) \\ &\quad = \sum_{h=1}^{2} d_t(\pi_{\mathrm{good},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h))\,. && (d_t(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2)) = 0) \end{aligned} \tag{17}$$

Whenever every trajectory in $\mathcal{D}_t$ follows one of these two state sequences,

$$\begin{aligned} Q_t(\pi_{\mathrm{bad}};\mathcal{D}_t) &= \sum_{\tau \in \mathcal{D}_t} \omega_{t,\tau} \phi_t\left(\sum_{h=1}^{2} d_t(\pi_{\mathrm{bad},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h))\right) && \text{(Equation (14))} \\ &\leq \sum_{\tau \in \mathcal{D}_t} \omega_{t,\tau} \phi_t\left(\sum_{h=1}^{2} d_t(\pi_{\mathrm{good},h}(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h))\right) \\ &\qquad && \text{(Equations (16) and (17), } \omega_{t,\tau} \geq 0,\ \phi_t \text{ nondecreasing)} \\ &= Q_t(\pi_{\mathrm{good}};\mathcal{D}_t). && \text{(Equation (14))} \end{aligned} \tag{18}$$

Since the update keeps the current policy whenever it is an empirical minimizer, (18) gives

$$\pi_t = \pi_{\mathrm{bad}} \quad \Longrightarrow \quad \pi_{t+1} = \pi_{\mathrm{bad}},$$

including ties caused by clipping.

Persistence and failure probability. The only remaining state sequence is $s_{1,1} \to s_{2,2}$. Under π_{bad}, its probability is

$$P^{\pi_{\mathrm{bad}}}\{s_1 = s_{1,1}, s_2 = s_{2,2}\} = \tfrac{1}{2}\pi_{\mathrm{bad},1}(a_2 \mid s_{1,1}) = \frac{\epsilon^2}{2}.$$

Let $\mathcal{E}_m$ be the event that $\pi_1 = \pi_{\mathrm{bad}}$ and none of the first m episodes takes $s_{1,1} \to s_{2,2}$. On $\mathcal{E}_{m-1}$, every completed refit keeps π_{bad} by (18), so episode m also uses π_{bad}. Therefore,

$$\mathbb{P}(\mathcal{E}_0) = \tfrac{1}{2}, \qquad \mathbb{P}(\mathcal{E}_m \mid \mathcal{E}_{m-1}) = 1 - \epsilon^2/2 \quad (m \in [T]), \qquad \widehat{\pi} = \pi_{\mathrm{bad}} \text{ on } \mathcal{E}_T. \tag{19}$$

For arbitrary batch sizes and allowed trajectory weights,

$$\begin{aligned}
\mathbb{P}\{L(\widehat{\pi}) \geq \lambda/4\} &\geq \mathbb{P}(\mathcal{E}_T) && \text{(Equations (15) and (19))} \\
&= \frac{1}{2}\left(1 - \frac{\epsilon^2}{2}\right)^T \geq \frac{1}{2}\left(1 - \frac{T\epsilon^2}{2}\right) && \text{(Bernoulli's inequality)} \\
&= \frac{1}{2}\left(1 - \frac{\lceil \lambda^3 \rceil e^{-2\lambda}}{2}\right) \geq \frac{1}{4}. && \text{(Sufficiently large } \lambda\text{)}
\end{aligned}$$

□

B.2 Online reverse-KL learning

B.2.1 Clipping and prefix truncation

Lemma B.1 (Clipped objectives and truncation)**.** *If $L(\pi) < \infty$ for every candidate, then*

$$\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] \asymp \overline{L}(\pi), \qquad D^2_{\mathrm{H}}(P^{\pi}, P^{\pi^\star}) \leq 6\overline{L}(\pi). \tag{20}$$

For every $\pi, \pi' \in \Pi$, $0 \leq \mathrm{Trun}_{\pi'}[\overline{L}(\pi)] \leq \overline{L}(\pi)$ and

$$\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] \geq \frac{1}{16}\overline{L}(\pi) - D^2_{\mathrm{H}}(P^{\pi^\star}, P^{\pi'}). \tag{21}$$

Consequently,

$$\overline{L}(\pi) \leq 64\left(\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] + \overline{L}(\pi')\right).$$

To prove Lemma B.1, we first introduce several auxilary lemmas to control the behaviour of the clipped loss.

Lemma B.2 (Change of measure for a bounded loss)**.** *For probability laws R, S and a measurable $f \in [0,1]$, $\mathrm{KL}(R\|S) \leq \mathbb{E}_R[f]$ implies $\mathbb{E}_S[f] \asymp \mathbb{E}_R[f]$.*

Proof. Lemma D.1, applied to $g = \lambda f$, gives for every $\lambda \in \mathbb{R}$,

$$\mathrm{KL}(R\|S) \geq \lambda\mathbb{E}_R[f] - \log \mathbb{E}_S\left[e^{\lambda f}\right] \geq \lambda\mathbb{E}_R[f] - (e^\lambda - 1)\mathbb{E}_S[f].$$

The second inequality uses convexity of the exponential on $[0,1]$. The choices $\lambda = 2$ and $\lambda = -1$, respectively, give $(e^2 - 1)\mathbb{E}_S[f] \geq \mathbb{E}_R[f]$ and $(1 - e^{-1})\mathbb{E}_S[f] \leq 2\mathbb{E}_R[f]$. □

Each term in $\mathrm{Trun}_{\pi'}[\overline{L}(\pi)]$ can be written as $\min\{e_h(\xi_h;\pi)P^{\pi}(\xi_h), e_h(\xi_h;\pi)P^{\pi'}(\xi_h)\}$. For a fixed candidate π and any trajectory distribution P, we therefore define an auxiliary distribution $\mathcal{R}_\pi[P]$ that assigns probability $e_h(\xi_h;\pi)P(\xi_h)$ to each prefix outcome (h, ξ_h), $h \in [H]$. The distribution also includes complete-trajectory outcomes (∞, τ) for the remaining probability, and is defined by

$$\begin{aligned}
\mathcal{R}_\pi[P](h, \xi_h) &= e_h(\xi_h;\pi)P(\xi_h), \\
\mathcal{R}_\pi[P](\infty, \tau) &= \left(1 - \sum_{h=1}^{H} e_h(\xi_h;\pi)\right)P(\tau).
\end{aligned} \tag{22}$$

The label ∞ distinguishes a complete trajectory from a prefix. The difference between τ and the last prefix ξ_H is that ξ_H do not have a_H but τ has. Given $\tau \sim P$, these probabilities correspond to retaining its prefix (h, ξ_h) with probability $e_h(\xi_h; \pi)$ and retaining (∞, τ) with the remaining probability. Since $\sum_{h=1}^{H} e_h(\xi_h; \pi) = \min\{1, \sum_{h=1}^{H} k_h(s_h; \pi)\}$, all probabilities are nonnegative and sum to one.

Lemma B.3 (KL between auxiliary distributions)**.** *For $L(\pi) < \infty$,*

$$\mathrm{KL}(\mathcal{R}_\pi[P^\pi] \| \mathcal{R}_\pi[P^{\pi^\star}]) = \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H}\left(1 - \sum_{v=1}^{h} k_v(s_v; \pi)\right)_+ k_h(s_h; \pi)\right] \le \overline{L}(\pi). \tag{23}$$

Proof. Fix $\pi \in \Pi$. For each $\widetilde{\pi} \in \Pi \cup \{\pi^\star\}$, define the joint law $\mathbb{P}_{\widetilde{\pi}}$ by $\tau \sim P^{\widetilde{\pi}}$ and $\sigma \in [H] \cup \{\infty\}$ with conditional distribution

$$\mathbb{P}_{\widetilde{\pi}}(\sigma = h \mid \tau) = e_h(\xi_h; \pi), \qquad h \in [H],$$
$$\mathbb{P}_{\widetilde{\pi}}(\sigma = \infty \mid \tau) = 1 - \sum_{h=1}^{H} e_h(\xi_h; \pi).$$

By (2), $e_h(\xi_h; \pi) \ge 0$ and $\sum_{h=1}^{H} e_h(\xi_h; \pi) \le 1$, so these probabilities sum to one. For $h \in [H]$,

$$\mathbb{P}_{\widetilde{\pi}}(\sigma \ge h \mid \tau) = 1 - \sum_{v=1}^{h-1} e_v(\xi_v; \pi), \tag{24}$$

$$\mathbb{P}_{\widetilde{\pi}}(\sigma > h \mid \tau) = 1 - \sum_{v=1}^{h} e_v(\xi_v; \pi) = \left(1 - \sum_{v=1}^{h} k_v(s_v; \pi)\right)_+. \tag{25}$$

where the second equation uses $1 - \min\{1, x\} = (1 - x)_+$. Consequently, whenever $\sum_{v=1}^{h-1} e_v(\xi_v; \pi) < 1$,

$$\mathbb{P}_{\widetilde{\pi}}(\sigma = h \mid \sigma \ge h, \tau) = \frac{e_h(\xi_h; \pi)}{1 - \sum_{v=1}^{h-1} e_v(\xi_v; \pi)} = \mathbb{P}_{\widetilde{\pi}}(\sigma = h \mid \sigma \ge h, \xi_h).$$

The last equality holds because the ratio depends only on ξ_h.

Define the retained record by

$$Y = \begin{cases} (\sigma, \xi_\sigma), & \sigma \in [H], \\ (\infty, \tau), & \sigma = \infty. \end{cases}$$

Since $e_h(\xi_h; \pi)$ depends only on ξ_h,

$$\begin{aligned} \mathbb{P}_{\widetilde{\pi}}(Y = (h, \xi_h)) &= P^{\widetilde{\pi}}(\xi_h)\mathbb{P}_{\widetilde{\pi}}(\sigma = h \mid \xi_h) \\ &= P^{\widetilde{\pi}}(\xi_h) e_h(\xi_h; \pi) = \mathcal{R}_\pi[P^{\widetilde{\pi}}](h, \xi_h), \\ \mathbb{P}_{\widetilde{\pi}}(Y = (\infty, \tau)) &= P^{\widetilde{\pi}}(\tau)\left(1 - \sum_{h=1}^{H} e_h(\xi_h; \pi)\right) = \mathcal{R}_\pi[P^{\widetilde{\pi}}](\infty, \tau). \end{aligned}$$

Thus $Y \sim \mathcal{R}_\pi[P^{\widetilde{\pi}}]$. From Equation 22, for any outcome with $\mathcal{R}_\pi[P^\pi](Y) > 0$, the common stopping factor cancels from the likelihood ratio. When $\sigma \in [H]$,

$$\frac{\mathcal{R}_\pi[P^\pi](Y)}{\mathcal{R}_\pi[P^{\pi^\star}](Y)} = \frac{e_\sigma(\xi_\sigma; \pi) P^\pi(\xi_\sigma)}{e_\sigma(\xi_\sigma; \pi) P^{\pi^\star}(\xi_\sigma)} = \frac{P^\pi(\xi_\sigma)}{P^{\pi^\star}(\xi_\sigma)}.$$

When $\sigma = \infty$, the common factor $1 - \sum_{h=1}^{H} e_h(\xi_h; \pi)$ cancels in the same way, leaving $P^\pi(\tau)/P^{\pi^\star}(\tau)$. The two policies share the initial-state distribution and environment transitions, so

$$\log \frac{\mathcal{R}_\pi[P^\pi](Y)}{\mathcal{R}_\pi[P^{\pi^\star}](Y)} = \sum_{h=1}^{H} \mathbf{1}\{\sigma > h\} \log \frac{\pi_h(a_h \mid s_h)}{\pi_h^\star(a_h \mid s_h)}.$$

Thus, we have

$$\begin{aligned}
&\mathrm{KL}(\mathcal{R}_\pi[P^\pi]\|\mathcal{R}_\pi[P^{\pi^\star}]) \\
&= \mathbb{E}_{\mathbb{P}_\pi}\left[\sum_{h=1}^{H} \mathbf{1}\{\sigma > h\} \log \frac{\pi_h(a_h \mid s_h)}{\pi_h^\star(a_h \mid s_h)}\right] \\
&= \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H} \mathbb{P}_\pi(\sigma > h \mid \tau) \log \frac{\pi_h(a_h \mid s_h)}{\pi_h^\star(a_h \mid s_h)}\right] \\
&= \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H} \left(1 - \sum_{v=1}^{h} k_v(s_v;\pi)\right)_+ \log \frac{\pi_h(a_h \mid s_h)}{\pi_h^\star(a_h \mid s_h)}\right] \qquad \text{(Equation 25)} \\
&= \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H} \left(1 - \sum_{v=1}^{h} k_v(s_v;\pi)\right)_+ k_h(s_h;\pi)\right].
\end{aligned}$$

By (2), $e_h(\xi_h;\pi) = k_h(s_h;\pi)$ when $\sum_{v=1}^{h} k_v(s_v;\pi) < 1$, otherwise, $(1 - \sum_{v=1}^{h} k_v(s_v;\pi))_+ = 0$. Thus,

$$\begin{aligned}
\left(1 - \sum_{v=1}^{h} k_v(s_v;\pi)\right)_+ k_h(s_h;\pi) &\le \mathbf{1}\left\{\sum_{v=1}^{h} k_v(s_v;\pi) < 1\right\} k_h(s_h;\pi) \\
&= \mathbf{1}\left\{\sum_{v=1}^{h} k_v(s_v;\pi) < 1\right\} e_h(\xi_h;\pi) \le e_h(\xi_h;\pi).
\end{aligned}$$

Hence

$$\mathrm{KL}(\mathcal{R}_\pi[P^\pi]\|\mathcal{R}_\pi[P^{\pi^\star}]) \le \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] = \overline{L}(\pi).$$

The construction also permits infinite local KL under a source $\widetilde{\pi}$, because $k_h(s_h;\pi) = \infty, \sum_{v=1}^{h} e_v(\xi_v;\pi) = 1, \mathbb{P}_{\widetilde{\pi}}(\sigma > h \mid \tau) = 0$.. □

Lemma B.4 (Shared mass on an event)**.** *Let R, S be probability distributions on a countable space $\mathcal{X}$ with $\mathrm{KL}(R\|S) < \infty$, and let $D \subseteq \mathcal{X}$. Write $z = R(D)$ and $v = \sum_{x\in D} \min\{R(x), S(x)\}$. For $z > 0$,*

$$\mathrm{KL}(R\|S) \ge z \log(z/v) - z. \tag{26}$$

Consequently, $\mathrm{KL}(R\|S) \le z$ implies $v \gtrsim z$.

Proof. For $z = 0$, $v = 0$. For $z > 0$, $0 < v \le z$. Define

$$g(x) = \begin{cases} \log \dfrac{R(x)}{\min\{R(x), S(x)\}}, & x \in D,\ R(x) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}
\mathbb{E}_S\left[e^g\right] &= 1 + \sum_{\substack{x\in D \\ R(x)>S(x)}} (R(x) - S(x)) \\
&= 1 + \sum_{x\in D} \big(R(x) - \min\{R(x), S(x)\}\big) = 1 + z - v, \\
\mathbb{E}_R\left[g\right] &= -z \sum_{\substack{x\in D \\ R(x)>0}} \frac{R(x)}{z} \log \frac{\min\{R(x), S(x)\}}{R(x)} \\
&\ge -z \log\left(\sum_{\substack{x\in D \\ R(x)>0}} \frac{R(x)}{z} \frac{\min\{R(x), S(x)\}}{R(x)}\right) = z \log(z/v).
\end{aligned}$$

The inequality follows from Jensen's inequality. Since $\mathbb{E}_S[e^g] = 1 + z - v \in [1, 2]$, Lemma D.1 gives

$$\begin{aligned}\mathrm{KL}(R\|S) &\geq \mathbb{E}_R[g] - \log \mathbb{E}_S[e^g] \\ &\geq z\log(z/v) - \log(1+z-v) \\ &\geq z\log(z/v) - (z-v) \geq z\log(z/v) - z.\end{aligned}$$

Consequently, $\mathrm{KL}(R\|S) \leq z \implies z\log(z/v) - z \leq z \implies v \geq e^{-2}z$. □

Proof of Lemma B.1. Fix $\pi \in \Pi$, and let $D = \{(h, \xi_h) : h \in [H]\}$ be the set of prefix outcomes in (22). For every $\widetilde{\pi} \in \Pi \cup \{\pi^\star\}$, Equation (22) gives

$$\mathcal{R}_\pi[P^{\widetilde{\pi}}](D) = \sum_{h=1}^{H}\sum_{\xi_h} e_h(\xi_h;\pi)P^{\widetilde{\pi}}(\xi_h) = \mathbb{E}_{P^{\widetilde{\pi}}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right].$$

Lemma B.3 therefore implies

$$\mathrm{KL}(\mathcal{R}_\pi[P^\pi]\|\mathcal{R}_\pi[P^{\pi^\star}]) \leq \overline{L}(\pi) = \mathcal{R}_\pi[P^\pi](D).$$

Applying Lemma B.2 to $\mathcal{R}_\pi[P^\pi]$ and $\mathcal{R}_\pi[P^{\pi^\star}]$ and $f(y) = \mathbf{1}\{y \in D\}$ as the indicator of D yields

$$\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] = \mathcal{R}_\pi[P^{\pi^\star}](D) \asymp \mathcal{R}_\pi[P^\pi](D) = \overline{L}(\pi). \tag{27}$$

Now we control the trajectory Hellinger distance.

$$\begin{aligned}D_{\mathrm{H}}^2(P^\pi, P^{\pi^\star}) &= \sum_\tau \left(1 - \sum_{h=1}^{H} e_h(\xi_h;\pi)\right)\left(\sqrt{P^\pi(\tau)} - \sqrt{P^{\pi^\star}(\tau)}\right)^2 + \sum_\tau \left(\sum_{h=1}^{H} e_h(\xi_h;\pi)\right)\left(\sqrt{P^\pi(\tau)} - \sqrt{P^{\pi^\star}(\tau)}\right)^2 \\ &= \sum_\tau \left(\sqrt{\mathcal{R}_\pi[P^\pi](\infty,\tau)} - \sqrt{\mathcal{R}_\pi[P^{\pi^\star}](\infty,\tau)}\right)^2 + \sum_\tau \left(\sum_{h=1}^{H} e_h(\xi_h;\pi)\right)\left(\sqrt{P^\pi(\tau)} - \sqrt{P^{\pi^\star}(\tau)}\right)^2 \\ &\qquad\qquad \text{(Equation (22))} \\ &\leq D_{\mathrm{H}}^2(\mathcal{R}_\pi[P^\pi], \mathcal{R}_\pi[P^{\pi^\star}]) + \overline{L}(\pi) + \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] \\ &\qquad\qquad ((\sqrt{p} - \sqrt{q})^2 \leq p + q) \\ &\leq \mathrm{KL}(\mathcal{R}_\pi[P^\pi]\|\mathcal{R}_\pi[P^{\pi^\star}]) + \overline{L}(\pi) + \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] \\ &\leq 2\overline{L}(\pi) + \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] \leq 6\overline{L}(\pi).\end{aligned} \tag{28}$$

The last line uses Lemma B.3 and the explicit bound $\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] \leq \frac{2}{1-e^{-1}}\overline{L}(\pi) \leq 4\overline{L}(\pi)$ from the proof of Lemma B.2.

For the same event $D = \{(h, \xi_h) : h \in [H]\}$, Equation (22) gives

$$\begin{aligned}\sum_{z\in D} \min\{\mathcal{R}_\pi[P^\pi](z), \mathcal{R}_\pi[P^{\pi'}](z)\} &= \sum_{h=1}^{H}\sum_{\xi_h} e_h(\xi_h;\pi)\min\{P^\pi(\xi_h), P^{\pi'}(\xi_h)\} \\ &= \mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] \in [0, \overline{L}(\pi)].\end{aligned}$$

The interval bound uses $\min\{P^\pi(\xi_h), P^{\pi'}(\xi_h)\} \leq P^\pi(\xi_h)$. By (23), Lemma B.4 applied to D gives

$$\sum_{z\in D} \min\{\mathcal{R}_\pi[P^\pi](z), \mathcal{R}_\pi[P^{\pi^\star}](z)\} \geq e^{-2}\overline{L}(\pi).$$

Equation (64) and (66) in Lemma D.2 give

$$\begin{aligned} e^{-1}\sqrt{\overline{L}(\pi)} &\le \sqrt{\sum_{z\in D}\min\{\mathcal{R}_\pi[P^\pi](z), \mathcal{R}_\pi[P^{\pi^\star}](z)\}} \\ &\le \sqrt{\sum_{z\in D}\min\{\mathcal{R}_\pi[P^\pi](z), \mathcal{R}_\pi[P^{\pi'}](z)\}} + D_{\mathrm{H}}(\mathcal{R}_\pi[P^{\pi^\star}], \mathcal{R}_\pi[P^{\pi'}]) \\ &\le \sqrt{\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right]} + D_{\mathrm{H}}(P^{\pi^\star}, P^{\pi'}). \end{aligned}$$

Squaring and using $(x+y)^2 \le 2x^2 + 2y^2$ and $e^{-2} > 1/8$ gives

$$\begin{aligned} e^{-2}\overline{L}(\pi) &\le 2\,\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] + 2D_{\mathrm{H}}^2(P^{\pi^\star}, P^{\pi'}), \\ \mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] &\ge \frac{1}{16}\overline{L}(\pi) - D_{\mathrm{H}}^2(P^{\pi^\star}, P^{\pi'}), \\ \overline{L}(\pi) &\le e^2\left(\sqrt{\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right]} + \sqrt{6\overline{L}(\pi')}\right)^2 \\ &\le 7e^2\left(\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] + \overline{L}(\pi')\right) \\ &\le 64\left(\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] + \overline{L}(\pi')\right). \end{aligned}$$

The last three inequalities use (28), Cauchy–Schwarz, and $7e^2 < 64$, respectively. □

B.2.2 Exponential-weights analysis

Throughout this subsection, $\mathbb{E}_t$ conditions on the history before episode t and averages over both $\pi_t \sim w_t$ and its trajectory. For uniform initialization, $\ell_t(\pi) \in [-1, 1]$, and $0 < \eta \le \log 2$, standard exponential-weights analysis (Lattimore & Szepesvári, 2020, Eqs. (27.2)–(27.3)) gives

$$-\sum_{t=1}^{T}\ell_t(\pi^\dagger) \le \frac{\log|\Pi|}{\eta} - \sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi) + \eta\sum_{t=1}^{T}\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2 \tag{29}$$

for every fixed $\pi^\dagger \in \Pi$ and every realized loss sequence. The following lemma relates the loss terms to $\overline{L}$.

Lemma B.5 (Conditional loss moments)**.** *For Algorithm 1, $|\ell_t(\pi)| \le 1$ and*

$$\mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)\right] = 0. \tag{30}$$

With $a = 1/16$ and a numerical $C > 0$,

$$\begin{aligned} -\mathbb{E}_t\left[\ell_t(\pi)\right] &\ge a\sum_{\pi'\in\Pi} w_t(\pi')\overline{L}(\pi') - 2L(\pi), \\ \mathbb{E}_t\left[\ell_t(\pi)\right] &\ge a\overline{L}(\pi) - C\sum_{\pi'\in\Pi} w_t(\pi')\overline{L}(\pi'). \end{aligned} \tag{31}$$

The second moments satisfy

$$\begin{aligned} \mathbb{E}_t\left[\ell_t(\pi)^2\right] &\le \overline{L}(\pi) + \sum_{\pi'\in\Pi} w_t(\pi')\overline{L}(\pi'), \\ \mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2\right] &\le 2\sum_{\pi'\in\Pi} w_t(\pi')\overline{L}(\pi'). \end{aligned} \tag{32}$$

Proof. The definition of ℓ_t gives the conditional mean and cancellation

$$\begin{aligned} \mathbb{E}_t\left[\ell_t(\pi) \mid \pi_t = \pi'\right] &= \mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] - \mathrm{Trun}_{\pi}\left[\overline{L}(\pi')\right], \\ \mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)\right] &= \sum_{\pi,\pi'\in\Pi} w_t(\pi)w_t(\pi')\Big(\mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] - \mathrm{Trun}_{\pi}\left[\overline{L}(\pi')\right]\Big) = 0. \end{aligned} \tag{33}$$

The truncation bound in Lemma B.1, together with $\overline{L}(\pi) \le L(\pi)$ and $D_{\mathrm{H}}^2(P^{\pi}, P^{\pi^\star}) \le L(\pi)$, yields

$$-\mathbb{E}_t\left[\ell_t(\pi) \mid \pi_t = \pi'\right] \ge a\overline{L}(\pi') - D_{\mathrm{H}}^2(P^{\pi}, P^{\pi^\star}) - \overline{L}(\pi) \ge a\overline{L}(\pi') - 2L(\pi),$$
$$\mathbb{E}_t\left[\ell_t(\pi) \mid \pi_t = \pi'\right] \ge a\overline{L}(\pi) - D_{\mathrm{H}}^2(P^{\pi'}, P^{\pi^\star}) - \overline{L}(\pi') \ge a\overline{L}(\pi) - C\overline{L}(\pi').$$

Here the last inequality uses $D_{\mathrm{H}}^2(P^{\pi'}, P^{\pi^\star}) \lesssim \overline{L}(\pi')$ from the same lemma. Averaging over $\pi' \sim w_t$ gives the two risk bounds.

Each weighted sum in (6) lies in $[0,1]$, so $|\ell_t(\pi)| \le 1$. The inequality $(x-y)^2 \le x + y$ for $x, y \in [0,1]$ gives the remaining bounds in succession,

$$\mathbb{E}_t\left[\ell_t(\pi)^2 \mid \pi_t = \pi'\right] \le \mathrm{Trun}_{\pi'}\left[\overline{L}(\pi)\right] + \mathrm{Trun}_{\pi}\left[\overline{L}(\pi')\right] \le \overline{L}(\pi) + \overline{L}(\pi'),$$
$$\mathbb{E}_t\left[\ell_t(\pi)^2\right] \le \overline{L}(\pi) + \sum_{\pi' \in \Pi} w_t(\pi')\overline{L}(\pi'),$$
$$\mathbb{E}_t\left[\sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)^2\right] \le \sum_{\pi, \pi' \in \Pi} w_t(\pi)w_t(\pi')\left(\overline{L}(\pi) + \overline{L}(\pi')\right) = 2\sum_{\pi \in \Pi} w_t(\pi)\overline{L}(\pi).$$

□

Proof of Theorem 3.2. The case $|\Pi| = 1$ is immediate. For $|\Pi| \ge 2$, fix $\pi^\dagger \in \arg\min_{\pi \in \Pi} L(\pi)$ and write $S_T = \sum_{t=1}^T \sum_{\pi \in \Pi} w_t(\pi)\overline{L}(\pi)$. The cumulative weighted clipped risk S_T is random because w_t depends on the training history.

Cumulative training risk. The first inequality in (31), summed for $\pi^\dagger$, bounds aS_T by $2TL^\star - \sum_{t=1}^T \mathbb{E}_t\left[\ell_t(\pi^\dagger)\right]$. Separating the observed losses from their conditional expectations and applying (29) gives

$$\begin{aligned} aS_T &\le 2TL^\star - \sum_{t=1}^T \mathbb{E}_t\left[\ell_t(\pi^\dagger)\right] \\ &= 2TL^\star - \sum_{t=1}^T \ell_t(\pi^\dagger) + \sum_{t=1}^T \Big(\ell_t(\pi^\dagger) - \mathbb{E}_t\left[\ell_t(\pi^\dagger)\right]\Big) \\ &\le 2TL^\star + \frac{\log|\Pi|}{\eta} + \underbrace{\left(-\sum_{t=1}^T \sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)\right)}_{\text{weighted loss}} + \underbrace{\eta \sum_{t=1}^T \sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)^2}_{\text{quadratic term}} \\ &\quad + \underbrace{\sum_{t=1}^T \Big(\ell_t(\pi^\dagger) - \mathbb{E}_t\left[\ell_t(\pi^\dagger)\right]\Big)}_{\text{comparator sampling error}}. \end{aligned} \tag{34}$$

For any fixed $0 < \lambda \le 1/2$, the following bounds hold simultaneously with probability at least $1 - \delta/2$, by Lemma D.4 and a union bound.

For the weighted loss,

$$\begin{aligned} &\left|\sum_{t=1}^T \sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)\right| \\ &\le \lambda \sum_{t=1}^T \mathbb{E}_t\left[\left(\sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)\right)^2\right] + \frac{\log(8/\delta)}{\lambda} && \text{(Equations (30) and (67))} \\ &\le \lambda \sum_{t=1}^T \mathbb{E}_t\left[\sum_{\pi \in \Pi} w_t(\pi)\ell_t(\pi)^2\right] + \frac{\log(8/\delta)}{\lambda} && \text{(Jensen's inequality)} \\ &\le 2\lambda S_T + \frac{\log(8/\delta)}{\lambda}. && \text{(Equation (32))} \end{aligned} \tag{35}$$

For the quadratic term, Lemma B.5 gives $0 \le \sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2 \le 1$, so

$$\begin{aligned}
&\eta \sum_{t=1}^{T} \sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2 \\
&\quad\le \eta \sum_{t=1}^{T} \mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2\right] + \eta\frac{\log(8/\delta)}{\lambda} && \text{(Equation (67))} \\
&\qquad + \eta\lambda \sum_{t=1}^{T} \mathbb{E}_t\left[\left(\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2\right)^2\right] \\
&\quad\le \eta(1+\lambda) \sum_{t=1}^{T} \mathbb{E}_t\left[\sum_{\pi\in\Pi} w_t(\pi)\ell_t(\pi)^2\right] + \eta\frac{\log(8/\delta)}{\lambda} && (x^2 \le x \text{ for } x\in[0,1]) \\
&\quad\le 2\eta(1+\lambda)S_T + \eta\frac{\log(8/\delta)}{\lambda}. && \text{(Equation (32))} \qquad (36)
\end{aligned}$$

For the comparator sampling error,

$$\begin{aligned}
&\sum_{t=1}^{T}\Big(\ell_t(\pi^\dagger) - \mathbb{E}_t\left[\ell_t(\pi^\dagger)\right]\Big) \\
&\quad\le \lambda \sum_{t=1}^{T} \mathbb{E}_t\left[\ell_t(\pi^\dagger)^2\right] + \frac{\log(8/\delta)}{\lambda} && \text{(Equation (67))} \\
&\quad\le \lambda\big(T\overline{L}(\pi^\dagger) + S_T\big) + \frac{\log(8/\delta)}{\lambda} && \text{(Equation (32))} \\
&\quad\le \lambda(TL^\star + S_T) + \frac{\log(8/\delta)}{\lambda}. && (\overline{L}(\pi^\dagger) \le L^\star) \qquad (37)
\end{aligned}$$

Substituting (35), (36), and (37) for the weighted loss, quadratic term, and comparator sampling error in (34), respectively, gives

$$\begin{aligned}
aS_T &\le 2TL^\star + \frac{\log|\Pi|}{\eta} + 2\lambda S_T + \frac{\log(8/\delta)}{\lambda} \\
&\qquad + 2\eta(1+\lambda)S_T + \eta\frac{\log(8/\delta)}{\lambda} + \lambda(TL^\star + S_T) + \frac{\log(8/\delta)}{\lambda}, \qquad (38) \\
\big[a - 3\lambda - 2\eta(1+\lambda)\big]S_T &\le (2+\lambda)TL^\star + \frac{\log|\Pi|}{\eta} + (2+\eta)\frac{\log(8/\delta)}{\lambda}.
\end{aligned}$$

For $\eta = 1/256$, $a - 2\eta = 7/128 > 0$. A sufficiently small fixed $\lambda > 0$ therefore makes the coefficient on S_T positive, yielding $S_T \lesssim TL^\star + \log(|\Pi|/\delta)$.

Selection by final weight. The maximum-weight output satisfies $w_{T+1}(\widehat{\pi}) \ge 1/|\Pi| = w_1(\widehat{\pi})$. For every candidate with $w_{T+1}(\pi) \ge w_1(\pi)$, the weight update and Jensen's inequality bound its cumulative observed loss by the weighted loss,

$$\begin{aligned}
\sum_{t=1}^{T}\ell_t(\pi) &= -\frac{1}{\eta}\log\frac{w_{T+1}(\pi)}{w_1(\pi)} - \frac{1}{\eta}\sum_{t=1}^{T}\log\left(\sum_{\pi'\in\Pi} w_t(\pi')e^{-\eta\ell_t(\pi')}\right) \\
&\le -\frac{1}{\eta}\log\frac{w_{T+1}(\pi)}{w_1(\pi)} + \sum_{t=1}^{T}\sum_{\pi'\in\Pi} w_t(\pi')\ell_t(\pi') \qquad (39) \\
&\le \sum_{t=1}^{T}\sum_{\pi'\in\Pi} w_t(\pi')\ell_t(\pi').
\end{aligned}$$

The last inequality uses $w_{T+1}(\pi) \ge w_1(\pi)$. To convert this observed-loss bound to a risk bound, we apply (68) with failure probability $\delta/(2|\Pi|)$ to each candidate and take a union bound over Π. The

uniform bound over Π and the preceding concentration bounds hold jointly with probability at least $1-\delta$. On this event, every candidate with $w_{T+1}(\pi) \geq 1/|\Pi|$ satisfies

$$
\begin{aligned}
aT\overline{L}(\pi) &\leq \sum_{t=1}^{T} \mathbb{E}_t\left[\ell_t(\pi)\right] + CS_T && \text{(Equation (31))} \\
&\leq \sum_{t=1}^{T} \ell_t(\pi) + \lambda \sum_{t=1}^{T} \mathbb{E}_t\left[\ell_t(\pi)^2\right] + CS_T + \frac{\log(2|\Pi|/\delta)}{\lambda} && \text{(Equation (68))} \\
&\leq \sum_{t=1}^{T} \ell_t(\pi) + \lambda T\overline{L}(\pi) + (C+\lambda)S_T + \frac{\log(2|\Pi|/\delta)}{\lambda} && \text{(Equation (32))} \\
&\leq \sum_{t=1}^{T} \sum_{\pi' \in \Pi} w_t(\pi')\ell_t(\pi') + \lambda T\overline{L}(\pi) + (C+\lambda)S_T + \frac{\log(2|\Pi|/\delta)}{\lambda} && \text{(Equation (39))} \\
&\leq \lambda T\overline{L}(\pi) + (C+3\lambda)S_T + \frac{\log(2|\Pi|/\delta) + \log(8/\delta)}{\lambda}. && \text{(Equation (35))}
\end{aligned}
$$

Rearranging with $\lambda < a$ and using (38) gives, simultaneously for every π with $w_{T+1}(\pi) \geq 1/|\Pi|$,

$$
\begin{aligned}
\overline{L}(\pi) &\leq \frac{(C+3\lambda)S_T + \lambda^{-1}\big(\log(2|\Pi|/\delta) + \log(8/\delta)\big)}{(a-\lambda)T} \\
&\lesssim L^\star + \frac{\log(|\Pi|/\delta)}{T}.
\end{aligned}
\tag{40}
$$

In particular, $\widehat{\pi}$ satisfies this bound. The clipped-risk conversions in Lemmas B.1 and D.3 then yield

$$
\begin{aligned}
L(\widehat{\pi}) \leq B\overline{L}(\widehat{\pi}) &\lesssim B\left(L^\star + \frac{\log(|\Pi|/\delta)}{T}\right) && \text{under Assumption 2.1,} \\
L(\widehat{\pi}) \lesssim B_\infty\overline{L}(\widehat{\pi}) &\lesssim B_\infty\left(L^\star + \frac{\log(|\Pi|/\delta)}{T}\right) && \text{under Assumption 2.2.}
\end{aligned}
$$

□

B.3 Online Hellinger learning from full feedback

Algorithm 2 replaces the local-KL score in Algorithm 1 by a prefix-weighted local Hellinger score before clipping.

Algorithm 2 Online Hellinger learning with full teacher feedback

Require: Finite candidate class Π, episode budget T.

1: Set $w_1(\pi) = 1/|\Pi|$ for every $\pi \in \Pi$ and $\eta = 1/128$.

2: **for** $t = 1, \dots, T$ **do**

3: Sample a policy $\pi_t \sim w_t$ and generate a complete trajectory $\tau_t = (s_{t,1}, a_{t,1}, \dots, s_{t,H}, a_{t,H}) \sim P^{\pi_t}$ by executing π_t for the entire episode.

4: Observe $\pi_h^\star(\cdot \mid s_{t,h})$ for every $h = 1, \dots, H$. Write $\xi_{t,h} = (s_{t,1}, a_{t,1}, \dots, s_{t,h})$.

5: **for** each $\pi \in \Pi$ **do**

6: For every $h = 1, \dots, H$ with $P^\pi(\xi_{t,h}) > 0$, compute

$$g_h(\xi_{t,h}; \pi) = \sqrt{\prod_{v=1}^{h-1} \frac{\pi_v^\star(a_{t,v} \mid s_{t,v})}{\pi_v(a_{t,v} \mid s_{t,v})}} D^2_{\mathrm{H}}(\pi_h(\cdot \mid s_{t,h}), \pi_h^\star(\cdot \mid s_{t,h}))\,.$$

Set $g_h(\xi_{t,h}; \pi) = 0$ when $P^\pi(\xi_{t,h}) = 0$. The empty product equals one.

7: For every $h = 1, \dots, H$, compute the clipped increment

$$e_h^{\mathrm{H}}(\xi_{t,h}; \pi) = \min\left\{1, \sum_{v=1}^{h} g_v(\xi_{t,v}; \pi)\right\} - \min\left\{1, \sum_{v=1}^{h-1} g_v(\xi_{t,v}; \pi)\right\}.$$

8: **end for**

9: For every $\pi \in \Pi$, compute

$$\ell_t(\pi) = \sum_{h=1}^{H} \min\left\{1, \prod_{v=1}^{h-1} \frac{\pi_v(a_{t,v} \mid s_{t,v})}{\pi_{t,v}(a_{t,v} \mid s_{t,v})}\right\} \left(e_h^{\mathrm{H}}(\xi_{t,h}; \pi) - e_h^{\mathrm{H}}(\xi_{t,h}; \pi_t)\right),$$

where $\pi_{t,v}$ is the action distribution of the sampled policy π_t at step v.

10: Update all candidate weights by $w_{t+1}(\pi) = \dfrac{w_t(\pi)\exp(-\eta \ell_t(\pi))}{\sum_{\pi' \in \Pi} w_t(\pi')\exp(-\eta \ell_t(\pi'))}$.

11: **end for**

12: Return $\widehat{\pi} \in \arg\max_{\pi \in \Pi} w_{T+1}(\pi)$.

Write $L_{\mathrm{H}}(\pi) = D^2_{\mathrm{H}}(P^\pi, P^{\pi^\star})$ and $\overline{L}_{\mathrm{H}}(\pi) = \mathbb{E}_{P^\pi}\left[\sum_{h=1}^{H} e_h^{\mathrm{H}}(\xi_h; \pi)\right]$.

Theorem B.6 (Online Hellinger guarantee)**.** *For a finite class Π, Algorithm 2 returns $\widehat{\pi} \in \Pi$ such that, with probability at least $1 - \delta$,*

$$L_{\mathrm{H}}(\widehat{\pi}) \lesssim \min_{\pi \in \Pi} L_{\mathrm{H}}(\pi) + \frac{\log(|\Pi|/\delta)}{T}. \tag{41}$$

Clipping and prefix truncation. For a fixed candidate π, write $P = P^\pi$, $Q = P^{\pi^\star}$, and $G_h = \sum_{v=1}^{h} g_v(\xi_v; \pi)$, with $G_0 = 0$. For $c > 0$, define

$$e_h^{(c)}(\xi_h) = \min\{c, G_h\} - \min\{c, G_{h-1}\}, \qquad \overline{L}_c = \mathbb{E}_P\left[\min\{c, G_H\}\right].$$

The proof uses $c = 1$ to compare $\overline{L}_{\mathrm{H}}(\pi)$ with $L_{\mathrm{H}}(\pi)$, and $c = 1/4$ to bound prefix truncation. For any trajectory distribution U, the auxiliary distribution $\mathcal{R}_c[U]$ assigns probabilities

$$\begin{aligned} \mathcal{R}_c[U](h, \xi_h) &= \frac{e_h^{(c)}(\xi_h)}{c} U(\xi_h), \qquad h \in [H], \\ \mathcal{R}_c[U](\infty, \tau) &= \left(1 - \frac{G_H}{c}\right)_+ U(\tau). \end{aligned} \tag{42}$$

Equivalently, conditional on $\tau \sim U$, the outcome is (h, ξ_h) with probability $e_h^{(c)}(\xi_h)/c$ and (∞, τ) with probability $(1 - G_H/c)_+$. These probabilities sum to one because $\sum_{h=1}^{H} e_h^{(c)}/c = \min\{1, G_H/c\}$. The fixed candidate π determines G_h and $e_h^{(c)}$ for all input distributions U, including $U = P^{\pi'}$.

Lemma B.7 (Hellinger identities for auxiliary distributions)**.** *For every candidate π and $c > 0$,*

$$\begin{aligned} L_{\mathrm{H}}(\pi) &= \mathbb{E}_P\left[\sum_{h=1}^{H} g_h(\xi_h;\pi)\right], \\ D^2_{\mathrm{H}}(\mathcal{R}_c[P],\mathcal{R}_c[Q]) &= \mathbb{E}_P\left[\sum_{h=1}^{H}\left(1-\frac{G_h}{c}\right)_+ g_h(\xi_h;\pi)\right] \le \overline{L}_c. \end{aligned} \tag{43}$$

Proof. Define the prefix affinities $A_h = \sum_{\xi_h}\sqrt{P(\xi_h)Q(\xi_h)}$ for $h \in [H]$ and $A_{H+1} = \sum_{\tau}\sqrt{P(\tau)Q(\tau)}$. The common initial law gives $A_1 = 1$. For $h \in [H]$,

$$\begin{aligned} 2(A_h - A_{h+1}) &= \sum_{\xi_h}\sqrt{P(\xi_h)Q(\xi_h)}\left(2 - 2\sum_a \sqrt{\pi_h(a \mid s_h)\pi_h^\star(a \mid s_h)}\right) \\ &\qquad \text{(Common transition laws)} \\ &= \sum_{\xi_h}\sqrt{P(\xi_h)Q(\xi_h)}D^2_{\mathrm{H}}(\pi_h(\cdot \mid s_h), \pi_h^\star(\cdot \mid s_h)) \qquad \text{(Definition of } D^2_{\mathrm{H}}) \\ &= \mathbb{E}_P\left[g_h(\xi_h;\pi)\right]. \qquad \text{(Algorithm 2)} \end{aligned} \tag{44}$$

Prefixes with $P(\xi_h) = 0$ contribute zero to both sides. Consequently,

$$\begin{aligned} L_{\mathrm{H}}(\pi) &= 2(1 - A_{H+1}) && \text{(Definition of } D^2_{\mathrm{H}}) \\ &= 2\sum_{h=1}^{H}(A_h - A_{h+1}) && (A_1 = 1) \\ &= \mathbb{E}_P\left[\sum_{h=1}^{H} g_h(\xi_h;\pi)\right]. && \text{(Equation (44))} \end{aligned}$$

For the auxiliary distributions, the corresponding affinities before each stopping decision are

$$A_h^{(c)} = \sum_{\xi_h}\left(1-\frac{G_{h-1}}{c}\right)_+\sqrt{P(\xi_h)Q(\xi_h)}, \qquad A_{H+1}^{(c)} = \sum_{\tau}\left(1-\frac{G_H}{c}\right)_+\sqrt{P(\tau)Q(\tau)}.$$

Here $A_1^{(c)} = 1$. Since $(1 - G_{h-1}/c)_+ = e_h^{(c)}/c + (1 - G_h/c)_+$,

$$\begin{aligned} A_h^{(c)} - A_{h+1}^{(c)} &= \sum_{\xi_h}\frac{e_h^{(c)}(\xi_h)}{c}\sqrt{P(\xi_h)Q(\xi_h)} \\ &\quad + \sum_{\xi_h}\left(1-\frac{G_h}{c}\right)_+\sqrt{P(\xi_h)Q(\xi_h)}\left(1 - \sum_a\sqrt{\pi_h(a \mid s_h)\pi_h^\star(a \mid s_h)}\right) \\ &\qquad \text{(Common transition laws)} \\ &= \sum_{\xi_h}\frac{e_h^{(c)}(\xi_h)}{c}\sqrt{P(\xi_h)Q(\xi_h)} \qquad \text{(Definition of } g_h) \\ &\quad + \frac{1}{2}\mathbb{E}_P\left[\left(1-\frac{G_h}{c}\right)_+ g_h(\xi_h;\pi)\right]. \end{aligned} \tag{45}$$

$$
\begin{aligned}
&D_{\mathrm{H}}^2(\mathcal{R}_c[P], \mathcal{R}_c[Q]) \\
&\quad = 2\left(1 - A_{H+1}^{(c)} - \sum_{h=1}^{H}\sum_{\xi_h} \frac{e_h^{(c)}(\xi_h)}{c}\sqrt{P(\xi_h)Q(\xi_h)}\right) && \text{(Equation (42))} \\
&\quad = 2\sum_{h=1}^{H}\left(A_h^{(c)} - A_{h+1}^{(c)} - \sum_{\xi_h} \frac{e_h^{(c)}(\xi_h)}{c}\sqrt{P(\xi_h)Q(\xi_h)}\right) && (A_1^{(c)} = 1) \\
&\quad = \mathbb{E}_P\left[\sum_{h=1}^{H}\left(1 - \frac{G_h}{c}\right)_+ g_h(\xi_h;\pi)\right] && \text{(Equation (45))} \\
&\quad \le \mathbb{E}_P\left[\sum_{h=1}^{H} e_h^{(c)}(\xi_h)\right] = \overline{L}_c. && ((1-G_h/c)_+ g_h \le e_h^{(c)})
\end{aligned}
$$

The pointwise inequality holds because $G_h < c$ implies $e_h^{(c)} = g_h$, whereas $G_h \ge c$ makes its left-hand side zero. □

The definition (4), with e_h^{H} in place of e_h, defines $\mathrm{Trun}_{\pi'}\left[\overline{L}_{\mathrm{H}}(\pi)\right]$.

Lemma B.8 (Hellinger clipping and truncation)**.** *For every $\pi, \pi' \in \Pi$,*

$$\overline{L}_{\mathrm{H}}(\pi) \asymp L_{\mathrm{H}}(\pi), \qquad \overline{L}_{\mathrm{H}}(\pi) \lesssim \mathrm{Trun}_{\pi'}\left[\overline{L}_{\mathrm{H}}(\pi)\right] + \overline{L}_{\mathrm{H}}(\pi'). \tag{46}$$

More precisely,

$$\mathrm{Trun}_{\pi'}\left[\overline{L}_{\mathrm{H}}(\pi)\right] \ge \overline{L}_{\mathrm{H}}(\pi)/32 - L_{\mathrm{H}}(\pi')/4. \tag{47}$$

Also $\overline{L}_{\mathrm{H}}(\pi) \le L_{\mathrm{H}}(\pi)$ and $0 \le \mathrm{Trun}_{\pi'}[\overline{L}_{\mathrm{H}}(\pi)] \le \overline{L}_{\mathrm{H}}(\pi)$.

Proof. Clipping and (43) give $\overline{L}_{\mathrm{H}}(\pi) \le L_{\mathrm{H}}(\pi)$. The bounds $0 \le \mathrm{Trun}_{\pi'}[\overline{L}_{\mathrm{H}}(\pi)] \le \overline{L}_{\mathrm{H}}(\pi)$ follow from (4). For the reverse inequality, let $D = \{(h, \xi_h) : h \in [H]\}$ be the set of prefix outcomes. Then $\mathcal{R}_1[P](D) = \overline{L}_{\mathrm{H}}(\pi)$, and

$$
\begin{aligned}
\sqrt{\mathbb{E}_Q\left[\min\{1, G_H\}\right]} &= \sqrt{\mathcal{R}_1[Q](D)} && \text{(Equation (42))} \\
&\le \sqrt{\mathcal{R}_1[P](D)} + D_{\mathrm{H}}(\mathcal{R}_1[P], \mathcal{R}_1[Q]) && \text{(Equation (65), } F = \mathbf{1}_D) \\
&\le 2\sqrt{\overline{L}_{\mathrm{H}}(\pi)}. && \text{(Equation (43))} \qquad (48)
\end{aligned}
$$

Splitting each trajectory's contribution with $1 = (1 - G_H)_+ + \min\{1, G_H\}$ gives

$$
\begin{aligned}
L_{\mathrm{H}}(\pi) &= \sum_{\tau}\left(\sqrt{\mathcal{R}_1[P](\infty,\tau)} - \sqrt{\mathcal{R}_1[Q](\infty,\tau)}\right)^2 \\
&\quad + \sum_{\tau} \min\{1, G_H\}(\sqrt{P(\tau)} - \sqrt{Q(\tau)})^2 && \text{(Equation (42))} \\
&\le D_{\mathrm{H}}^2(\mathcal{R}_1[P], \mathcal{R}_1[Q]) + \overline{L}_{\mathrm{H}}(\pi) + \mathbb{E}_Q\left[\min\{1, G_H\}\right] && ((\sqrt{p} - \sqrt{q})^2 \le p + q) \\
&\le 6\overline{L}_{\mathrm{H}}(\pi). && \text{(Equations (43) and (48))} \qquad (49)
\end{aligned}
$$

For truncation, set $c = 1/4$, $R = \mathcal{R}_c[P]$, $S = \mathcal{R}_c[Q]$, $S' = \mathcal{R}_c[P^{\pi'}]$, and $z = R(D) = \overline{L}_c/c$. The overlap inequality first compares R with S, and then S with S':

$$
\begin{aligned}
\frac{1}{2}\sqrt{z} &\le \sqrt{\sum_{y \in D} \min\{R(y), S(y)\}} && \text{(Equations (43) and (66), } c = 1/4) \\
&\le \sqrt{\sum_{y \in D} \min\{R(y), S'(y)\}} + D_{\mathrm{H}}(S, S') && \text{(Equation (66))} \\
&\le \sqrt{\sum_{y \in D} \min\{R(y), S'(y)\}} + \sqrt{L_{\mathrm{H}}(\pi')}. && \text{(Equation (64))} \qquad (50)
\end{aligned}
$$

Since $e_h^{(c)} \le e_h^{\mathrm{H}}$ and $\min\{c, G_H\} \ge c \min\{1, G_H\}$,

$$\begin{aligned}
&\mathrm{Trun}_{\pi'}\left[\overline{L}_{\mathrm{H}}(\pi)\right] \\
&\ge \sum_{h=1}^{H} \sum_{\xi_h} \min\{P(\xi_h), P^{\pi'}(\xi_h)\} e_h^{(c)}(\xi_h) && (e_h^{(c)} \le e_h^{\mathrm{H}}) \\
&= c \sum_{y \in D} \min\{R(y), S'(y)\} && \text{(Equation (42))} \\
&\ge c\left(\frac{z}{8} - L_{\mathrm{H}}(\pi')\right) && \text{(Equation (50), } (x+y)^2 \le 2x^2 + 2y^2) \\
&= \frac{\overline{L}_c}{8} - \frac{L_{\mathrm{H}}(\pi')}{4} \ge \frac{\overline{L}_{\mathrm{H}}(\pi)}{32} - \frac{L_{\mathrm{H}}(\pi')}{4}. && (cz = \overline{L}_c,\ \overline{L}_c \ge c\overline{L}_{\mathrm{H}}(\pi))
\end{aligned}$$

This proves (47). Rearranging (47) and using $L_{\mathrm{H}}(\pi') \le 6\overline{L}_{\mathrm{H}}(\pi')$ from (49) gives

$$\overline{L}_{\mathrm{H}}(\pi) \le 32\,\mathrm{Trun}_{\pi'}\left[\overline{L}_{\mathrm{H}}(\pi)\right] + 48\overline{L}_{\mathrm{H}}(\pi'),$$

which proves the second relation in (46). □

Proof of Theorem B.6. The conditional-mean, cancellation, and second-moment calculations in Lemma B.5 use only the form of the loss and nonnegative increments satisfying $\sum_{h=1}^{H} e_h \le 1$. Replacing e_h by e_h^{H} therefore gives $|\ell_t(\pi)| \le 1$ and (33), (30), and (32) with $\overline{L}_{\mathrm{H}}$ in place of $\overline{L}$.

Applying (47) to each direction of the conditional mean in (33), and using $\overline{L}_{\mathrm{H}} \le L_{\mathrm{H}} \le 6\overline{L}_{\mathrm{H}}$, gives

$$\begin{aligned}
-\mathbb{E}_t\left[\ell_t(\pi)\right] &\ge \frac{1}{32} \sum_{\pi' \in \Pi} w_t(\pi') \overline{L}_{\mathrm{H}}(\pi') - \frac{5}{4} L_{\mathrm{H}}(\pi), \\
\mathbb{E}_t\left[\ell_t(\pi)\right] &\ge \frac{1}{32} \overline{L}_{\mathrm{H}}(\pi) - \frac{5}{2} \sum_{\pi' \in \Pi} w_t(\pi') \overline{L}_{\mathrm{H}}(\pi').
\end{aligned}$$

Thus (31) holds with $a = 1/32$, $C = 5/2$, and $L_{\mathrm{H}}, \overline{L}_{\mathrm{H}}$ in place of $L, \overline{L}$, since $5/4 \le 2$. The stepsize $\eta = 1/128$ satisfies $a - 2\eta = 1/64 > 0$. The analysis of concentration and selection by final weight in Appendix B.2.2, from (29) through (40), therefore gives, with probability at least $1 - \delta$,

$$L_{\mathrm{H}}(\widehat{\pi}) \le 6\overline{L}_{\mathrm{H}}(\widehat{\pi}) \lesssim \min_{\pi \in \Pi} L_{\mathrm{H}}(\pi) + \frac{\log(|\Pi|/\delta)}{T},$$

where the first inequality is (49). □

Corollary B.9 (Reverse KL via online Hellinger learning)**.** *Under Assumption 2.2, Algorithm 2 satisfies, with probability at least* $1 - \delta$*,*

$$L(\widehat{\pi}) \lesssim B_\infty \left(\min_{\pi \in \Pi} D_{\mathrm{H}}^2(P^{\pi}, P^{\pi^\star}) + \frac{\log(|\Pi|/\delta)}{T}\right). \tag{51}$$

Proof. Theorem B.6 and Lemma D.3 give (51). Moreover, $\min_{\pi \in \Pi} L_{\mathrm{H}}(\pi) \le L^\star$ by (63), recovering the reverse-KL guarantee with approximation term $B_\infty L^\star$. □

B.4 Estimation and approximation lower bounds

Lower bounds for Theorem 3.3 and Section 4. ***Estimation.*** *For every integer $d \geq 1$, $T \geq 1$, $B \geq 1$, and $\delta \in (0, 1/8]$, every online or offline full-feedback learner using T episodes has a realizable instance with $H = 1$, $|\mathcal{A}| = 2$, and $|\Pi| = 2^d$ satisfying Assumption 2.1 on which, with probability greater than δ,*

$$L(\widehat{\pi}) = \Omega\left(B \min\left\{1, \frac{\log(|\Pi|/\delta)}{T}\right\}\right). \tag{52}$$

Under Assumption 2.2, the same lower bound holds with $B_\infty > 1$ replacing B, including for offline action feedback.

Approximation at optimal estimation. *Fix $\kappa > 0$. There is an integer $k = k(\kappa) \geq 8$ such that, for $B \geq 2 \log k$ and $T \geq k$, any learner returning $\widehat{\pi} \in \Pi$ whose guarantee at confidence $3/4$ has $C_{\mathrm{est}} \leq \kappa B$ on all $H = 2$ instances satisfying Assumption 2.1 with $|\mathcal{A}| = |\Pi| = k$ must have $C_{\mathrm{app}} = \Omega(B)$. This holds for both online and offline full feedback. Under Assumption 2.2, the same conclusion holds with B_∞ replacing B, including for offline action feedback.*

B.4.1 Estimation lower bound

Proof of the estimation lower bound. MDP and policy class. The one-step MDP has

$$\mathcal{A} = \{a_1, a_2\}, \qquad \mathcal{S}_1 = \{s_{1,0}, s_{1,1}, \ldots, s_{1,d}\}, \qquad \Pi = \{\pi_f : f \in \{0,1\}^d\}.$$

For $\rho \in (0, 1]$, the initial law is $\mu(s_{1,x}) = \rho/d$ for $x \in [d]$ and $\mu(s_{1,0}) = 1 - \rho$. The episode ends after the first action. For $\beta \geq 1$, define distributions on (a_1, a_2) by

$$R_0(\beta) = \left(\frac{e^\beta}{1+e^\beta}, \frac{1}{1+e^\beta}\right), \qquad R_1(\beta) = \left(\frac{1}{1+e^\beta}, \frac{e^\beta}{1+e^\beta}\right).$$

For $f \in \{0,1\}^d$ and $x \in [d]$,

$$\pi_{f,1}(\cdot \mid s_{1,x}) = R_{f_x}(\beta), \qquad \pi_{f,1}(\cdot \mid s_{1,0}) = (1/2, 1/2).$$

The teacher is $\pi^\star_\theta = \pi_\theta$ for $\theta \in \{0,1\}^d$, so $L^\star = 0$.

Boundedness. For $x \in [d]$ and each feasible $\tau = (s_1, a_1)$,

$$k_1(s_{1,x}; \pi_f) = \mathrm{KL}(R_{f_x}(\beta) \| R_{\theta_x}(\beta)) = \beta \tanh(\beta/2) \mathbf{1}\{f_x \neq \theta_x\} \leq \beta, \quad \text{(Bernoulli KL formula)}$$

$$\frac{P^{\pi_f}(\tau)}{P^{\pi^\star_\theta}(\tau)} = \frac{\pi_{f,1}(a_1 \mid s_1)}{\pi^\star_{\theta,1}(a_1 \mid s_1)} \in \{1, e^{-\beta}, e^\beta\}. \quad \text{(Common initial law)}$$

Thus $\beta = B$ verifies Assumption 2.1, and $\beta = B_\infty$ verifies Assumption 2.2, for either initial law used below.

Risk at an unvisited state. Write $L_\theta(\pi) = \mathrm{KL}(P^\pi \| P^{\pi^\star_\theta})$. Let θ be uniform on $\{0,1\}^d$, and let M_x count training episodes starting at $s_{1,x}$. For any binary distribution v and $U_2 = (1/2, 1/2)$,

$$\frac{1}{2}\sum_{z=0}^{1} \mathrm{KL}(v \| R_z(\beta)) = \mathrm{KL}(v \| U_2) + b(\beta) \geq b(\beta), \qquad b(\beta) := \log\cosh(\beta/2) \asymp \beta. \tag{53}$$

Let $v_x = \widehat{\pi}_1(\cdot \mid s_{1,x})$. Conditional on $M_x = 0$, the training observations and v_x are independent of θ_x for either trajectory source. Equation (53) gives

$$\mathbb{P}\{\mathrm{KL}(v_x \| R_{\theta_x}(\beta)) \geq b(\beta) \mid M_x = 0\} \geq \frac{1}{2}.$$

Class-size term. Take $\rho = \min\{1, d/(8T)\}$ and set

$$W = \frac{1}{d}\sum_{x=1}^{d} \mathbf{1}\{M_x = 0,\ \mathrm{KL}(v_x \| R_{\theta_x}(\beta)) \geq b(\beta)\}.$$

Then

$$\begin{aligned}
\mathbb{E}[W] &\geq \frac{1}{2d}\sum_{x=1}^{d}\mathbb{P}\{M_x = 0\} = \frac{1}{2}\left(1-\frac{\rho}{d}\right)^T && \text{(Independent initial-state draws)}\\
&\geq \frac{1}{2}\left(1-\frac{T\rho}{d}\right) \geq \frac{7}{16}, && \text{(Bernoulli's inequality)}\\
L_\theta(\widehat{\pi}) &= \sum_{s\in\mathcal{S}_1}\mu(s)\,\mathrm{KL}(\widehat{\pi}_1(\cdot\mid s)\|\pi^\star_{\theta,1}(\cdot\mid s)) && (H=1)\\
&\geq \frac{\rho}{d}\sum_{x=1}^{d}\mathrm{KL}(v_x\|R_{\theta_x}(\beta)) \geq \rho b(\beta)W, && \text{(Definition of } W)\\
\mathbb{P}\left\{L_\theta(\widehat{\pi}) > \frac{\rho b(\beta)}{8}\right\} &\geq \mathbb{P}\{W > 1/8\} \geq \frac{\mathbb{E}[W]-1/8}{1-1/8} \geq \frac{5}{14} > \frac{1}{4}. && (0\leq W\leq 1)
\end{aligned}$$

Here $\rho b(\beta) \asymp \beta\min\{1, \log|\Pi|/T\}$. Averaging over the finite teacher prior gives a fixed teacher with this failure probability.

Confidence term. Keep the same policy class and take $\rho = \min\{1/2, \log(1/(2\delta))/(4T)\}$. Replace the initial law by

$$\mu(s_{1,1}) = \rho, \qquad \begin{cases}\mu(s_{1,0}) = 1-\rho, & d=1,\\ \mu(s_{1,0}) = 0, \quad \mu(s_{1,x}) = (1-\rho)/(d-1) \quad (2\leq x\leq d), & d\geq 2.\end{cases}$$

Only θ_1 is unknown: it is uniform on $\{0,1\}$, and $\theta_x = 0$ for $x \geq 2$. Thus observations away from $s_{1,1}$ reveal no information about θ_1, and

$$\begin{aligned}
\mathbb{P}\{L_\theta(\widehat{\pi}) \geq \rho b(\beta)\} &\geq \tfrac{1}{2}\mathbb{P}\{M_1 = 0\} = \tfrac{1}{2}(1-\rho)^T && \text{(Equation (53))}\\
&\geq \tfrac{1}{2}e^{-2T\rho} \geq \tfrac{1}{2}\sqrt{2\delta} > \delta. && (\log(1-\rho)\geq -2\rho,\ \delta\leq 1/8)
\end{aligned}$$

Here $\rho b(\beta) \asymp \beta\min\{1, \log(1/\delta)/T\}$, and prior averaging again fixes a teacher. Choosing the larger of the two risk thresholds yields

$$\beta\max\left\{\min\left\{1, \frac{\log|\Pi|}{T}\right\}, \min\left\{1, \frac{\log(1/\delta)}{T}\right\}\right\} \asymp \beta\min\left\{1, \frac{\log(|\Pi|/\delta)}{T}\right\}.$$

Both arguments allow full feedback from either trajectory source. Restricting feedback gives the offline action-feedback bound. □

B.4.2 Approximation lower bound

Proof of the approximation lower bound. MDP and policy class. Let $k \geq 8$ be chosen below in terms of κ, and set $\rho = (k-1)/(8T)$. The two-step MDP has

$$\begin{gathered}
\mathcal{A} = \{a_1,\dots,a_k\}, \qquad \mathcal{S}_1 = \{s_{1,0}, s_{1,1}\}, \qquad \mathcal{S}_2 = \{s_{2,0}, s_{2,1}, \dots, s_{2,k}\},\\
\mu(s_{1,1}) = \rho, \qquad \mu(s_{1,0}) = 1-\rho,\\
\mathsf{P}_1(s_{2,j}\mid s_{1,1}, a_j) = 1, \qquad \mathsf{P}_1(s_{2,0}\mid s_{1,0}, a_j) = 1 \quad (j\in[k]).
\end{gathered}$$

The episode ends after the second action. The class is $\Pi = \{\pi_i : i\in[k]\}$, and the teacher is indexed by $\theta\in[k]$. Their first-step distributions at $s_{1,1}$ are

$$\pi_{i,1}(a_j\mid s_{1,1}) = \begin{cases}3/4, & j=i,\\ 1/[4(k-1)], & j\neq i,\end{cases} \qquad \pi^\star_{\theta,1}(a_j\mid s_{1,1}) = 1/k.$$

For $z\in\{0,1\}$, define distributions on $\mathcal{A}$ by

$$R_z(a_1) = \frac{\exp((B/2)\mathbf{1}\{z=0\})}{1+e^{B/2}}, \qquad R_z(a_j) = \frac{\exp((B/2)\mathbf{1}\{z=1\})}{(1+e^{B/2})(k-1)} \quad (2\leq j\leq k).$$

At the second layer,

$$\pi_{i,2}(\cdot\mid s_{2,j}) = R_{\mathbf{1}\{j=i\}}, \qquad \pi^\star_{\theta,2}(\cdot\mid s_{2,j}) = R_{\mathbf{1}\{j=\theta\}} \quad (j\in[k]).$$

Every policy is uniform at $s_{1,0}$ and $s_{2,0}$. The learner may know the initial law and transitions and explore adaptively, but cannot reset within an episode or query unvisited states. Its output belongs to Π.

Boundedness and trajectory risks. Write $L_\theta(\pi) = \mathrm{KL}(P^\pi \| P^{\pi^\star_\theta})$ and $d_k = \mathrm{KL}(\mathrm{Ber}(3/4) \| \mathrm{Ber}(1/k)) \le \log k$. For $B \ge 2\log k$,

$$\mathrm{KL}(R_0\|R_1) = \mathrm{KL}(R_1\|R_0) = \frac{B}{2}\tanh(B/4) \in [B/4, B/2], \qquad \text{(Bernoulli KL formula, } k \ge 8)$$

$$k_1(s_{1,1};\pi_i) = d_k, \qquad k_2(s_{2,j};\pi_i) = \mathrm{KL}(R_0\|R_1)\mathbf{1}\{\mathbf{1}\{j=i\} \ne \mathbf{1}\{j=\theta\}\}.$$

(The specified action distributions)

All action probabilities are positive, and every feasible trajectory satisfies

$$\sum_{h=1}^{2} k_h(s_h;\pi_i) \le \log k + B/2 \le B,$$

$$\log \frac{P^{\pi_i}(\tau)}{P^{\pi^\star_\theta}(\tau)} = \sum_{h=1}^{2} \log \frac{\pi_{i,h}(a_h \mid s_h)}{\pi^\star_{\theta,h}(a_h \mid s_h)} \le \log k + B/2 \le B.$$

(First-step ratio $\le k$, second-step ratio $\le e^{B/2}$)

Thus both assumptions hold, with $B_\infty = B$. If $i \ne \theta$, only branches $j = i$ and $j = \theta$ have nonzero second-step KL. The KL chain rule gives

$$\begin{aligned} L_\theta(\pi_i) &= \rho d_k + \rho\left(\frac{3}{4} + \frac{1}{4(k-1)}\right)\mathrm{KL}(R_0\|R_1)\mathbf{1}\{i \ne \theta\}, \\ L^\star &= L_\theta(\pi_\theta) = \rho d_k. \end{aligned} \tag{54}$$

Identification error. Let θ be uniform on $[k]$. The teacher's first-step distribution is independent of θ. A trajectory taking a_j at $s_{1,1}$ reaches $s_{2,j}$, where full feedback reveals exactly $\mathbf{1}\{j=\theta\}$. One equality test also simulates an offline teacher trajectory: its first action is uniform, and its second follows $R_{\mathbf{1}\{j=\theta\}}$. Let $M = \sum_{t=1}^{T} \mathbf{1}\{s_{t,1} = s_{1,1}\}$ and write $\widehat{\pi} = \pi_{\widehat{i}}$. Conditional on the initial-state sequence, M is fixed and θ remains uniform on $[k]$. Therefore,

$$\begin{aligned} \mathbb{P}\{\widehat{i} \ne \theta\} &\ge \mathbb{E}\left[1 - \frac{M+1}{k}\right] && \text{(Lemma D.6, conditionally then averaged)} \\ &= \frac{7(k-1)}{8k} \ge \frac{3}{4}, && (\mathbb{E}[M] = T\rho = (k-1)/8,\ k \ge 8) \\ L_\theta(\widehat{\pi}) - L^\star &\ge \frac{3\rho B}{16} \quad \text{on } \{\widehat{i} \ne \theta\}. && \text{(Equation (54), } \mathrm{KL}(R_0\|R_1) \ge B/4) \end{aligned}$$

Prior averaging fixes a teacher for which this event has probability at least $3/4$.

Approximation coefficient. The event above intersects the success event of any guarantee at confidence $3/4$, so

$$\begin{aligned} \frac{3\rho B}{16} &\le (C_{\mathrm{app}} - 1)L^\star + C_{\mathrm{est}}\frac{\log(4k)}{T} && \text{(Guarantee at confidence } 3/4) \\ \frac{3B}{16} &\le (C_{\mathrm{app}} - 1)d_k + \frac{8C_{\mathrm{est}}\log(4k)}{k-1}. && (L^\star = \rho d_k,\ \rho = (k-1)/(8T)) \end{aligned} \tag{55}$$

Choose the fixed integer $k = k(\kappa) \ge 8$ so that $8\kappa\log(4k)/(k-1) \le 3/32$. Then

$$\begin{aligned} \frac{8C_{\mathrm{est}}\log(4k)}{k-1} &\le \frac{8\kappa B\log(4k)}{k-1} \le \frac{3B}{32}, && (C_{\mathrm{est}} \le \kappa B) \\ (C_{\mathrm{app}} - 1)d_k &\ge \frac{3B}{32}, && \text{(Equation (55))} \\ C_{\mathrm{app}} &\ge 1 + \frac{3B}{32\log k} = \Omega(B). && (d_k \le \log k, \text{ fixed } k = k(\kappa)) \end{aligned}$$

Replacing B by B_∞ gives the ratio-bound result. Restricting the offline observations to action feedback preserves the lower bound. □

B.5 THE ADDITIONAL COST OF ONLINE ACTION FEEDBACK

Theorem 3.5 (detailed statement). *For $T \geq 1$, $B_\infty > 1$, and $|\Pi|, |\mathcal{A}| \geq 2$, let $m = \min\{|\mathcal{A}|, |\Pi|, e^{B_\infty}\}$. Every online action-feedback learner with arbitrary causal exploration and output in Π has a realizable $H = 1$ instance with a full-support candidate class satisfying Assumption 2.2 such that, with probability at least $1/4$,*

$$L(\widehat{\pi}) = \Omega\left(B_\infty \min\left\{1, \frac{m \log |\Pi|}{T \log(em)}\right\}\right). \tag{56}$$

Proof of Theorem 3.5. MDP and policy class. Write $A = |\mathcal{A}|$, $N = |\Pi|$, and set

$$m = \min\{A, N, e^{B_\infty}\}, \qquad k = \lfloor m \rfloor, \qquad d = \lfloor \log N / \log k \rfloor, \qquad \rho = \min\left\{1, \frac{(k-1)d}{8T}\right\}.$$

The one-step MDP has

$$\mathcal{A} = \{a_1, \ldots, a_A\}, \qquad \mathcal{S}_1 = \{s_{1,0}, s_{1,1}, \ldots, s_{1,d}, s_{1,d+1}\},$$
$$\mu(s_{1,x}) = \rho/d \quad (x \in [d]), \qquad \mu(s_{1,0}) = 1 - \rho, \qquad \mu(s_{1,d+1}) = 0.$$

The episode ends after the first action. The state $s_{1,d+1}$ will distinguish additional candidates if $k^d < N$. For $f \in [k]^d$ and $x \in [d]$, define

$$\pi_{f,1}(a_j \mid s_{1,x}) = \begin{cases} \dfrac{A+k}{2A} \dfrac{\exp(B_\infty \mathbf{1}\{j = f_x\})}{e^{B_\infty} + k - 1}, & 1 \leq j \leq k, \\ \dfrac{1}{2A}, & k < j \leq A. \end{cases} \tag{57}$$

The last $A - k$ actions have total probability $(A-k)/(2A)$; the first k have the remaining probability $(A+k)/(2A) \geq 1/2$. At $s_{1,0}$ and $s_{1,d+1}$, every π_f is uniform. The class contains all k^d policies π_f. The remaining $N - k^d$ candidates agree with $\pi_{(1,\ldots,1)}$ except at $s_{1,d+1}$, where they use distinct, nonuniform, strictly positive action distributions. The teacher is $\pi^\star_\theta = \pi_\theta$, $\theta \in [k]^d$, so $L^\star = 0$.

Boundedness and trajectory risks. Write $L_\theta(\pi) = \mathrm{KL}(P^\pi \| P^{\pi^\star_\theta})$. All action probabilities are positive. On a trajectory starting at $s_{1,x}$ with action a_j,

$$\frac{P^{\pi_f}(\tau)}{P^{\pi^\star_\theta}(\tau)} = \begin{cases} \exp(B_\infty[\mathbf{1}\{j = f_x\} - \mathbf{1}\{j = \theta_x\}]), & x \in [d],\ j \leq k, \\ 1, & \text{otherwise} \end{cases} \leq e^{B_\infty}.$$

This also holds for the additional candidates, whose trajectory law is $P^{\pi_{(1,\ldots,1)}}$. For $f_x \neq \theta_x$,

$$\begin{aligned} k_1(s_{1,x}; \pi_f) &= B_\infty\big(\pi_{f,1}(a_{f_x} \mid s_{1,x}) - \pi_{f,1}(a_{\theta_x} \mid s_{1,x})\big) \quad \text{(Only } a_{f_x} \text{ and } a_{\theta_x} \text{ have nonunit ratios)} \\ &= \frac{A+k}{2A} \frac{B_\infty(e^{B_\infty} - 1)}{e^{B_\infty} + k - 1} \asymp B_\infty. \quad (k \leq e^{B_\infty}, B_\infty > 1, 1/2 \leq (A+k)/(2A) \leq 1) \end{aligned}$$

Consequently,

$$L_\theta(\pi_f) = \frac{\rho(A+k)}{2Ad} \frac{B_\infty(e^{B_\infty} - 1)}{e^{B_\infty} + k - 1} \sum_{x=1}^{d} \mathbf{1}\{f_x \neq \theta_x\} \asymp \frac{\rho B_\infty}{d} \sum_{x=1}^{d} \mathbf{1}\{f_x \neq \theta_x\}. \tag{58}$$

Identification error and risk. Let θ be uniform on $[k]^d$. By (57), querying a_j at $s_{1,x}$ reveals exactly $\mathbf{1}\{j = \theta_x\}$ for $j \leq k$, and reveals no information about θ_x for $j > k$. Feedback at $s_{1,0}$ is independent of θ. Let $M_x = \sum_{t=1}^{T} \mathbf{1}\{s_{t,1} = s_{1,x}\}$ and write the output as $\pi_{\widehat{f}}$, assigning label $(1, \ldots, 1)$ to the additional candidates. Conditional on the initial-state sequence and all other labels, M_x is fixed and θ_x remains uniform on $[k]$. Lemma D.6, followed by averaging, gives

$$\begin{aligned} \mathbb{P}\{\widehat{f}_x \neq \theta_x\} &\geq \mathbb{E}\left[1 - \frac{M_x + 1}{k}\right] && \text{(Lemma D.6)} \\ &\geq 1 - \frac{(k-1)/8 + 1}{k} = \frac{7(k-1)}{8k} \geq \frac{7}{16}. && (\mathbb{E}[M_x] = T\rho/d \leq (k-1)/8,\ k \geq 2) \end{aligned}$$

For $W = d^{-1}\sum_{x=1}^{d} \mathbf{1}\{\widehat{f}_x \neq \theta_x\}$,

$$\mathbb{E}[W] \geq 7/16, \qquad \mathbb{P}\{W > 1/8\} \geq \frac{\mathbb{E}[W] - 1/8}{1 - 1/8} \geq \frac{5}{14}, \qquad (0 \leq W \leq 1)$$

$$L_\theta(\widehat{\pi}) \gtrsim \rho B_\infty \quad \text{on } \{W > 1/8\}, \qquad \text{(Equation (58))}$$

$$\rho B_\infty \asymp B_\infty \min\left\{1, \frac{kd}{T}\right\} \qquad (k \geq 2 \text{ and the definition of } \rho)$$

$$\asymp B_\infty \min\left\{1, \frac{m \log N}{T \log(em)}\right\}. \qquad (k \asymp m,\ d \asymp \log N / \log k,\ \log k \asymp \log(em))$$

Averaging over the finite teacher prior gives a fixed teacher with failure probability at least $5/14 > 1/4$. □

C Offline learning

C.1 Full feedback

Proof of Proposition 4.1. If $\pi^\star \in \Pi$, both empirical objectives have minimum zero. Nonnegativity of the local KL losses gives, for every $\pi \in \Pi$,

$$\sum_{t=1}^{T}\sum_{h=1}^{H} k_h(s^\star_{t,h}; \pi) = 0 \quad \Longleftrightarrow \quad \sum_{t=1}^{T} \min\left\{1, \sum_{h=1}^{H} k_h(s^\star_{t,h}; \pi)\right\} = 0.$$

Thus (11) and (12) have the same minimizers, and Theorem 4.2 with $L^\star = 0$ proves the first claim. The lower bound (52) shows that this estimation rate is minimax optimal.

MDP and policy class. For the misspecified instance, fix $B \geq 32$ and set

$$\epsilon = \frac{1}{B}, \qquad p = \frac{1}{16 \log B}, \qquad q = \frac{p}{B}.$$

The $H = 2$ MDP has action set $\mathcal{A} = \{a_1, a_2\}$ and state layers $\mathcal{S}_1 = \{s_{1,1}, s_{1,2}\}$ and $\mathcal{S}_2 = \{s_{2,0}, s_{2,1}, s_{2,2}\}$. The initial law is $\mu(s_{1,1}) = \mu(s_{1,2}) = 1/2$, and

$$\mathsf{P}_1(s_{2,j} \mid s_{1,j}, a_1) = 1, \qquad \mathsf{P}_1(s_{2,0} \mid s_{1,j}, a_2) = 1 \quad (j \in \{1, 2\}).$$

The episode ends after the second action. The table gives the probability of a_1 at each state, with its complement assigned to a_2. The candidate class is $\Pi = \{\pi_{\mathrm{good}}, \pi_{\mathrm{bad}}\}$.

	$s_{1,1}$	$s_{1,2}$	$s_{2,0}$	$s_{2,1}$	$s_{2,2}$
$\pi^\star$	ϵ	q	0	$e^{-B/2}$	$e^{-B/2}$
π_{good}	0	q	0	1	$e^{-B/2}$
π_{bad}	ϵ	p	0	$e^{-B/2}$	1

Boundedness. The only nonzero local KL losses are

$$k_1(s_{1,1}; \pi_{\mathrm{good}}) = \ell := -\log(1 - \epsilon), \qquad k_2(s_{2,1}; \pi_{\mathrm{good}}) = B/2,$$

$$k_1(s_{1,2}; \pi_{\mathrm{bad}}) = r := p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}, \quad k_2(s_{2,2}; \pi_{\mathrm{bad}}) = B/2.$$

Since $0 < q < p < 1$,

$$\epsilon \leq \ell \leq 2\epsilon, \qquad 0 \leq r \leq p \log(p/q) = \frac{1}{16}, \qquad (\epsilon \leq 1/2,\ p/q = B)$$

$$\sum_{h=1}^{2} k_h(s_h; \pi) \leq \frac{B}{2} + \max\{\ell, r\} \leq B \qquad \text{for every } \pi \in \Pi \text{ and } \tau \in \mathcal{T}_{\mathrm{env}}.$$

This verifies Assumption 2.1, including trajectories that a candidate never generates.

Population objectives and empirical selection. Under the teacher, $s_{2,1}$ and $s_{2,2}$ have probabilities $\epsilon/2$ and $q/2$. Under π_{good}, $s_{2,1}$ is never visited, whereas π_{bad} visits $s_{2,2}$ with probability $p/2$. Consequently,

	$\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{2} k_h(s_h;\pi)\right]$	$L(\pi)$
π_{good}	$(\ell+\epsilon B/2)/2$	$\ell/2$
π_{bad}	$(r+qB/2)/2$	$(r+pB/2)/2$

The untruncated objective ranks the candidates in the wrong order,

$$\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{2} k_h(s_h;\pi_{\mathrm{bad}})\right] \le \frac{1}{32}+\frac{p}{4} \le \frac{1}{16}, \qquad (r \le 1/16,\, q = p/B)$$

$$\mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{2} k_h(s_h;\pi_{\mathrm{good}})\right] \ge \frac{\epsilon B}{4} = \frac{1}{4}, \qquad (\ell \ge 0,\, \epsilon = 1/B)$$

$$\frac{L(\pi_{\mathrm{bad}})}{L(\pi_{\mathrm{good}})} = \frac{r+pB/2}{\ell} \ge \frac{pB}{4\epsilon} = \frac{B^2}{64\log B} > 1. \qquad (r \ge 0,\, \ell \le 2\epsilon,\, B \ge 32)$$

In particular, $L^\star = L(\pi_{\mathrm{good}}) > 0$. For each $\pi \in \Pi$, the strong law of large numbers gives, with probability one,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{2} k_h(s^\star_{t,h};\pi) \longrightarrow \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{2} k_h(s_h;\pi)\right].$$

The population objective of π_{bad} is smaller by at least $3/16$, so

$$\mathbb{P}\{\widehat{\pi}_{\mathrm{raw}} = \pi_{\mathrm{bad}}\} \longrightarrow 1.$$

Consequently, any guarantee with an estimation term vanishing as $T \to \infty$ requires

$$C_{\mathrm{app}} \ge \frac{L(\pi_{\mathrm{bad}})}{L^\star} \ge \frac{B^2}{64\log B} = \Omega\left(\frac{B^2}{\log(eB)}\right).$$

□

Proof of Theorem 4.2. Each teacher trajectory contributes $\sum_{h=1}^{H} e_h(\xi^\star_{t,h};\pi) \in [0,1]$ to (12), and the teacher trajectories are independent. Let $\pi^\dagger \in \arg\min_{\pi\in\Pi} L(\pi)$. With probability at least $1-\delta$,

$$\begin{aligned}
\overline{L}(\widehat{\pi}) &\lesssim \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\widehat{\pi})\right] && \text{(Equation (20))}\\
&\lesssim \min_{\pi\in\Pi} \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi)\right] + \frac{\log(|\Pi|/\delta)}{T} && \text{(Lemma D.5)}\\
&\le \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} e_h(\xi_h;\pi^\dagger)\right] + \frac{\log(|\Pi|/\delta)}{T} && (\pi^\dagger \in \Pi)\\
&\lesssim \overline{L}(\pi^\dagger) + \frac{\log(|\Pi|/\delta)}{T} && \text{(Equation (20))}\\
&\le L^\star + \frac{\log(|\Pi|/\delta)}{T}. && (\overline{L}(\pi^\dagger) \le L(\pi^\dagger) = L^\star) \qquad (59)
\end{aligned}$$

The expectations under $P^{\pi^\star}$ evaluate each displayed candidate on an independent teacher trajectory. Under Assumption 2.1,

$$\begin{aligned}
L(\widehat{\pi}) &\le B\overline{L}(\widehat{\pi}) && \text{(Equation (3))}\\
&\lesssim B\left[L^\star + \frac{\log(|\Pi|/\delta)}{T}\right]. && \text{(Equation (59))}
\end{aligned}$$

Under Assumption 2.2,

$$\begin{aligned} L(\widehat{\pi}) &\lesssim B_\infty D_{\mathrm{H}}^2(P^{\widehat{\pi}}, P^{\pi^\star}) && \text{(Lemma D.3, } B_\infty > 1\text{)} \\ &\lesssim B_\infty \overline{L}(\widehat{\pi}) && \text{(Equation (20))} \\ &\lesssim B_\infty \left[L^\star + \frac{\log(|\Pi|/\delta)}{T} \right]. && \text{(Equation (59))} \end{aligned}$$

□

C.2 Action feedback

Proposition C.1 (Insufficiency of the accumulated KL bound)**.** *For every $B \geq 1$ and integer $T \geq 1$, there exist an $H = 1$ MDP with one state and three actions and a class Π of two policies such that every possibly randomized offline action-feedback learner returning $\widehat{\pi} \in \Pi$ has a teacher $\pi^\star \in \Pi$ for which Assumption 2.1 holds and*

$$L^\star = 0, \qquad \mathbb{P}\{L(\widehat{\pi}) \geq B/2\} \geq \frac{1}{4}.$$

Proof. Fix $B \geq 1$ and $T \geq 1$.

MDP and policy class. The state and action sets are $\mathcal{S} = \{s_{1,1}\}$ and $\mathcal{A} = \{a_1, a_2, a_3\}$, with $\mu(s_{1,1}) = 1$. The episode ends after the first action. Let

$$\epsilon = \frac{1}{4T}, \qquad r = \epsilon e^{-B/\epsilon}, \qquad \Pi = \{\pi_0, \pi_1\},$$

where the action probabilities at $s_{1,1}$ are

	a_1	a_2	a_3
π_0	$1-\epsilon-r$	ϵ	r
π_1	$1-\epsilon-r$	r	ϵ

The teacher is $\pi^\star = \pi_\theta$ for an unknown $\theta \in \{0, 1\}$.

Boundedness and candidate risks. All action probabilities are positive, and for either teacher,

$$\begin{aligned} k_1(s_{1,1}; \pi_{1-\theta}) &= \mathrm{KL}(P^{\pi_{1-\theta}} \| P^{\pi_\theta}) = \epsilon \log \frac{\epsilon}{r} + r \log \frac{r}{\epsilon} && \text{(The two policies agree on } a_1\text{)} \\ &= (\epsilon - r)\frac{B}{\epsilon} = B\big(1 - e^{-B/\epsilon}\big) \in [B/2, B]. && (B/\epsilon = 4BT \geq 4) \end{aligned} \tag{60}$$

Since $k_1(s_{1,1}; \pi_\theta) = 0$, Assumption 2.1 holds for both candidates and $L^\star = L(\pi_\theta) = 0$.

Indistinguishable observations. Write $\mathbb{P}_\theta$ for the law of the training data and the learner's output when $\pi^\star = \pi_\theta$. For $E = \{a^\star_{t,1} = a_1 \text{ for all } t \in [T]\}$,

$$\begin{aligned} \mathbb{P}_\theta(E) &= (1 - \epsilon - r)^T && \text{(Independent teacher trajectories)} \\ &\geq 1 - T(\epsilon + r) \geq 1 - 2T\epsilon = \frac{1}{2}. && \text{(Bernoulli's inequality, } r \leq \epsilon\text{)} \end{aligned}$$

On E, both teachers produce the same T observations $(s_{1,1}, a_1, 1 - \epsilon - r)$, including the reported action probabilities. Thus the conditional output distributions coincide, and $\widehat{\pi} \in \{\pi_0, \pi_1\}$ implies

$$\mathbb{P}_0(\widehat{\pi} \neq \pi_0 \mid E) + \mathbb{P}_1(\widehat{\pi} \neq \pi_1 \mid E) = 1.$$

Consequently,

$$\begin{aligned} \max_{\theta \in \{0,1\}} \mathbb{P}_\theta\Big\{ \mathrm{KL}(P^{\widehat{\pi}} \| P^{\pi_\theta}) \geq B/2 \Big\} &= \max_{\theta \in \{0,1\}} \mathbb{P}_\theta\{\widehat{\pi} \neq \pi_\theta\} && \text{(Equation (60))} \\ &\geq \frac{1}{2} \sum_{\theta=0}^{1} \mathbb{P}_\theta(\{\widehat{\pi} \neq \pi_\theta\} \cap E) && \text{(Maximum bounds the average)} \\ &= \frac{(1 - \epsilon - r)^T}{2} \geq \frac{1}{4}. && \text{(Identical observations on } E\text{)} \end{aligned}$$

□

Proof of Proposition 4.3. The estimator in Rohatgi et al. (2025, Appendix C, Eq. (11)) maximizes the empirical log-likelihood of $(P^{\pi} + P^{\pi^\star})/2$ over $\pi \in \Pi$. Removing terms independent of π gives the equivalent objective

$$\widehat{\pi} \in \arg\min_{\pi\in\Pi} -\sum_{t=1}^{T} \log\left(1 + \frac{P^{\pi}(\tau_t^\star)}{P^{\pi^\star}(\tau_t^\star)}\right). \tag{61}$$

For every observed teacher trajectory, the common initial-state and transition probabilities cancel, so

$$\frac{P^{\pi}(\tau_t^\star)}{P^{\pi^\star}(\tau_t^\star)} = \prod_{h=1}^{H} \frac{\pi_h(a_{t,h}^\star \mid s_{t,h}^\star)}{\pi_h^\star(a_{t,h}^\star \mid s_{t,h}^\star)}. \tag{62}$$

Thus action feedback suffices to evaluate (61) without knowing the dynamics.

Under Assumption 2.2, with probability at least $1-\delta$,

$$\begin{aligned}
L(\widehat{\pi}) &\lesssim B_\infty D_{\mathrm{H}}^2(P^{\widehat{\pi}}, P^{\pi^\star}) && \text{(Lemma D.3, } B_\infty > 1\text{)} \\
&\lesssim B_\infty \left[\min_{\pi\in\Pi} D_{\mathrm{H}}^2(P^{\pi}, P^{\pi^\star}) + \frac{\log(|\Pi|/\delta)}{T}\right] && \text{(Rohatgi et al., 2025, Proposition C.1)} \\
&\le B_\infty \left[L^\star + \frac{\log(|\Pi|/\delta)}{T}\right]. && \text{(Equation (63))}
\end{aligned}$$

□

D Auxiliary lemmas

D.1 KL and Hellinger inequalities

All distributions below are on a common discrete space. We write $P \ll Q$ when $Q(z) = 0$ implies $P(z) = 0$. Ratios $P(z)/Q(z)$ are evaluated where $Q(z) > 0$, and equal zero when $P(z) = 0$.

Lemma D.1 (KL variational inequality (Polyanskiy & Wu, 2024, Theorem 4.6))**.** *Let R, S be probability distributions with $\mathrm{KL}(R\|S) < \infty$. For every real-valued function g satisfying $0 < \mathbb{E}_S[e^g] < \infty$,*

$$\mathrm{KL}(R\|S) \ge \mathbb{E}_R[g] - \log \mathbb{E}_S[e^g].$$

Lemma D.2 (Elementary Hellinger inequalities)**.** *For probability distributions P, Q,*

$$D_{\mathrm{H}}^2(P,Q) = \sum_z (\sqrt{P(z)} - \sqrt{Q(z)})^2 \le \mathrm{KL}(P\|Q). \tag{63}$$

For conditional probabilities $r(y \mid x) \ge 0$ with $\sum_y r(y \mid x) = 1$, define $\widetilde{P}(y) = \sum_x r(y \mid x)P(x)$ and $\widetilde{Q}(y) = \sum_x r(y \mid x)Q(x)$. Then

$$D_{\mathrm{H}}(\widetilde{P}, \widetilde{Q}) \le D_{\mathrm{H}}(P, Q). \tag{64}$$

For a function F with $0 \le F(z) \le M$,

$$\sqrt{\mathbb{E}_P[F]} \le \sqrt{\mathbb{E}_Q[F]} + \sqrt{M}\, D_{\mathrm{H}}(P, Q). \tag{65}$$

For distributions R, S, S' and a subset D of their common space,

$$\sqrt{\sum_{z\in D} \min\{R(z), S(z)\}} \le \sqrt{\sum_{z\in D} \min\{R(z), S'(z)\}} + D_{\mathrm{H}}(S, S'). \tag{66}$$

Proof. For (63), only $\mathrm{KL}(P\|Q) < \infty$ needs proof. The scalar inequality $(\sqrt{x}-1)^2 \le x\log x - x + 1$ for $x \ge 0$ gives

$$\begin{aligned}
D_{\mathrm{H}}^2(P,Q) &= \mathbb{E}_Q\left[\left(\sqrt{P/Q} - 1\right)^2\right] && (P \ll Q) \\
&\le \mathbb{E}_Q\left[(P/Q)\log(P/Q) - P/Q + 1\right] && \text{(Scalar inequality)} \\
&= \mathrm{KL}(P\|Q). && (\textstyle\sum_z P(z) = \sum_z Q(z) = 1)
\end{aligned}$$

For (64),

$$\begin{aligned} D_{\mathrm{H}}^2(\widetilde{P},\widetilde{Q}) &= 2 - 2\sum_y \sqrt{\Big(\sum_x r(y \mid x)P(x)\Big)\Big(\sum_x r(y \mid x)Q(x)\Big)} && \text{(Definition of } \widetilde{P},\widetilde{Q}) \\ &\le 2 - 2\sum_y \sum_x r(y \mid x)\sqrt{P(x)Q(x)} && \text{(Cauchy–Schwarz)} \\ &= 2 - 2\sum_x \sqrt{P(x)Q(x)} = D_{\mathrm{H}}^2(P,Q). && (\textstyle\sum_y r(y \mid x) = 1) \end{aligned}$$

For (65),

$$\begin{aligned} \sqrt{\mathbb{E}_P[F]} &= \|\sqrt{F}\sqrt{P}\|_2 && \text{(Definition of } \|\cdot\|_2) \\ &\le \|\sqrt{F}\sqrt{Q}\|_2 + \|\sqrt{F}(\sqrt{P} - \sqrt{Q})\|_2 && \text{(Triangle inequality)} \\ &\le \sqrt{\mathbb{E}_Q[F]} + \sqrt{M}\, D_{\mathrm{H}}(P,Q). && (0 \le F \le M) \end{aligned}$$

For (66), the minimum is taken pointwise, and

$$\begin{aligned} \sqrt{\sum_{z\in D} \min\{R(z), S(z)\}} &= \|\mathbf{1}_D \min\{\sqrt{R}, \sqrt{S}\}\|_2 && \text{(Definition of } \|\cdot\|_2) \\ &\le \|\mathbf{1}_D \min\{\sqrt{R}, \sqrt{S'}\}\|_2 + \|\mathbf{1}_D(\sqrt{S} - \sqrt{S'})\|_2 \\ &\qquad \text{(Triangle inequality and 1-Lipschitz minimum)} \\ &\le \sqrt{\sum_{z\in D} \min\{R(z), S'(z)\}} + D_{\mathrm{H}}(S,S'). && (\mathbf{1}_D \le 1) \end{aligned}$$

□

Lemma D.3 (Hellinger-to-KL conversion). *If $P \ll Q$ and $P(z)/Q(z) \le e^b$ for $b \ge 0$, then*

$$D_{\mathrm{H}}^2(P,Q) \le \mathrm{KL}(P\|Q) \lesssim (1+b) D_{\mathrm{H}}^2(P,Q), \qquad \mathrm{KL}(P\|Q) \le b.$$

Proof. For $0 \le x \le e^b$, the scalar comparison underlying Sason & Verdú (2016, Theorem 9), at Hellinger order $1/2$, gives $x \log x - x + 1 \lesssim (1+b)(\sqrt{x}-1)^2$. Hence

$$\begin{aligned} D_{\mathrm{H}}^2(P,Q) &\le \mathrm{KL}(P\|Q) && \text{(Equation (63))} \\ &= \mathbb{E}_Q[(P/Q)\log(P/Q) - P/Q + 1] && (\mathbb{E}_Q[P/Q] = 1) \\ &\lesssim (1+b)\mathbb{E}_Q\left[\left(\sqrt{P/Q} - 1\right)^2\right] && \text{(Scalar comparison, } P/Q \le e^b) \\ &= (1+b) D_{\mathrm{H}}^2(P,Q). && \text{(Definition of } D_{\mathrm{H}}^2) \end{aligned}$$

Also $\mathrm{KL}(P\|Q) = \mathbb{E}_P[\log(P/Q)] \le b$. □

D.2 Concentration and testing bounds

Let $(\mathcal{F}_t)_{t=0}^T$ be the training-history filtration. The history $\mathcal{F}_{t-1}$ includes all observations and randomization before episode t, and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. In particular, π_t and its trajectory are both averaged over unless π_t is explicitly included in the conditioning.

Lemma D.4 (A fixed-time exponential inequality). *Let X_t be $\mathcal{F}_t$-measurable with $|X_t| \le 2$. For fixed T, $0 < \lambda \le 1/2$, and $\epsilon \in (0,1)$, each of*

$$\sum_{t=1}^T X_t \le \sum_{t=1}^T \mathbb{E}_t[X_t] + \lambda \sum_{t=1}^T \mathbb{E}_t\left[X_t^2\right] + \frac{\log(1/\epsilon)}{\lambda}, \tag{67}$$

$$\sum_{t=1}^T \mathbb{E}_t[X_t] \le \sum_{t=1}^T X_t + \lambda \sum_{t=1}^T \mathbb{E}_t\left[X_t^2\right] + \frac{\log(1/\epsilon)}{\lambda} \tag{68}$$

holds with probability at least $1-\epsilon$.

Proof. The exponential inequality of Foster et al. (2024, Lemma D.1), applied to $-\lambda X_t$, gives, with probability at least $1-\epsilon$,

$$\begin{aligned}
\sum_{t=1}^{T} X_t &\le \frac{1}{\lambda}\sum_{t=1}^{T} \log \mathbb{E}_t\left[e^{\lambda X_t}\right] + \frac{\log(1/\epsilon)}{\lambda} && \text{(Foster et al., 2024, Lemma D.1)}\\
&\le \frac{1}{\lambda}\sum_{t=1}^{T} \log\big(1+\lambda\mathbb{E}_t\left[X_t\right] + \lambda^2\mathbb{E}_t\left[X_t^2\right]\big)\\
&\quad + \frac{\log(1/\epsilon)}{\lambda} && (e^x \le 1+x+x^2,\ |x|\le 1)\\
&\le \sum_{t=1}^{T}\mathbb{E}_t\left[X_t\right] + \lambda\sum_{t=1}^{T}\mathbb{E}_t\left[X_t^2\right] + \frac{\log(1/\epsilon)}{\lambda}. && (\log(1+x)\le x)
\end{aligned}$$

This is (67). Applying the same inequality to $-X_t$ gives (68). □

Lemma D.5 (Empirical minimization of bounded scores)**.** *Let the vectors $(Z_t(\pi))_{\pi\in\Pi}$ be independent and identically distributed across t, with $0 \le Z_t(\pi) \le 1$ and $\mu(\pi) = \mathbb{E}\left[Z_t(\pi)\right]$. Dependence across candidates within an episode is allowed. Every empirical minimizer $\widehat{\pi} \in \arg\min_{\pi\in\Pi}\sum_{t=1}^{T} Z_t(\pi)$ satisfies, with probability at least $1-\delta$,*

$$\mu(\widehat{\pi}) \lesssim \min_{\pi\in\Pi}\mu(\pi) + \frac{\log(|\Pi|/\delta)}{T}.$$

Proof. For every π, independence across episodes and boundedness give $\mathbb{E}_t\left[Z_t(\pi)\right] = \mu(\pi)$ and $\mathbb{E}_t\left[Z_t(\pi)^2\right] \le \mu(\pi)$. With $\lambda = 1/2$ and $\epsilon = \delta/(2|\Pi|)$, (67) and (68) hold simultaneously for all $\pi \in \Pi$ with probability at least $1-\delta$. On this event,

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{T} Z_t(\pi) - \mu(\pi)\right| &\le \frac{1}{2T}\sum_{t=1}^{T}\mathbb{E}_t\left[Z_t(\pi)^2\right] + \frac{2\log(2|\Pi|/\delta)}{T} && \text{(Equations (67) and (68))}\\
&\le \frac{1}{2}\mu(\pi) + \frac{2\log(2|\Pi|/\delta)}{T}. && (Z_t(\pi)^2 \le Z_t(\pi)) \qquad (69)
\end{aligned}$$

For $\pi_\mu \in \arg\min_{\pi\in\Pi}\mu(\pi)$,

$$\begin{aligned}
\frac{1}{2}\mu(\widehat{\pi}) &\le \frac{1}{T}\sum_{t=1}^{T} Z_t(\widehat{\pi}) + \frac{2\log(2|\Pi|/\delta)}{T} && \text{(Equation (69))}\\
&\le \frac{1}{T}\sum_{t=1}^{T} Z_t(\pi_\mu) + \frac{2\log(2|\Pi|/\delta)}{T} && \text{(Definition of } \widehat{\pi})\\
&\le \frac{3}{2}\mu(\pi_\mu) + \frac{4\log(2|\Pi|/\delta)}{T}. && \text{(Equation (69))}
\end{aligned}$$

For $|\Pi| \ge 2$, $\log(2|\Pi|/\delta) \lesssim \log(|\Pi|/\delta)$. For $|\Pi| = 1$, $\mu(\widehat{\pi}) = \min_{\pi\in\Pi}\mu(\pi)$. □

Lemma D.6 (Adaptive equality tests)**.** *Let Θ be uniform on $[k]$ and $n \ge 0$ be an integer. A possibly randomized learner obtains information about Θ only through at most n tests. Each test selects $j \in [k]$ using previous answers and reveals $\mathbf{1}\{j = \Theta\}$. Every final estimate $\widehat{\Theta} \in [k]$ satisfies*

$$\mathbb{P}\{\widehat{\Theta} \ne \Theta\} \ge 1 - \frac{n+1}{k}.$$

Proof. The claim is trivial for $n \ge k-1$. For a deterministic learner, let $j_1, \dots, j_q$, $q \le n$, and b be its tests and final estimate when every answer is zero. Then

$$\begin{aligned}
\Theta \notin \{j_1,\dots,j_q\} \quad &\Longrightarrow \quad \widehat{\Theta} = b, && \text{(Every test returns zero)}\\
\{\widehat{\Theta} = \Theta\} &\subseteq \{\Theta \in \{j_1,\dots,j_q,b\}\},\\
\mathbb{P}\{\widehat{\Theta} = \Theta\} &\le \frac{|\{j_1,\dots,j_q,b\}|}{k} \le \frac{q+1}{k} \le \frac{n+1}{k}. && \text{(Uniform } \Theta)
\end{aligned}$$

For a randomized learner, the same bound holds conditional on its internal random choices and hence also after averaging. □