When Does On-Policy Interaction Help? Representational Tradeoffs in Value-Based Imitation Learning

Luca Viano* luca.viano@ai.ethz.ch
Antoine Moulin* antoine.moulin@berkeley.edu
Audrey Huang audreyh5@illinois.edu
Volkan Cevher volkan.cevher@epfl.ch
Philip Amortila p.amortila@berkeley.edu
Dylan J. Foster dylanfoster@microsoft.com

Abstract

Imitation learning (IL)—training an agent to replicate expert behavior from demonstrations—underpins applications from robotics to language model training. Standard approaches such as Behavior Cloning (`BC`) are known to suffer from compounding errors and performance plateaus, particularly when the learner cannot perfectly represent the expert's policy (as is typical, e.g., in distillation). Two interventions are widely understood empirically to improve performance: querying the expert *interactively* along the learner's own trajectories, and using value function estimation en route to generating a policy.

We investigate the nature of these improvements and their potentially surprising interplay. Our main finding is that expert interaction relaxes the *representational* demands on the learner: one only needs a model capable of realizing the expert's value function, bypassing the (often stricter) requirement of realizing the expert's policy itself. Concretely, we introduce `OVI`, an interactive on-policy IL algorithm that is statistically and computationally efficient whenever the learner can represent the expert's value function. We complement this with a negative result showing that interaction is necessary. Namely, without stronger assumptions beyond expert-value realizability alone, any offline IL algorithm must scale with the complexity of the expert policy class. Our findings bear out empirically. `OVI` outperforms offline policy-based (`BC`), interactive policy-based (`DAgger`), and offline value-based IL methods, with the largest gains when the learner network is substantially less expressive than the expert's.

1 Introduction

Training a model to replicate the behavior of a more capable expert—a process known as *imitation learning* (IL)—has become a central paradigm in modern AI. In language modeling, distillation trains a smaller language model (LM) to reproduce the outputs of a larger one (Agarwal et al., 2024; Lu and Thinking Machines Lab, 2025); in robotics and autonomous navigation, agents learn complex manipulation skills directly from expert demonstrations (Osa et al., 2018; Pomerleau, 1988). Predominant approaches such as Behavior Cloning (`BC`, Pomerleau, 1991) often suffer from slow convergence, training instability, and performance plateaus even given ample offline expert data (de Haan et al., 2019; Block et al., 2024; Laskey et al., 2017; Hu et al., 2025; Spencer et al., 2021). These shortcomings are commonly attributed to two interrelated phenomena: error propagation along the horizon (Ross and Bagnell, 2010; Ross et al., 2011), and the inability of the learner to faithfully

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

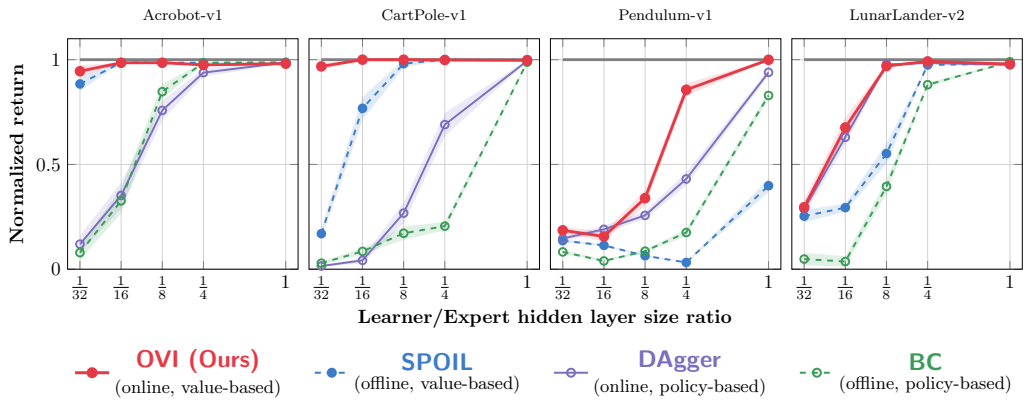


Figure 1: **`OVI` achieves higher return with smaller learner networks.** Average normalized return (y-axis, 1 and 0 correspond to the expert's and to a bad policy's returns, respectively) as a function of the learner-to-expert network-width ratio (x-axis), over 50 seeds and 10 expert trajectories for offline methods or 10 rounds of expert queries along learner trajectories for interactive methods. The expert network has width 64. We compare with offline and interactive policy-based methods (`BC`; Pomerleau, 1988; `DAgger`; Ross et al., 2011), and the offline value-based method `SPOIL` (Moulin et al., 2025a). **When the learner shrinks, `OVI` outperforms each of the following methods alone: (1) offline value-based and (2) interactive policy-based—thereby supporting the theory** that predicts that representational gains can emerge only from expert interaction and value-based design jointly (see Section 3.3).

capture the expert's policy (Espinosa-Dice et al., 2025; Rohatgi et al., 2025). Two interventions are commonly observed empirically to mitigate each of these problems in turn.

The first intervention is *on-policy interaction*, i.e., the act of querying the expert along the learner's own trajectories. *Interactive* (or *on-policy*) IL is understood to avoid error amplification by learning to correct mistakes on-policy. This benefit has classically been formalized through improved horizon dependence (Ross et al., 2011; Rajaraman et al., 2021a). However, recent work by Foster et al. (2024) establishes that when the learner can accurately represent the expert policy, the apparent gap between `BC` and interactive IL disappears, making the precise role of interaction less clear than it once seemed.

A second, parallel intervention is *value-based* IL, where an estimated *value or reward function* is used to assist policy learning (Garg et al., 2021; Swamy et al., 2021, 2022a), subsuming most inverse reinforcement learning (IRL) and apprenticeship learning methods (Abbeel and Ng, 2004; Syed and Schapire, 2007; Ho and Ermon, 2016). Rather than fitting the expert's full distribution over actions, including potentially arbitrary or exogenous choices with equal value, value-based IL seeks to recover only enough information to match the expert's return. This raises the hope that a value-based learner can succeed with a weaker representational burden, e.g., needing only to represent the expected value of each action even when representing the expert policy itself is too demanding. While previous works have sought to develop a theoretical understanding of value-based IL (Abbeel and Ng, 2004; Syed and Schapire, 2007; Joshi et al., 2026; Moulin et al., 2025a), existing methods either do not address the sequential setting or require structural assumptions on the environment; as a result, our understanding of the benefits of value-based IL remains incomplete.

In this work, we investigate how value functions and interaction help in IL. Our results reveal a potentially surprising interplay: **interaction relaxes representational demands on the learner.**

Leveraging on-policy imitation, one only needs a model capable of realizing the expert's value function, bypassing the (often stricter) requirement of realizing the expert's policy itself. Furthermore, this is not possible in offline IL.

Our contributions are detailed below.

Interaction permits efficient value-based IL (Section 3). We introduce `OVI` (Algorithm 1), the first algorithm that is statistically and computationally oracle-efficient under only Q^{π_E}*-realizability*—a natural representational assumption requiring only that the learner can represent the expert's value function—*provided that the learner can interactively query the expert on-policy.* `OVI` queries the expert along the learner's own trajectories and leverages a saddle-point formulation of IL to identify

a useful value function, building on prior work (Moulin et al., 2025a; Swamy et al., 2021). We complement this with a study on the representational advantages of Q^{π_E}-realizability (Section 3.2): we show in several natural problem classes that representing the expert's value function can be significantly easier than representing its policy, and furthermore that the weaker assumption of *reward realizability* is in general insufficient for (offline or interactive) IL.

Interactive value-based IL empirically enables learning with fewer parameters (Section 3.3). We validate these findings empirically in Gymnasium (Towers et al., 2025); see Figure 1. Consistent with our theory, `OVI` outperforms both interactive policy-based methods (`DAgger`; Ross et al., 2011) and offline value-based methods (`SPOIL`; Moulin et al., 2025a), especially as the learner's network shrinks relative to the expert's—a regime in which policy realizability is likely to be increasingly violated. The code for these experiments is available at `https://github.com/lviano/ovi`.

Interaction is necessary for value-based IL (Section 4). We prove that without conditions stronger than Q^{π_E}-realizability, all offline IL algorithms given only an offline dataset of expert trajectories necessarily scale with the number of states in the environment or the complexity of the expert policy class (Theorem 4.1). We further establish that an important subclass of value-based learners (including popular empirical algorithms) can fail to compete with the expert policy in the offline setting even in the limit of infinite data (Theorem G.1). Intuitively, the hardness arises because, unlike in traditional policy-based IL (e.g., via `BC` or `DAgger`), value-based learners match the expert's return without matching its state visitation distribution (cf. Figure 2); our results indicate that this can be drastically more efficient under interaction but fails to overcome well-studied distribution shift challenges (Ross et al., 2011; Rajaraman et al., 2020; Foster et al., 2024) present in offline IL.

Computational-representational tradeoffs in chain-of-thought learning (Section 5). Lastly, we provide a complementary perspective on the benefits of interaction in IL by showing that `OVI` can achieve an exponential computational improvement over existing reward-based approaches (Joshi et al., 2026) when specialized to chain-of-thought learning. Prior theoretical approaches for value-based IL in chain-of-thought learning extract an imitating policy from a learned outcome reward function, but require intractable exponential-weights-type updates and response-level sampling. We show that, leveraging interaction and value function realizability, `OVI` uses value functions to decompose the search over complete responses into a series of tractable token-level policy updates. This gives new theoretical understanding and motivation for the successful use of *on-policy* updates (Agarwal et al., 2024; Gu et al., 2024; Yang et al., 2025) and *process reward models* (Lightman et al., 2024; Uesato et al., 2023) in LM training.

When does interaction help in IL? Going beyond the classical understanding of improved horizon dependence, our results reveal additional (and perhaps stronger) **representational** benefits of interaction. Taken together, they provide a comprehensive picture of the interplay between interaction and representation in value-based IL.

2 Problem Setting: Value-Based Imitation Learning

We first provide general definitions of Markov decision processes (MDPs) and the general IL problem, and then formally describe our problem setting of *value-based imitation learning*.

2.1 Imitation Learning

Episodic MDPs. An *episodic* MDP is defined as a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, r, \nu_0)$, where $\mathcal{X}$ is a (potentially large) finite state space, $\mathcal{A}$ is a finite action space with $A \geq 2$ actions, $H \in \mathbb{N}$ is the horizon, $P = (P_h \colon \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X}))_{h=1}^H$ is the transition kernel, $r = (r_h \colon \mathcal{X} \times \mathcal{A} \to [0,1])_{h=1}^H$ is the reward function, and $\nu_0 \in \Delta(\mathcal{X})$ is the initial state distribution. A nonstationary Markov policy (henceforth simply *policy*) is a sequence of *decision rules* $\pi = (\pi_h \colon \mathcal{X} \to \Delta(\mathcal{A}))_{h=1}^H$, it induces a distribution $\mathbb{P}^\pi$ over trajectories $(\boldsymbol{x}_h, \boldsymbol{a}_h, \boldsymbol{r}_h)_{h=1}^H$ via the following interaction protocol: an initial state $\boldsymbol{x}_1 \sim \nu_0$ is drawn, and then for each $h = 1, \dots, H$, we have $\boldsymbol{a}_h \sim \pi_h(\cdot \mid \boldsymbol{x}_h)$, $\boldsymbol{r}_h = r_h(\boldsymbol{x}_h, \boldsymbol{a}_h)$, and $\boldsymbol{x}_{h+1} \sim P_h(\cdot \mid \boldsymbol{x}_h, \boldsymbol{a}_h)$. We denote by $\mathbb{E}^\pi$ the corresponding expectation. The expected return of a policy π is given by $J^\pi = \mathbb{E}^\pi\left[\sum_{h=1}^H \boldsymbol{r}_h\right]$, and the value functions are defined for any state-action pair (x, a) by $V_h^\pi(x) = \mathbb{E}^\pi\left[\sum_{h'=h}^H \boldsymbol{r}_{h'} \mid \boldsymbol{x}_h = x\right]$, and $Q_h^\pi(x,a) = \mathbb{E}^\pi\left[\sum_{h'=h}^H \boldsymbol{r}_{h'} \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]$. They satisfy the Bellman equations: $Q_h^\pi(x,a) = r_h(x,a) + \mathbb{E}_{\boldsymbol{x}' \sim P_h(\cdot|x,a)}[V_{h+1}^\pi(\boldsymbol{x}')]$, and $V_h^\pi(x) = \mathbb{E}_{\boldsymbol{a} \sim \pi_h(\cdot|x)}[Q_h^\pi(x, \boldsymbol{a})]$, where $V_{H+1}^\pi = 0$. We define the occupancy measures via $d_h^\pi(x,a) =$

$\mathbb{P}^\pi[\boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$ and $d_h^\pi(x) = \mathbb{P}^\pi[\boldsymbol{x}_h = x]$. To disentangle the effects of reward scaling from the horizon H, we assume rewards are normalized such that for some $Q_{\mathtt{max}} > 0$ and any feasible trajectory, $\sum_{h=1}^H \boldsymbol{r}_h \in [0, Q_{\mathtt{max}}]$ (Jiang and Agarwal, 2018; Wang et al., 2020; Zhang et al., 2021).[2] For any policy π and function $f\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we sometimes use the notation $f(x, \pi) = \sum_a \pi(a \,|\, x) f(x, a)$.

Offline and interactive imitation learning. In the IL setting that we consider, there is an unknown *expert policy* π_{E} and an unknown MDP $\mathcal{M}$, and the goal is to learn a policy that performs as well as π_{E}, as measured by expected return. Formally, for error tolerance $\varepsilon > 0$ and failure probability $\delta > 0$, the learner's objective is to output a policy π_{out} such that, with probability at least $1 - \delta$,

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon. \tag{1}$$

Central to our paper is the distinction between offline and interactive IL. In *offline IL*, the learner is given a dataset of pre-collected independent expert trajectories $\mathcal{D}_{\mathsf{E}} = \{((\boldsymbol{x}^i_{\mathsf{E},h}, \boldsymbol{a}^i_{\mathsf{E},h}))_{h=1}^H\}_{i=1}^{n_{\mathsf{E}}}$, where $(\boldsymbol{x}^i_{\mathsf{E},h}, \boldsymbol{a}^i_{\mathsf{E},h}) \sim d_h^{\pi_{\mathsf{E}}}$ and n_{E} is the number of samples. The learner cannot interact with the MDP or the expert further. In contrast, in *interactive IL* the learner can query the expert π_{E} for actions while rolling out in the unknown MDP $\mathcal{M}$—importantly, without observing the rewards. For the purpose of our paper, we formalize *on-policy interaction* as follows: in each round of interaction, the learner selects a policy π to generate a state trajectory $(\boldsymbol{x}_h)_{h=1}^H$ where $\boldsymbol{x}_h \sim d_h^\pi$ and might query expert actions at states along the trajectory, i.e., $\boldsymbol{a}_{\mathsf{E},h} \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_h)$.

2.2 Value-Based Imitation Learning

Our focus is on *value-based imitation learning*, by which we mean methods that derive policies from learned value or reward functions. Classical *policy-based* methods (e.g., `BC`, `DAgger`) aim to imitate the full conditional action distribution of the expert π_{E} and to match its performance as a consequence. By contrast, *value-based* learners aim to recover only enough information to match the expert's return, which may be a simpler task whenever there is redundancy in the expert's trajectory distribution (Example 3.1) or relevant structure in the dynamics that can be exploited (Example 3.2).

In particular, our goal is to develop value-based IL algorithms that succeed with *only realizability of the expert's value function*, a natural representational assumption for value-based IL. Stated in learning-theoretic terms, the learner has access to a value function class $\mathcal{Q} \subseteq \{(Q_h\colon \mathcal{X} \times \mathcal{A} \to [0, Q_{\mathtt{max}}])_{h=1}^H\}$ that can be used to recover a good policy, and we assume only the following on its expressivity.

Assumption 2.1 ($Q^{\pi_{\mathsf{E}}}$-realizability)**.** *The class $\mathcal{Q}$ contains the expert's value function, i.e., $Q^{\pi_{\mathsf{E}}} \in \mathcal{Q}$.*

By contrast, policy-based methods, such as `BC` and `DAgger`, commonly require policy realizability assumptions (Foster et al., 2024), which render them inapplicable in this setting. In Section 3.2, we highlight several natural settings in which *representing the expert's value function is drastically less stringent than realizing the expert's policy*. Accordingly, in Section 3.1 we will develop algorithms for interactive IL whose statistical rates depend only on the statistical capacity of the value function class $\mathcal{Q}$ (e.g., $\log(|\mathcal{Q}|)$ for finite function classes) and not the policy class used by the learner.

Value-based realizability assumptions comparable to Assumption 2.1 have been analyzed in previous works, although they further required assumptions about the dynamics of the MDP. In particular, Joshi et al. only consider the $H = 1$ ("contextual bandit") setting, where Assumption 2.1 simplifies to realizability of the reward function, and Moulin et al. require closure-type assumptions on $\mathcal{Q}$, i.e., that it can realize Q^π for any policy π generated by the learner.[3] In Section 3.2, we further establish that mere reward realizability is insufficient for general environments when $H > 1$ (Theorem H.1), indicating that $Q^{\pi_{\mathsf{E}}}$-realizability is a natural minimal assumption for value-based IL.

Comparing interactive value-based IL to related settings. Finally, we clarify the differences between our value-based IL setting and related IL settings considered in the literature. All of the settings mentioned below allow the learner to roll out in the MDP, but use protocols for interacting with the expert that differ from ours. In *inverse RL*, the learner interacts with the MDP or with known MDP dynamics, but does not query the expert interactively (Abbeel and Ng, 2004; Syed and Schapire, 2007; Ziebart et al., 2008). In *value-feedback* algorithms such as `AggreVaTe` (Ross and Bagnell, 2014; Sun et al., 2017), the learner interacts with the MDP and can observe the expert value

[2] In general (the *dense reward* setting), $Q_{\mathtt{max}}$ can be as large as H, but in some settings (e.g., in LM reasoning tasks with binary outcome-level rewards) we can have $Q_{\mathtt{max}} = \mathcal{O}(1)$ (the *sparse reward* setting).

[3] Such "completeness" assumptions implicitly place restrictions on the dynamics (Chen and Jiang, 2019; Foster et al., 2022; Jiang and Xie, 2025), and by contrast with Assumption 2.1 cannot be satisfied by simply taking a more expressive $\mathcal{Q}$ class.

$Q^{\pi_{\mathtt{E}}}(\boldsymbol{x}_h, \boldsymbol{a}_h)$, whereas it is unclear how to directly estimate these values in our setting given that the learner does not observe rewards. Lastly, in a setting we call *RL with expert advice*, the learner both interacts with the MDP and queries the expert (or a reference policy that covers it well) interactively, but additionally observes rewards in the MDP (Amortila et al., 2022; Tiapkin et al., 2024; Foster et al., 2025). See Appendix B for a comprehensive discussion of related work.

3 Interaction Enables $Q^{\pi_{\mathtt{E}}}$-Realizable Value-Based Imitation

The main result of this section is `OVI`, the first provably efficient algorithm for value-based IL under $Q^{\pi_{\mathtt{E}}}$-realizability. In Section 3.1, we describe the algorithm and its guarantees. We then show in Section 3.2 that $Q^{\pi_{\mathtt{E}}}$-realizability can be weaker than policy realizability, while reward realizability alone is insufficient. Finally, in Section 3.3 we present empirical evidence that `OVI` outperforms both policy-based methods (`DAgger`; Ross et al., 2011) and offline value-based methods (`SPOIL`; Moulin et al., 2025a) in regimes where policy realizability is likely to be violated.

3.1 `OVI`: Value-Based Imitation Learning With Only $Q^{\pi_{\mathtt{E}}}$-Realizability

Algorithm design. `OVI` treats the problem of minimizing the gap in Eq. (1) as a min-max game between a π-player and a Q-player. The starting point is the performance difference lemma (Lemma E.9)

$$J^{\pi_{\mathtt{E}}} - J^{\pi} = \sum_{h=1}^{H} \mathcal{L}_h^{d^{\pi}}(\pi_h, Q_h^{\pi_{\mathtt{E}}}),$$

$$\text{where, given } d = (d_h \in \Delta(\mathcal{X}))_{h=1}^{H},\ \mathcal{L}_h^{d}(\pi, Q) := \underbrace{\sum_{x \in \mathcal{X}} d_h(x)}_{\texttt{(I)}} \underbrace{\sum_{a \in \mathcal{A}} (\pi_{\mathtt{E},h}(a \,|\, x) - \pi(a \,|\, x)) Q(x, a)}_{\texttt{(II)}}.$$

This form is not enough yet to derive an algorithm because $Q^{\pi_{\mathtt{E}}}$ is unknown. However, assuming $Q^{\pi_{\mathtt{E}}} \in \mathcal{Q}$, we bound the suboptimality by taking the supremum over $\mathcal{Q}_h := \{Q_h : \exists Q' \in \mathcal{Q}, Q'_h = Q_h\}$ for every h on the right hand side, $J^{\pi_{\mathtt{E}}} - J^{\pi} \le \sum_{h=1}^{H} \sup_{Q_h \in \mathcal{Q}_h} \mathcal{L}_h^{d^{\pi}}(\pi_h, Q_h)$. Accordingly, what remains is to find an approximate saddle point of the expected advantage function, i.e., a policy $\pi_{\mathtt{out}}$ such that, for all $h \in [H]$, $\sup_{Q_h \in \mathcal{Q}_h} \mathcal{L}_h^{d^{\pi_{\mathtt{out}}}}(\pi_{\mathtt{out},h}, Q_h) \lesssim \varepsilon / H$.

Algorithmically, the main point of departure from prior value-based IL algorithms is the *layer-wise* learning of a good policy, in the vein of the `Forward` algorithm (Ross and Bagnell, 2010). To understand this design choice, we recall that the approach employed by past work (Moulin et al., 2025a) is to compute a sequence $(\pi^k, Q^k)_{k=1}^{K}$ where π^k performs online mirror ascent (Beck and Teboulle, 2003) and Q^k performs a best response to (an empirical estimate of) the loss function $\mathcal{L}$. This technique does not seem to apply in the $Q^{\pi_{\mathtt{E}}}$-realizability setting, as the sequence π^k is also dictating the current sampling distribution d under which the error is measured (see `(I)` above), preventing us from simultaneously forming an empirical estimate of $\mathcal{L}$ and controlling the error in `(II)` via online learning. This is addressed in `OVI` by learning $\pi_{\mathtt{out}}$ one layer at a time (note the outer loop over $h = 1, \dots, H$ in line 2 of Algorithm 1), as $d_h^{\pi_{\mathtt{out}}}(x)$ only depends on $(\pi_{\mathtt{out},1}, \dots, \pi_{\mathtt{out},h-1})$ and we can thus fix the sampling distribution for stage h before optimizing for $\pi_{\mathtt{out},h}$.

We view `OVI` as a principled algorithmic realization of the "On-Q" moment matching template introduced by Swamy et al. (2021). The pseudocode is given in Algorithm 1. The main sample complexity guarantee is stated below. It establishes that `OVI` only requires a number of expert queries that scales with the statistical capacity of the value function class $\mathcal{Q}$, as represented by its log-cardinality $\log|\mathcal{Q}|$.

Theorem 3.1 (Sample complexity of `OVI`)**.** *Let Assumption 2.1 hold for a finite class $\mathcal{Q}$. Then, for any $\varepsilon, \delta \in (0,1)$, `OVI` (Algorithm 1) with parameters $n_{\mathtt{E}} = \widetilde{\mathcal{O}}(H^4 Q_{\mathtt{max}}^4 \log(A) \log(|\mathcal{Q}|/\delta)\varepsilon^{-4})$, $\eta = (\log(A)/(K Q_{\mathtt{max}}^2))^{1/2}$, and $K = \mathcal{O}(H^2 Q_{\mathtt{max}}^2 \log(A)\varepsilon^{-2})$, outputs a policy $\pi_{\mathtt{out}}$ such that with probability at least $1-\delta$, $J^{\pi_{\mathtt{E}}} - J^{\pi_{\mathtt{out}}} \le \varepsilon$ after*

$$\widetilde{\mathcal{O}}\left(\frac{H^5 Q_{\mathtt{max}}^4 \log(A) \log(|\mathcal{Q}|/\delta)}{\varepsilon^4}\right) \qquad \textit{expert queries.}$$

Note that the total number of learner interactions with the environment obeys the same bound with an extra factor of H. Up to the usual polynomial factors in H, $1/\varepsilon$, and $\log(\delta^{-1})$, the main

Algorithm 1 `OVI`: **O**n-Policy **V**alue-Based **I**mitation Learning

1: **input:** Learning rate η, iterations K, expert queries per stage n_{E}.
2: **for** $h = 1, \ldots, H$ **do**
3: Create dataset at stage h: for $i \in [n_{\mathsf{E}}]$, sample $\boldsymbol{x}_h^i \sim d_h^{\pi_{\mathsf{out}}}$, and query $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^i)$.
4: Initialize $\pi_h^1 = \mathrm{Unif}(\mathcal{A})$.
5: **for** $k = 1, \ldots, K$ **do**
6: Set $Q_h^k \in \arg\max_{Q_h \in \mathcal{Q}_h} \sum_{i=1}^{n_{\mathsf{E}}} (Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i) - Q_h(\boldsymbol{x}_h^i, \pi_h^k))$.
7: Set $\pi_h^{k+1}(a \mid x) \propto \pi_h^k(a \mid x) e^{\eta Q_h^k(x,a)}$, for any state-action pair (x, a).
8: Create the output policy at layer h: $\pi_{\mathsf{out},h} = \frac{1}{K} \sum_{k=1}^{K} \pi_h^k$.

appealing feature of `OVI` is that its sample complexity is controlled by the statistical complexity of *only the value function class* $\mathcal{Q}$, through its log-cardinality. In particular, the guarantee imposes no statistical assumption on the learner's policy class. We show in Section 3.2 that, in many natural settings, a function class $\mathcal{Q}$ satisfying $Q^{\pi_{\mathsf{E}}}$-realizability can be drastically smaller than a policy class Π satisfying π_{E}-realizability, while it is the latter that governs the sample complexity of policy-based methods such as `BC` or `DAgger`. We note that the dependence on the problem horizon is larger than that obtained by policy-based methods such as `BC` or `DAgger` (Foster et al., 2024); it is an interesting question whether this can be improved further or if this is fundamental to value-based IL.

Theorem F.1 extends the guarantee of Theorem 3.1 to infinite classes via covering numbers. The result also shows that, when $\mathcal{Q}$ is convex, the expert-query complexity improves to $\widetilde{\mathcal{O}}(\varepsilon^{-2})$.

Computationally, `OVI` only requires access to a standard linear maximization oracle for the Q updates (Line 6 in Algorithm 1) and standard softmax policy updates for the π player (Line 7 in Algorithm 1). One drawback is that the layer-wise updates learn a potentially nonstationary policy and introduce computation and memory requirements that scale linearly with the horizon. However, in our experiments (Section 3.3), we find that a stationary approximation of `OVI` performs well. We refer to Appendix D for experimental details.

Improved rates for nonconvex classes. The rate attained by `OVI` for nonconvex classes $\mathcal{Q}$ has a sub-optimal $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ dependence. In Appendix I, we introduce the variant `OVI+` (Algorithm 3) that reuses the expert labels across iterations while collecting fresh learner rollouts to estimate the learner term in the Q-player's objective. It improves the number of expert queries to $\widetilde{\mathcal{O}}(H^3 Q_{\mathsf{max}}^2 \log(|\mathcal{Q}|/\delta)\varepsilon^{-2})$, but the number of learner rollouts remains of order $\widetilde{\mathcal{O}}(H^5 Q_{\mathsf{max}}^4 \log(A) \log(|\mathcal{Q}|/\delta)\varepsilon^{-4})$. A promising direction for reducing the remaining learner-rollout cost is recursive variance reduction of the learner term, which we leave to future work (see Appendix I). Both the number of expert queries and learner rollouts can be improved to an $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ rate via a variant of `OVI`, named Q-`OVI` (Algorithm 4), where the roles of the value functions and the policies are reversed, i.e., the Q player performs online learning and the π player performs a best response (Theorem J.1). While statistically more efficient, this relies on explicitly discretizing the value-function class $\mathcal{Q}$, making the method computationally unattractive for large or continuous classes such as neural networks. It is an interesting question to obtain a rate of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for both expert queries and learner rollouts in a computationally efficient manner. Another advantage of Q-`OVI` is that it can be used to compute a stationary expert-matching policy (see Appendix J.3), whereas `OVI` learns a nonstationary policy. Whether a computationally efficient value-based IL algorithm can compute such a stationary policy remains open.

3.2 Representational Advantages of $Q^{\pi_{\mathsf{E}}}$-Realizability

`OVI` requires only access to a function class that realizes the expert's Q-function, $Q^{\pi_{\mathsf{E}}}$. We show that, in many natural settings, $Q^{\pi_{\mathsf{E}}}$-realizability is substantially weaker than the assumptions used by prior policy- and value-based approaches. We also establish that reward realizability alone is insufficient.

Values can be easier to represent than policies. Classical IL algorithms, such as `BC`, `DAgger`, and their derivatives, are *policy-based*: they operate by fitting the expert's entire conditional distribution of actions, and their performance relies on a *policy realizability* assumption (Foster et al., 2024).

Assumption 3.1 (Policy realizability)**.** *The policy class* Π *contains the expert policy, i.e.,* $\pi_{\mathsf{E}} \in \Pi$.

This assumption is natural if the goal is to reproduce the expert's trajectory distribution *in a reward-free manner*. But the objective in IL is weaker: to match the expert's return. For this objective, it may be unnecessary and wasteful to attempt to fit the specific details of the expert distribution (e.g., how it breaks ties among equally good actions, or how it responds to features that are irrelevant for value). The examples below formalize this intuition.

Joshi et al. (2026) consider the special case of our setting in which $H = 1$, and show that value-based IL can be strictly weaker than policy-based IL (note that when $H = 1$, $Q^{\pi_{\mathsf{E}}}$-realizability reduces to reward realizability). They show that when the expert is optimal and realized by the policy class Π, the class can be used to construct a realizable reward-function class $\mathcal{R}$[4] satisfying $|\mathcal{R}| \leq |\Pi|$, simply by taking $\mathcal{R} = \{r : \exists \pi \in \Pi, r(x,a) = \mathbb{1}\{a \in \mathsf{supp}(\pi(\cdot \mid x))\}\}$. By contrast, they show that the existence of a small realizable reward class $\mathcal{R}$ cannot be used to construct a small realizable policy class Π under which BC will succeed.[5] Our first example extends this to the multi-step setting.

Example 3.1 (Learning to answer from correct demonstrations, with $H > 1$)**.** *Given a finite class Π that realizes an optimal expert π_{E}, we show in Appendix C.1 that it is possible to construct a class $\mathcal{Q}$ such that $|\mathcal{Q}| \leq |\Pi|$ and $\mathcal{Q}$ realizes a function $\widetilde{Q}^{\pi_{\mathsf{E}}}$ which assigns maximal value to actions within the expert support, defined by $\widetilde{Q}^{\pi_{\mathsf{E}}}(x,a) = Q_{\mathsf{max}}\, \mathbb{1}\{a \in \mathsf{supp}(\pi_{\mathsf{E}}(\cdot \mid x))\}$. OVI still succeeds, even if we cannot guarantee that $Q^{\pi_{\mathsf{E}}} \in \mathcal{Q}$ (see Appendix C.1.1). Conversely, as above, when given a small $\mathcal{Q}$ class satisfying $Q^{\pi_{\mathsf{E}}}$-realizability, the learner cannot construct a small policy class Π satisfying $\pi_{\mathsf{E}} \in \Pi$. We come back to this example in Section 5.* ◁

We note that this reduction is enabled by the relaxed $Q^{\pi_{\mathsf{E}}}$-realizability condition detailed in Appendix C.1.1. In the $H > 1$ setting, the closest comparable value-based IL work is the SPOIL algorithm of Moulin et al., which requires that $\mathcal{Q}$ can realize Q^{π} for all policies π generated by the learner. In that setting, it is highly unclear whether there exists a similar reduction from a small policy class Π to a small function class $\mathcal{Q}$ that satisfies this closure-type condition. Thus, under prior value expressivity conditions, value-based IL may not be representationally weaker than policy-based IL.

Our second example shows that, even with suboptimal experts, the same separation can arise from structure in the dynamics, as the value function may ignore high-dimensional information needed to reproduce the expert's exact actions.

Example 3.2 (Structure in the dynamics)**.** *In Appendix C.2, we study an Exogenous Block MDP (Efroni et al., 2022; Mhammedi et al., 2024; Amortila et al., 2024b), where the dynamics and reward depend only on low-dimensional* endogenous *state components, while the observations contain high-dimensional exogenous noise. We construct a family of MDPs and expert policies and show that any expert-agnostic policy class Π satisfying Assumption 3.1 must be exponentially large. In contrast, the same two-dimensional class $\mathcal{Q}$ satisfies Assumption 2.1 for the entire family.* ◁

Although the two assumptions need not be comparable in general, we expect value realizability to be weaker than policy realizability in a wide range of analogous settings. This is empirically supported by our experiments in Section 3.3.

Reward realizability is insufficient. Even with known rewards, matching an optimal expert to expected suboptimality ε can require $\Omega(|\mathcal{X}|/\varepsilon)$ rounds of interaction. The reward can identify desirable states without revealing which actions reach them. See Appendix H for the result and proof.

All in all, **we view $Q^{\pi_{\mathsf{E}}}$-realizability as a minimal representational assumption for efficient learning**: it is more informative than reward realizability alone, which is insufficient, but less demanding than other assumptions (e.g., policy realizability, realizability of Q^{π} for infinitely-many policies π, or realizability of rewards and transitions). We next provide empirical evidence that the minimality of $Q^{\pi_{\mathsf{E}}}$-realizability translates to more effective IL when using learners with limited expressivity.

3.3 Experiments

We test whether the representational benefits suggested by Section 3.2 translate into practical performance when the learner has limited capacity. In each of four Gymnasium environments (Towers et al., 2025), the expert is an RL-trained policy represented by a network with two hidden layers of 64 neurons each, while the learners use the same depth but have widths in $\{2, 4, 8, 16, 32, 64\}$. We

[4] Up to a reward shaping term that does not alter the optimal policy.

[5] While the learner knows that policies in Π must be supported on reward-maximizing actions, the set of such policies is prohibitively large.

use network width as a proxy for representational capacity: as the learner's network becomes smaller relative to that of the expert, representing the expert's action distribution should become increasingly difficult, and policy-based methods such as BC and DAgger may therefore degrade relative to OVI. The code for these experiments is available at https://github.com/lviano/ovi.

We evaluate OVI, SPOIL, BC, and DAgger across the above range of widths. We use 10 expert trajectories for the offline methods, 10 learner-trajectory labeling rounds for the interactive methods, and 50 seeds (see Appendix D for more experimental details). Figure 1 shows that when the learner also has 64 neurons per layer, BC and DAgger perform well in all four environments. Their performance degrades, however, as the learner network shrinks. In contrast, OVI dominates across all learner sizes and attains good performance even at the smallest width. SPOIL generally deteriorates faster than OVI, especially in CartPole-v1 and Acrobot-v1, consistent with the possibility that Q^{π_E}-realizability is easier to satisfy in these experiments than the closure-type realizability conditions required by SPOIL. We use a stationary approximation to OVI in these experiments; see Algorithm 2 for pseudocode.

Finally, we compare DAgger and OVI in a synthetic Q^{π_E}-realizable MDP to highlight that, despite both IL algorithms being interactive, OVI implements a different imitation mechanism than DAgger.

(Figure 2) DAgger vs. OVI: a different imitation mechanism. Although also an interactive IL method, DAgger is surprisingly suboptimal in this environment, which is an extended version of the simple Q^{π_E}-realizable instance used in Theorem G.1 (defined for Theorem 4.1 and illustrated in Figure 6). As displayed in Figure 2, the reward suboptimality, i.e., $J^{\pi_E} - J^{\pi_{\text{out}}}$, of OVI (solid red line) converges significantly faster than the suboptimality of DAgger (solid purple line). The advantage of OVI stems from the fact that it directly matches the expert's return, while DAgger, being a policy-based method, attempts to match the expert's actions almost everywhere, which is a harder problem in the MDP under consideration because there are $\mathcal{O}(2^{|\mathcal{X}|})$ possible optimal policies from which the expert can choose. In comparison, OVI acts approximately[6] greedily with respect to its learned Q^{π_E}, which is sufficient to achieve the optimal return with far fewer samples but at the cost of potentially learning a policy different from the expert's. To illustrate the difference between the learning mechanisms of these two algorithms, we plot in Figure 2 their trajectory-level TV distance to the expert, $\mathcal{D}_{\text{TV}}(\mathbb{P}^{\pi_E}, \mathbb{P}^{\pi_{\text{out}}})$, which decreases under DAgger (dashed purple line) but remains constant for OVI (dashed red line).

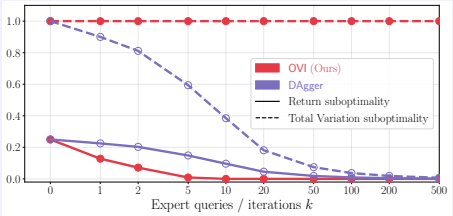


Figure 2: DAgger vs OVI in the MDP illustrated in Figure 6. Solid and dashed lines show suboptimality, $J^{\pi_E} - J^{\pi_{\text{out}}}$, and total variation between trajectory distributions, $\mathcal{D}_{\text{TV}}(\mathbb{P}^{\pi_E}, \mathbb{P}^{\pi_{\text{out}}})$, respectively.

4 Interaction Is Necessary for Q^{π_E}-Realizable Value-Based Imitation

In Section 3, we showed that interaction enables OVI to succeed under only Q^{π_E}-realizability, a weaker representational requirement than those imposed by comparable algorithms. Here, we prove that interaction is in fact *necessary*: under Q^{π_E}-realizability alone, *no offline IL method can learn efficiently*. We formally define an offline IL algorithm below.

Definition 4.1 (Offline IL Algorithm)**.** *An offline IL algorithm* Alg *is a (potentially randomized) mapping from an expert dataset $\mathcal{D}_E$ to a nonstationary policy $(\pi_{\text{out},h} \colon \mathcal{X} \to \Delta(\mathcal{A}))_{h=1}^{H}$.*

Our main negative result (Theorem 4.1), stated formally below, establishes that there exists a large family of MDPs (and corresponding optimal expert policies) with a small value function class exhibiting Q^{π_E}-realizability for each MDP in the class, but for which the sample complexity of any offline IL algorithm must scale with either the size of the state space or the number of MDPs in the class. For the statement below, we recall that, for an MDP $\mathcal{M}$, we write $J^{\pi}_{\mathcal{M}}$ for the return of a policy π, and $\pi^{\star}_{\mathcal{M}}$ for one of its optimal deterministic policies.

Theorem 4.1 (Main lower bound for offline IL)**.** *For any $X \in \mathbb{N}^{\star}$ and any $\varepsilon \in (0, \frac{1}{8}]$, there exists a family $\mathcal{F}$ of MDPs with: i) stochastic initial states but deterministic rewards and transitions, ii)*

[6] The OVI iterates become more and more greedy with respect to Q^{π_E} as K increases.

horizon $H = 2$, iii) state space size $|\mathcal{X}| = \mathcal{O}(X)$, and iv) family size $\log(|\mathcal{F}|) = \mathcal{O}(X)$, as well as a value function class $\mathcal{Q}$ with $|\mathcal{Q}| = 2$ such that for every $\mathcal{M} \in \mathcal{F}$, Assumption 2.1 is satisfied for expert policy $\pi_{\mathsf{E}} = \pi^\star_{\mathcal{M}}$. Let Alg *be any offline IL algorithm (Definition 4.1). Then, there exists an MDP $\mathcal{M} \in \mathcal{F}$ such that* Alg *needs $n_{\mathsf{E}} \geq \Omega(\frac{X}{\varepsilon})$ trajectories sampled offline with expert policy $\pi^\star_{\mathcal{M}}$, to output a policy π_{out} such that $\mathbb{E}[J^{\pi^\star_{\mathcal{M}}}_{\mathcal{M}} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}}] \leq \varepsilon$.*[7]

In other words, any offline imitation learner needs at least $\Omega(\min\{|\mathcal{X}|, \log(|\mathcal{F}|)\}/\varepsilon)$ samples to obtain a near-optimal policy, even when there exists a small value function class satisfying $Q^{\pi_{\mathsf{E}}}$-realizability for every MDP in the problem class.[8] We note that similar "tabular" lower bound rates have appeared for offline IL (Rajaraman et al., 2020, 2021b; Foster et al., 2024), though with a focus on establishing the optimal dependence on horizon. By comparison, our lower bound shows that the worst-case offline IL rate continues to depend on the number of states even when there exists a small value function class for the problem class, and even for $H = 2$.

Moreover, we show in Theorem G.1 that an even stronger lower bound can be proven against a large class of algorithms that defines the output policy as a function of learned Q-values.

Our lower bound (Theorem 4.1) show that interaction is necessary in general for learning with only $Q^{\pi_{\mathsf{E}}}$-realizability. However, we point out that the expert induces a distribution over the state space with sufficient coverage, the lower bound no longer holds. In particular, in Theorem K.1, we prove sample complexity guarantees for SPOIL under $Q^{\pi_{\mathsf{E}}}$-realizability, expert optimality and a standard coverage condition (Munos, 2003; Antos et al., 2008; Chen and Jiang, 2019; Xie et al., 2023; Amortila et al., 2024c; Jiang and Xie, 2025).

Finally, we show that additional representational benefits can arise when the dataset includes trajectories from *both* the learner and the expert, with the expert being queried at all such states (a technique commonly used in LM distillation; Agarwal et al., 2024; Li et al., 2026a). In particular, we show that a variant of OVI that mixes learner and expert trajectories succeeds whenever $\mathcal{Q}$ satisfies either $Q^{\pi_{\mathsf{E}}}$-realizability (Assumption 2.1) or the SPOIL-like closure condition, which requires $Q^\pi \in \mathcal{Q}$ for every policy π generated by the learner, without knowing in advance which condition holds (Theorem L.1).

5 Chain-of-Thought Learning: Computational-Representational Tradeoffs

The preceding sections showed that interaction relaxes the representational requirements for the learner. Here, we show how OVI can be instantiated in the setting of LM reasoning to achieve exponential improvements in computational efficiency over existing approaches (Joshi et al., 2026), provided that the expert's value can be realized, thereby providing a complementary perspective on the value of interaction in IL.

Reasoning tasks such as mathematics and code generation have become a central focus of modern LM research (Wei et al., 2022; Li et al., 2022). In these domains, objective evaluators (such as unit and integration tests for coding and formal verifiers for math) are often available and can be used as a "ground truth" reward function. To solve these problems, LMs are trained to generate a sequence of tokens, called a *chain of thought* (CoT), before outputting the final answer. To acquire CoT reasoning, LMs can be trained via IL, commonly via next-token prediction (Ouyang et al., 2022; Cobbe et al., 2021) on expert reasoning traces consisting of a prompt, the expert's complete chain-of-thought, and the correct final answer.

CoT learning (Malach, 2024; Joshi et al., 2025) has been formalized as follows. The LM is modeled as an autoregressive policy that, given a prompt $\boldsymbol{x} \in \mathcal{X}$, samples a response $\boldsymbol{y} \in \mathcal{Y}$ via $\boldsymbol{y} \sim \pi(\cdot \mid \boldsymbol{x})$, where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{H-1}, \boldsymbol{y}_H)$ is a sequence of H tokens, belonging to a token vocabulary Σ, generated autoregressively as $\boldsymbol{y}_h \sim \pi(\cdot \mid \boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{h-1})$, $h = 1, \ldots, H$, and the last token $\boldsymbol{y}_H$ is the final answer. This interaction defines a *token-level MDP* with horizon H, where states are the partial generations so far, $\boldsymbol{x}_h = (\boldsymbol{x}, \boldsymbol{y}_{1:h-1}) \in \mathcal{X} \times \Sigma^{h-1}$, the action is the next token $\boldsymbol{y}_h \in \Sigma$, and the next state is $\boldsymbol{x}_{h+1} = (\boldsymbol{x}, \boldsymbol{y}_{1:h})$. The learner is given access to a dataset $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\}$, where the prompt $\boldsymbol{x}$ is sampled from a given prompt distribution, and the response is generated by the expert, $\boldsymbol{y} \sim \pi_{\mathsf{E}}(\cdot \mid \boldsymbol{x})$. The reward function is a binary *outcome verifier* defined as $r \colon \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ so that the expected re-

[7] The expectation is over the training data and the internal randomness of Alg.

[8] By contrast, OVI succeeds on the MDPs defined in the proof of Theorem 4.1 with one expert query outside the support of its state distribution.

turn of a policy, $J^\pi = \mathbb{E}^\pi[r(\boldsymbol{x}, \boldsymbol{y})]$, is equivalent to its answer accuracy. The goal of IL then translates to the problem of producing a policy π_{out} whose final-answer accuracy competes with the expert's.

Non-interactive reward-based IL is intractable computationally. A natural simplification to CoT learning, exploiting that the dynamics are deterministic and known, is to treat the CoT problem as an $H = 1$ problem (i.e., a contextual bandit), where the prompt $\boldsymbol{x}$ is the context and the response $\boldsymbol{y} \in \Sigma^H$ is an action. Under a reward (or outcome verifier) realizability assumption (Assumption H.1), Joshi et al. (2026) estimate a reward $\widehat{r}$ from the demonstrations $\mathcal{D}$ using an exponential-weights-type procedure which, at each iteration, must solve $\arg\max_{y \in \Sigma^H} \widehat{r}(\boldsymbol{x}, y)$ to extract a response for a given prompt $\boldsymbol{x}$.[9] Their procedure presents two computational barriers: (i) the reward optimization is over the full response space $\mathcal{Y} = \Sigma^H$, which is computationally intractable for long horizons, and (ii) exponential weight updates over the reward class is infeasible as it requires enumerating over that class.

Computational benefits of value-based interactive IL. To exploit the inherently sequential nature of CoT learning, we can apply `OVI` directly to the token-level MDP generated by the interaction of the prompt distribution, the LM, and the outcome verifier r. At each step h, `OVI` learns a value function that naturally induces a softmax token sampling distribution (Line 7 of Algorithm 1). This decomposes the previously intractable search over the response space Σ^H into H local search steps. Moreover, each stage of the problem is solved by `OVI` using simple linear maximization oracles for the value function search (Line 6 of Algorithm 1) and softmax updates for the policy (Line 7 of Algorithm 1), which can be implemented at scale. Indeed, our approach suggests a loss function for the value network which can be minimized via backpropagation coupled with modern optimizers.

It is worth noting that the computational benefits of `OVI` over sequence-level methods like Joshi et al. (2026) come at the cost of a stronger representational requirement (realizing the expert's value function for each step rather than simply the reward of the final answer) as well as interactive query access to the expert. Both of these additional assumptions have been widely shown to improve performance; the first can naturally be interpreted as requiring an architecture capable of expressing a process verifier (or process reward model; Lightman et al., 2024; Uesato et al., 2023), as opposed to simply an outcome verifier, and the second corresponds to the setting of on-policy distillation (Agarwal et al., 2024; Gu et al., 2024; Yang et al., 2025) which has been widely employed in language modeling. Our results therefore provide a novel theoretical justification for the use of process reward models and on-policy interaction in LM training. Whether these gains translate to practical value-based IL algorithms for language models is a very interesting empirical question.

6 Conclusion

What is the value of interaction in IL? We investigated how interaction and value function estimation help in IL, revealing a perhaps surprising interplay: interaction relaxes the representational demands on the learner, allowing it to succeed while realizing only the expert's value function, as opposed to the expert's full policy or the value functions of all learner-generated policies. We formalized this through `OVI`, an interactive value-based IL algorithm that is statistically and computationally efficient under $Q^{\pi_{\mathsf{E}}}$-realizability alone, and a complementary lower bound establishing that interaction is necessary for IL under this minimal assumption. Specializing to chain-of-thought reasoning, we showed that these representational benefits can translate into exponential computational gains over non-interactive approaches (Joshi et al., 2026). These findings suggest that the benefits of interaction in IL extend well beyond the classical understanding of improved horizon dependence.

Several interesting future directions remain open. On the technical side, it would first be valuable to obtain an $\mathcal{O}(\varepsilon^{-2})$ rate for both expert queries and learner rollouts when $\mathcal{Q}$ is nonconvex with a computationally efficient algorithm. Second, while under π_{E}-realizability nearly horizon-free bounds are possible (Foster et al., 2024), whether the same is achievable for value-based IL is an interesting open question. Finally, it would be interesting to study value-based IL in the presence of value function misspecification; we expect that the relevant forms of misspecification differ from their counterparts in value-based RL (Chen and Jiang, 2019; Du et al., 2020; Amortila et al., 2023, 2024a; Maran and Szepesvári, 2025). On the empirical side, it would be valuable to understand whether `OVI` (or value-based IL methods more broadly) can be effective in language model distillation, where expert policy realizability is unlikely to hold.

[9] We note that our reward-based lower bound Theorem H.1 does not apply in CoT learning since the dynamics in the token-level MDP are known.

Acknowledgments

We thank Gergely Neu and Nan Jiang for interesting discussions about this project. LV was supported by the Swiss Data Science Center under fellowship number P22_03 and this work was done while LV was at EPFL. AM received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180). PA gratefully acknowledges the support of DARPA through award No. HR00112520022. Part of this research was performed while AM and AH were visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation (Grant No. DMS-2425650). Research by VC was sponsored by the Army Research Office under Grant Number W911NF-24-1-0048 and funded by the Swiss National Science Foundation (SNSF) under grant number 2000-1-240094.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004. URL https://icml.cc/Conferences/2004/proceedings/papers/335.pdf.

Pieter Abbeel, Dmitri Dolgov, Andrew Y. Ng, and Sebastian Thrun. Apprenticeship learning for motion planning with application to parking lot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008. URL https://ai.stanford.edu/~ang/papers/iros08-ApprenticeshipLearningParkingLotNavigation.pdf.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, 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://arxiv.org/abs/2306.13649.

Philip Amortila, Nan Jiang, Dhruv Madeka, and Dean Foster. A few expert queries suffices for sample-efficient RL with resets and linear value approximation. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=d19Dsqtw421.

Philip Amortila, Nan Jiang, and Csaba Szepesvári. The optimal approximation factors in misspecified off-policy value function estimation. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/amortila23a.html.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *Conference on Learning Theory*, 2024a. URL https://arxiv.org/abs/2401.12216.

Philip Amortila, Dylan J Foster, Nan Jiang, Akshay Krishnamurthy, and Zakaria Mhammedi. Reinforcement learning under latent dynamics: Toward statistical and algorithmic modularity. In *Advances in Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=qf2uZAdy1N.

Philip Amortila, Dylan J Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability. In *International Conference on Machine Learning*, 2024c. URL https://arxiv.org/abs/2403.06571.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008. URL https://link.springer.com/article/10.1007/s10994-007-5038-2.

Maria-Florina Balcan, Avrim Blum, Kiriaki Fragkia, Zhiyuan Li, and Dravyansh Sharma. Verify to amplify: Improving reasoning via learned chain-of-thought verification. *arXiv preprint arXiv:2603.03538*, 2026. URL https://arxiv.org/abs/2603.03538.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 2003. URL https://www.tau.ac.il/~becka/3.pdf.

Adam Block, Dylan J Foster, Akshay Krishnamurthy, Max Simchowitz, and Cyril Zhang. Butterfly effects of SGD noise: Error amplification in behavior cloning and autoregression. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CgPs04l9TO.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2011. URL https://proceedings.mlr.press/v15/boularias11a.html.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012. URL https://arxiv.org/abs/1204.5721.

Locke Cai, Max Ryabinin, and Ivan Provilkov. Escaping the verifier: Learning to reason via demonstrations. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pS1khvoxHT.

Haoqun Cao and Tengyang Xie. Understanding behavior cloning with action quantization. *arXiv preprint arXiv:2603.20538*, 2026. URL https://arxiv.org/abs/2603.20538.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

Ching-An Cheng, Xinyan Yan, Nolan Wagener, and Byron Boots. Fast policy learning through imitation and reinforcement. *arXiv preprint arXiv:1805.10413*, 2018. URL https://arxiv.org/abs/1805.10413.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/cheng22b.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Chris Cundy and Stefano Ermon. SequenceMatch: Imitation learning for autoregressive sequence modelling with backtracking. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FJWT0692hw.

Robert Dadashi, Léonard Hussenot, Damien Vincent, Sertan Girgin, Anton Raichuk, Matthieu Geist, and Olivier Pietquin. Continuous control with action quantization from demonstrations. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/dadashi22a.html.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/947018640bf36a2bb609d3557a285329-Paper.pdf.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.03016.

Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, 2022. URL https://proceedings.mlr.press/v178/efroni22a.html.

Nicolas Espinosa-Dice, Sanjiban Choudhury, Wen Sun, and Gokul Swamy. Efficient imitation under misspecification. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fn36V5qsCw.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, 2016. URL https://proceedings.mlr.press/v48/finn16.html.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, 2022. URL https://arxiv.org/abs/2111.10919.

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*, 2024. URL https://openreview.net/forum?id=8KPyJm4gt5.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? The computational role of the base model in exploration. In *Conference on Learning Theory*, 2025. URL https://proceedings.mlr.press/v291/foster25a.html.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkHywl-A-.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, Matthieu Geist, and Stefano Ermon. IQ-Learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.12142. Note: Read arXiv version for correct version and complete author list.

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://openreview.net/forum?id=5h0qf7IBZZ.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf.

Jonathan Ho, Jayesh Gupta, and Stefano Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning*, 2016. URL https://proceedings.mlr.press/v48/ho16.html.

Ronald A Howard. *Dynamic programming and Markov processes*. John Wiley, 1960.

Zheyuan Hu, Robyn Wu, Naveen Enock, Jasmine Jia ni Li, Riya Kadakia, Zackory Erickson, and Aviral Kumar. RaC: Robot learning for long-horizon tasks by scaling recovery and correction. In *Workshop on Making Sense of Data in Robotics: Composition, Curation, and Interpretability at Scale at CoRL 2025*, 2025. URL https://openreview.net/forum?id=y8wskVS7BV.

Chen Jia. Adversarial moment-matching distillation of large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0VeSCjRDBy.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference on Learning Theory*, 2018. URL https://proceedings.mlr.press/v75/jiang18a.html.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2025. URL https://arxiv.org/abs/2510.04088.

Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. In *Conference on Learning Theory*, 2025. URL https://arxiv.org/abs/2503.07932.

Nirmit Joshi, Gene Li, Siddharth Bhandari, Shiva Kasiviswanathan, Cong Ma, and Nathan Srebro. Learning to answer from correct demonstrations. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=69fIHgLjyH.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002. URL https://homes.cs.washington.edu/~sham/papers/rl/aoarl.pdf.

Rahul Kidambi, Jonathan Chang, and Wen Sun. MobILE: Model-based imitation learning from observation alone. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2102.10769.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hk4fpoA5Km.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyg-JC4FDr.

Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. DART: Noise injection for robust imitation learning. In *Conference on Robot Learning*, 2017. URL https://proceedings.mlr.press/v78/laskey17a.html.

Changhao Li, Rushi Qiang, Jiawei Huang, Chenxiao Gao, Chao Zhang, Niao He, and Bo Dai. Revisiting DAgger in the era of LLM-agents. *arXiv preprint arXiv:2605.12913*, 2026a. URL https://arxiv.org/abs/2605.12913.

Jiaxiang Li, Siliang Zeng, Hoi To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the SFT data: Reward learning from human demonstration improves SFT for LLM alignment. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=orxQccN8Fm.

Shangzhe Li, Dongruo Zhou, and Weitong Zhang. Near-optimal second-order guarantees for model-based adversarial imitation learning. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=PD8wnZOV1J.

Yichen Li and Chicheng Zhang. On efficient online imitation learning via classification. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=h2imPVlCCyN.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with AlphaCode. *Science*, 2022. URL https://arxiv.org/abs/2203.07814.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Sihan Liu, Gaurav Mahajan, Daniel Kane, Shachar Lovett, Gellért Weisz, and Csaba Szepesvári. Exponential hardness of reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2023. URL https://proceedings.mlr.press/v195/liu23b.html.

Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/liu22u.html.

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

Eran Malach. Auto-regressive next-token predictors are universal learners. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2309.06979.

Davide Maran and Csaba Szepesvári. Beyond least squares: Uniform approximation and the hidden cost of misspecification. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/f5ecc67e8b2f4941137bcbde901ac6be-Paper-Conference.pdf.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7sACcaOmGi.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015. URL https://www.nature.com/articles/nature14236.

Antoine Moulin, Gergely Neu, and Luca Viano. Inverse Q-learning done right: Offline imitation learning in Q^π-realizable MDPs. In *Advances in Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=tvEE9KQcLi.

Antoine Moulin, Gergely Neu, and Luca Viano. Optimistically optimistic exploration for provably efficient infinite-horizon reinforcement and imitation learning. In *Conference on Learning Theory*, 2025b. URL https://proceedings.mlr.press/v291/moulin25a.html.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003. URL https://cdn.aaai.org/ICML/2003/ICML03-074.pdf.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000. URL https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

Francesco Orabona. Online learning: A modern introduction using convex optimization. *arXiv preprint arXiv:1912.13213*, 2026. URL https://arxiv.org/abs/1912.13213v10.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018. URL https://arxiv.org/abs/1811.06711.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Dean A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.

Dean A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 1991. URL https://ieeexplore.ieee.org/document/6796843.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1e7875cf32d306989d80c14308f3a099-Abstract.html.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems*, 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/09dbc1177211571ef3e1ca961cc39363-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*, 2021b. URL https://arxiv.org/abs/2102.12948.

N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning*, 2006. URL https://dl.acm.org/doi/10.1145/1143844.1143936.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. SQIL: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*, 2019. URL https://arxiv.org/abs/1905.11108.

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. In *Conference on Learning Theory*, 2025. URL https://proceedings.mlr.press/v291/rohatgi25a.html.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014. URL https://arxiv.org/abs/1406.5979.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Conference on Computational Learning Theory*, 1998. URL https://dl.acm.org/doi/pdf/10.1145/279943.279964.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20798.

Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual RL: Unification and new methods for reinforcement and imitation learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xt9Bu66rqv.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. *arXiv preprint arXiv:2102.02872*, 2021. URL https://arxiv.org/abs/2102.02872.

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.

Hao Sun and Mihaela van der Schaar. Inverse reinforcement learning meets large language model post-training: Basics, advances, and opportunities. *arXiv preprint arXiv:2507.13158*, 2025. URL https://arxiv.org/abs/2507.13158.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/sun17d.html.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, 2019a. URL https://arxiv.org/abs/1811.08540.

Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International Conference on Machine Learning*, 2019b. URL https://proceedings.mlr.press/v97/sun19b.html.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/swamy21a.html.

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Sequence model imitation learning with unobserved contexts. In *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=3nbKUphLBg5.

Gokul Swamy, Nived Rajaraman, Matt Peng, Sanjiban Choudhury, J Bagnell, Steven Z Wu, Jiantao Jiao, and Kannan Ramchandran. Minimax optimal online imitation learning via replay estimation. In *Advances in Neural Information Processing Systems*, 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2e809adc337594e0fee330a64acbb982-Abstract-Conference.html.

Umar Syed and Robert E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/ca3ec598002d2e7662e2ef4bdd58278b-Abstract.html.

Umar Syed, Michael Bowling, and Robert E. Schapire. Apprenticeship learning using linear programming. In *International Conference on Machine Learning*, 2008. URL https://icml.cc/Conferences/2008/papers/645.pdf.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. Demonstration-regularized RL. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.17303.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *International Joint Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1805.01954.

Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://arxiv.org/abs/1905.13566.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, et al. Gymnasium: A standard interface for RL environments. *arXiv preprint arXiv:2407.17032*, 2025. URL https://arxiv.org/abs/2407.17032.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Yamamoto Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-based and outcome-based feedback, 2023. URL https://openreview.net/forum?id=MND1kmmNy0O.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9988f2c8e07c1f98af7ba9ca31ccae0b-Abstract-Conference.html.

Luca Viano, Stratis Skoulakis, and Volkan Cevher. Imitation learning in discounted linear MDPs without exploration assumptions. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DChQpB4AJy.

Stefano Viel, Luca Viano, and Volkan Cevher. IL-SOAR: Imitation learning with soft optimistic actor critic. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=NNr8DHb0L7.

Ruosong Wang, Simon S Du, Lin Yang, and Sham Kakade. Is long horizon RL more difficult than short horizon RL? In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.00527.

Joe Watson, Sandy Huang, and Nicolas Heess. Coherent soft imitation learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2f0435cffef91068ced08d7c7d8e643e-Paper-Conference.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Gellert Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, and Csaba Szepesvari. On query-efficient planning in MDPs under linear realizability of the optimal state-value function. In *Conference on Learning Theory*, 2021. URL https://proceedings.mlr.press/v134/weisz21a.html.

Runzhe Wu, Yiding Chen, Gokul Swamy, Kianté Brantley, and Wen Sun. Diffusing states and matching scores: A new framework for imitation learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kWRKNDU6uN.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015. URL https://arxiv.org/abs/1507.04888.

Markus Wulfmeier, Michael Bloesch, Nino Vieillard, Arun Ahuja, Jorg Bornschein, Sandy Huang, Artem Sokolov, Matt Barnes, Guillaume Desjardins, Alex Bewley, et al. Imitating language via scalable inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5d2eScRiRC.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.04157.

Tian Xu, Ziniu Li, Yang Yu, and Zhi-Quan Luo. Provably efficient adversarial imitation learning with unknown transitions. In *Conference on Uncertainty in Artificial Intelligence*, 2023. URL https://proceedings.mlr.press/v216/xu23c.html.

Tian Xu, Zhilong Zhang, Ruishuo Chen, Yihao Sun, and Yang Yu. Provably and practically efficient adversarial imitation learning with general function approximation. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7YdafFbhxL.

Tian Xu, Chenyang Wang, Xiaochen Zhai, Ziniu Li, Yi-Chen Li, and Yang Yu. Non-adversarial imitation learning provably free of compounding errors: The value flow mechanism. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=1UUJrgYr20.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, 2021. URL https://proceedings.mlr.press/v134/zhang21b.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

Contents of Appendix

A Notation Reference **21**

B Additional Related Work **22**

B.1 Theoretical Guarantees for Imitation Learning 22

B.2 Empirical Value-Based Imitation Learning 25

C Value-Based Imitation Learning versus Policy-Based: Motivating Examples **27**

C.1 Constructing $\mathcal{Q}$ from Π Reduces Complexity 27

C.2 Q^{π_E}-Realizability Can Be Strictly Weaker Than π_E-Realizability under Exogenous Noise 30

D Experiment Details **35**

D.1 Experiments on the Lower Bound Instance 35

D.2 Implementation with Neural Networks 35

D.3 Additional Details and Comments on the Gym Experiments 35

E Technical Tools **38**

E.1 Concentration 38

E.2 Optimization 41

E.3 Performance Difference Lemma 44

I Proofs of Main Results **45**

F Proofs from Section 3 **45**

F.1 Proof of Theorem 3.1 (Sample Complexity Guarantee for OVI) 46

G Proofs from Section 4 (Q^{π_E}-Realizability Is Insufficient to Learn Offline) **50**

G.1 Examples of offline algorithms using VI policies 50

G.2 Proof of Theorem G.1 (Lower Bound Against Offline Algorithms that Output VI Policies) 51

G.3 Proof of Theorem 4.1 (Lower Bound Against Any Offline IL Algorithm) 53

II Additional Results **56**

H Reward realizability is insufficient **56**

H.1 Proof of Theorem H.1 (Reward Realizability Is Insufficient) 57

I OVI+: Improved Bound on the Number of Expert Queries **60**

I.1 Design of OVI+ 60

I.2 Guarantees for OVI+ 60

J Q-OVI: An Inefficient Algorithm with Improved Statistical Guarantees **64**

J.1 Proof of Lemma J.1 (Suboptimality Gap Decomposition for Q-OVI) 65

J.2 Proof of Theorem J.1 (Sample Complexity Guarantee for Q-OVI) 66

J.3 Learning a Stationary Policy with Q-OVI 67

K Breaking the Offline Lower Bound with Coverage **70**

K.1 OVI with Arbitrary Sampling Distributions 70

K.2 Proof of Theorem K.1 (Offline IL with Coverage and Optimal Expert) 73

L On the Benefits of Mixing Expert and Learner Trajectories **76**

A Notation Reference

The table summarizes notation. The last column points to definitions or key uses.

Table 1: Notation reference.

Notation	Meaning	Reference
MDP and imitation learning setting		
$\mathcal{M}=(\mathcal{X},\mathcal{A},H,P,r,\nu_0)$	State/action spaces $\mathcal{X},\mathcal{A}$ ($A=\lvert\mathcal{A}\rvert$), horizon H, transitions P, rewards r, initial law ν_0.	Sec. 2
$\pi=(\pi_h)_{h=1}^H$	Nonstationary Markov policy. $\pi_h\colon\mathcal{X}\to\Delta(\mathcal{A})$ is its decision rule at stage h.	Sec. 2
$\mathbb{P}^\pi,\ \mathbb{E}^\pi,\ J^\pi$	Trajectory law, expectation, and expected return under π.	Sec. 2
$V_h^\pi,\ Q_h^\pi$	State and state-action value functions at stage h.	Sec. 2
$d_h^\pi(x),\ d_h^\pi(x,a)$	State and state-action occupancy measures at stage h.	Sec. 2
$\pi_{\mathsf{E}},\ \pi_{\mathsf{E},h},\pi_{\mathsf{out}},\ \pi_{\mathsf{out},h}$	Expert and output policies, with their decision rules at stage h.	Sec. 2
$\varepsilon,\ \delta$	Target suboptimality and failure probability.	Eq. (1)
$\mathcal{D}_{\mathsf{E}},\ \boldsymbol{x}_{\mathsf{E},h}^i,\boldsymbol{x}_h^i,\ \boldsymbol{a}_{\mathsf{E},h}^i$	Offline expert dataset/states, query states and expert actions.	Secs. 2.1 & 3.1
n_{E}	Offline trajectories, queried samples per stage, episodes in Def. H.1.	Secs. 2.1 & 3.1
$f(x,\pi)$	Action average $\sum_a\pi(a\,\lvert\,x)f(x,a)$.	Sec. 2
Value classes and realizability		
$\mathcal{Q},\ \mathcal{Q}_h$	Value-function class and its projection at stage h.	Secs. 2.2 & 3.1
Q_{max}	Upper bound on total trajectory reward and candidate values.	Sec. 2
$\Pi,\ \mathcal{R}$	Policy/reward classes. Realizability means $\pi_{\mathsf{E}}\in\Pi$, $r\in\mathcal{R}$.	Asps. 3.1 & H.1
$\widetilde{Q}^{\pi_{\mathsf{E}}}$	Witness for relaxed $Q^{\pi_{\mathsf{E}}}$-realizability.	Asp. C.1
$Q^{\Pi_{\mathcal{Q}}}$-realizability	$Q^\pi\in\mathcal{Q}$ for every policy in the specified class $\Pi_{\mathcal{Q}}$.	Asp. L.1
Objectives, empirical estimates, and complexity		
$\mathcal{L}_h^d(p,Q)$	Expert-minus-learner value under stagewise state law d_h.	Sec. 3.1
$\mathcal{L}_h^\pi$	Shorthand for $\mathcal{L}_h^{d^\pi}$.	App. F
$\widehat{\mathcal{L}}_h(p,Q),\ \Delta(\pi)$	Empirical version of $\mathcal{L}_h^{\pi_{\mathsf{out}}}$ and its uniform estimation error at π.	Lem. F.1
$\mathfrak{R}_h^\pi(x)$	Cumulative policy regret against the expert at stage h, state x.	Lem. F.1
$\mathcal{N}_\varepsilon(S,\mathrm{d}),\ \mathcal{C}_\varepsilon(S,\mathrm{d})$	Covering number and an ε-cover. Omitted metric means $\lVert\cdot\rVert_\infty$.	Def. E.1
$\mathcal{N}_\varepsilon^{\mathsf{max}}(\mathcal{Q})$	Largest stagewise covering number: $\max_h\mathcal{N}_\varepsilon(\mathcal{Q}_h,\lVert\cdot\rVert_\infty)$.	Thm. F.1
$\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q}\times\Pi_{\mathcal{Q}},\rho)$	Product-class covering number $\max_h\mathcal{N}_r(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)$.	Thm. L.1
$\lVert\cdot\rVert_{\infty,1},\ \rho$	Sup-state ℓ_1 policy norm, scaled value/policy product metric.	Eqs. (3) & (4)
$\Pi_{\mathcal{Q}},\ \Pi_{\mathcal{Q},h}$	Softmax policies and stagewise projection, K/rate-dependent, with stagewise hybrids in App. L.	Lem. E.3 & Asp. L.1
$r_\varepsilon,\ r_\varepsilon^{(1)},\ r_\varepsilon^{(\infty)}$	Result-dependent covering radii in sample-complexity bounds.	Apps. F, I & K
Algorithmic and appendix-specific notation		
$K,\ \eta,\ \eta_k,\ \pi_h^k,\ Q_h^k$	Stage iteration count, learning rates, and policy/value iterates.	Algs. 1 & 6
$Q^{1:K},\ \mathrm{LC},\ f$	Value functions, linear combination, and link of a VI policy.	Def. G.1
$\Pi_{\mathrm{Alg}},\ C_\infty,\ C_1$	Output-policy class, L_∞ and L_1 expert-coverage bounds.	Asps. K.1 & K.2
$d=(d_h)_{h=1}^H,$ $\widehat{\mathcal{L}}_h^d,\ \Delta_d(\pi)$	Stagewise sampling laws, possibly history-dependent. Empirical objective and estimation error.	Thm. K.2, proof
$n_{\mathsf{L}},\ \boldsymbol{x}_h^{jk}$	`OVI+` learner batch size per stage/iteration and fresh states.	Alg. 3
$\widehat{\mathcal{L}}_{h,k}(p,Q_h),\ \Delta_{h,k}(p)$	`OVI+` empirical objective, error uniform in Q_h at fixed h,k.	Thm. I.1, proof
α	Expert-state weight in the learner/expert occupancy mixture.	Alg. 7
$\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q)$	Sampled loss in the Q-`OVI` exponential-weights update.	App. J
$w_h^k,\ \mathfrak{R}_h^w,\ \widehat{Q}_h^{\pi_{\mathsf{E}}}$	Cover weights, regret, and a closest cover element to $Q_h^{\pi_{\mathsf{E}}}$.	Lem. J.1

B Additional Related Work

We provide an overview of closely related work. In Appendix B.1, we discuss theoretical results available for value-based IL methods, and in Appendix B.2 we cover value-based IL algorithms that focus on empirical performance. We use "value-based IL" broadly to include reward-learning, inverse reinforcement learning, apprenticeship learning, moment matching, and Q-function-based methods, since all use learned reward/value information to derive a policy rather than directly fitting the expert action distribution.

B.1 Theoretical Guarantees for Imitation Learning

As mentioned in the main text, our work aims to establish expert-sample complexity bounds when expert policy realizability need not hold. While we are the first to study this problem under Q^{π_E}-realizability alone, prior work has examined several alternative assumptions that we discuss below.

After behavior cloning was proposed in the seminal work of Pomerleau (1988, 1991), a parallel line of work formalized *inverse reinforcement learning* (IRL; Russell, 1998; Ng and Russell, 2000), whose goal is to infer a reward function that explains the expert's behavior, typically among many rewards for which the expert is optimal. Closely related work on *apprenticeship learning* (Abbeel and Ng, 2004; Abbeel et al., 2008; Syed and Schapire, 2007; Syed et al., 2008; Ziebart et al., 2008) uses this reward-based perspective to learn a policy that matches the expert's performance.

Ross and Bagnell (2014) introduced `AggreVaTe`, an interactive no-regret reduction that augments expert action queries with value information. That is, when the learner visits a state-action pair (x, a), $Q^{\pi_E}(x, a)$ is observed. This oracle is referred to as *value feedback* and is stronger than our setting, where the learner can query expert actions but does not observe rewards or expert values. Their framework also yields practical algorithms compatible with neural-network-based function approximation; see Sun et al. (2017); Cheng et al. (2018). Value-based feedback, together with access to an interactive expert, is known to allow for better horizon dependence in error propagation (Ross and Bagnell, 2014; Sun et al., 2017), which can lead to representational benefits in policy-based IL (Foster et al., 2024).

The closest conceptual predecessor to `OVI` is the moment-matching framework of Swamy et al. (2021), especially their "On-Q" template and the associated `DAeQuIL` algorithm. This framework also includes reward and "Off-Q" moment-matching templates, and analyzes imitation gaps through *moment recoverability* μ: matching action-value moments to precision ε yields a suboptimality bound of order $\mu H \varepsilon$, improving on the classical H^2 error-propagation factor in `BC`-style analyses. Empirically, Swamy et al. (2021) show that `DAeQuIL` can outperform `BC` and `DAgger` in a forest-navigation task, which they attribute to the mode-seeking behavior induced by action-value moment matching. However, despite the better error-propagation properties of interactive algorithms, Foster et al. (2024) showed that no interactive algorithm can uniformly improve upon an offline one across all possible policy classes. Therefore, the error-propagation-style analysis does not seem sufficient to capture the benefits of interaction. Our contribution is complementary: we instantiate the "On-Q" perspective with a computationally oracle-efficient algorithm, provide a sample-complexity analysis under Q^{π_E}-realizability, and show that this representational assumption alone is insufficient in the offline setting.

Next, we review value-based IL further, distinguishing (i) offline algorithms that operate on a fixed dataset of expert demonstrations from (ii) variants that can additionally access the environment and collect trajectories without observing rewards. Then, we discuss the main policy-based IL methods, followed by influential empirical work on value-based IL.

Value-based imitation learning with online MDP access or known dynamics. It is useful to distinguish three closely related access models. First, classical apprenticeship-learning and IRL methods (Abbeel and Ng, 2004; Syed and Schapire, 2007) assumed knowledge of a linear reward class containing the true unknown reward, as well as perfect knowledge of the MDP dynamics. In tabular problems with unknown dynamics, Syed and Schapire (2007) instead estimated the transition kernel only on state-action pairs frequently visited by the expert and sent all other transitions to a pessimistic dead state. They showed that this restricted model suffices to retain a guarantee, at the cost of a slightly worse sample-complexity bound. Also in the tabular setting, Rajaraman et al. (2020) characterize worst-case offline imitation rates with quadratic horizon dependence H^2 and show that,

for deterministic experts, access to the transition model can improve the horizon dependence by at least a $\sqrt{H}$ factor. Rajaraman et al. (2021b) later proved a matching lower bound in that setting.

A second line of work assumes reward-free online access to the MDP: the learner may roll out policies in the environment, but it does not observe the reward and typically obtains expert information through precollected demonstrations, expert occupancies, or expert states or features rather than through value feedback. Xu et al. (2023) removed the known-transition assumption discussed above in tabular adversarial IL and proposed a method achieving the minimax optimal expert-sample complexity by combining expert demonstrations with reward-free environment interaction. Shani et al. (2022) presented an approach based on alternating updates between sequences of rewards and policies. Viano et al. (2022) develop minimax/proximal-point methods for infinite-horizon IL in MDPs with linear, unknown rewards and dynamics under a strong exploratory assumption on the learner's policies, which was later removed by Viano et al. (2024). Moulin et al. (2025b) further develop optimistic exploration tools for infinite-horizon linear MDPs and apply them to IL. In state- or feature-matching variants, such methods can sometimes use expert states or features rather than expert actions, connecting them to imitation from observation alone (Sun et al., 2019b; Torabi et al., 2018, 2019; Kidambi et al., 2021; Viel et al., 2025). Beyond linear function approximation, Xu et al. (2024) extended the guarantees to the setting in which general function approximation is required to approximate the reward and the dynamics of the environment. Finally, recent work also studies instance-dependent and second-order guarantees (i.e., bounds that scale with the variance of the expert and learner value functions): Wu et al. (2025) give first- and second-order bounds in a score-matching framework, while Li et al. (2026b) prove near-optimal second-order guarantees for model-based adversarial IL.

Finally, another related but distinct line gives the learner online reward observations in addition to expert information, making the problem closer to RL with expert advice than to reward-free IL. For example, Tiapkin et al. (2024) studied the setting in which an offline record of state-action pairs collected from an optimal expert can be used to initialize the learner policy in RL. They show that in tabular and linear MDPs, this can achieve better sample complexity than RL without expert data. On the representation side, Amortila et al. (2022) study RL with "one-step" resets, access to an interactive optimal expert, and a class that realizes the optimal state-value function. They provide a computationally efficient algorithm to approximate the optimal policy, showing that expert advice can make otherwise difficult value-realizable settings tractable. This contrasts with planning or RL under value-function realizability alone, where query or computational barriers are known in related settings (Weisz et al., 2021; Liu et al., 2023).

These works are complementary to our setting: they exploit reward observations, known dynamics, reward-free environment interaction without learner-state expert action queries, or expert advice for RL, whereas `OVI` uses only reward-free MDP rollouts together with expert action queries on learner-visited states.

Value-based imitation learning with logged expert data. The closest offline counterpart to our setting is Moulin et al. (2025a). Their algorithm, `SPOIL`, shows that expert-policy realizability can be avoided in offline IL by imposing a closure-type value realizability assumption: the class $\mathcal{Q}$ must realize Q^π for every policy π in the implicit policy class $\Pi_\mathcal{Q}$ that the algorithm may optimize over or output. This assumption is qualitatively different from, and substantially stronger than, requiring only $Q^{\pi_E} \in \mathcal{Q}$, since it asks the value class to be closed under many learner-generated policies rather than only to contain the expert's value function. This distinction is important for interpreting our offline lower bound: the lower bound shows that Q^{π_E}-realizability alone is insufficient for offline value-based IL in general, while the offline guarantees of `SPOIL` rely on additional closure structure that is absent from our main setting. Moreover, we point out that Adversarially Trained Actor Critic (`ATAC`; Cheng et al. (2022)), originally developed for offline RL, recovers `SPOIL` when some hyperparameter (β in their paper) is set to zero, and mirror descent is used as their generic online learning algorithm. Cheng et al. (2022) made explicit that IL algorithms can be recovered from their formulation, but their analysis still relies on expert policy realizability.

A different logged-data setting was studied by Joshi et al. (2026), who model learning from correct demonstrations as offline IL in a contextual bandit, equivalently an MDP with horizon $H = 1$. They assume that the correctness, or reward, model belongs to a finite class, rather than assuming expert-policy realizability, and they design statistically efficient algorithms from offline demonstrations alone. Moreover, they improve the algorithm of Syed and Schapire (2007) when instantiated with

$H = 1$, thereby obtaining faster rates when the expert is optimal. Because $H = 1$, there is no sequential distribution shift and no nontrivial transition dynamics to learn or exploit.

A recent value-based IL method also investigates how value structure can mitigate compounding errors. Xu et al. (2026) show that IQ-Learn-style objectives can reduce to behavior cloning and still suffer compounding errors, and they introduce Bellman constraints as a mechanism for propagating value information from demonstrated states to states not covered by the demonstrations. In our setting, a variant of this idea could be used to filter a given realizable class $\mathcal{Q}$ by removing functions with "temporal" variations greater than one; this is guaranteed to reduce the complexity of $\mathcal{Q}$ while still providing realizability of Q^{π_E}. In practice, this could be implemented by adding a regularization term to the OVI objective that penalizes large temporal variations.

Policy-based imitation learning. Historically, the most common paradigm for tackling IL is to maximize the likelihood of the observed expert actions over a given policy class. In its offline form, BC fits the expert action distribution by supervised learning on expert trajectories, with early BC-style systems including ALVINN (Pomerleau, 1988, 1991). Interactive variants such as DAgger—very popular in practice—and related no-regret reductions instead collect expert labels on states visited by the learner, thereby addressing the covariate shift that appears when an offline cloned policy is deployed on its own trajectory distribution (Ross and Bagnell, 2010; Ross et al., 2011). Theoretical guarantees are typically expressed in terms of the complexity of a policy class Π that contains the expert policy π_E (see, for example, Rajaraman et al., 2021a; Swamy et al., 2022b). Li and Zhang (2022) study this classification-based online IL viewpoint more generally, including nonrealizable settings. More recently, Foster et al. (2024) refined the classical picture by showing that, under suitable policy realizability and log-loss learnability assumptions, offline BC can have much better horizon dependence than suggested by the standard compounding-error analysis.

Follow-up work investigates what happens when the expert policy is not exactly realizable by the learner's policy class. Rohatgi et al. (2025) analyze autoregressive learning and IL under policy misspecification, showing that error amplification reappears in misspecified settings and that next-token or BC-style objectives face computational-statistical tradeoffs. They also propose robustified BC variants whose guarantees quantify the dependence on horizon and misspecification. Espinosa-Dice et al. (2025) introduce GUITAR, an efficient interactive IL algorithm under misspecification that uses a structural completeness condition and environment access initialized from states sampled from the expert occupancy measure, with guarantees controlled by an advantage-class notion of misspecification rather than by statistical divergences.

Most recently, Sriraman et al. (2026) studied a setting in which expert samples may be corrupted: with some probability, the action shown to the learner is drawn from a different policy rather than the expert's. They show that any offline policy-based IL algorithm suffers exponential sample complexity, whereas their interactive algorithm, NAIL, avoids this exponential blow-up by additionally querying the expert.

In a spirit similar to ours, another recent work (Zhang et al., 2026) studies the advantage of on-policy interaction, albeit from a policy-based viewpoint. They show that when the policy class does not realize the expert policy, a coverage coefficient akin to ours is unavoidable for offline imitation. In contrast, online distillation can avoid such dependence if the expert logits are *aligned* with the reward direction. We consider their approach orthogonal to ours: they show under which assumptions existing algorithms are effective, whereas we introduce a setup (Q^{π_E}-realizable MDPs) in which new algorithmic ideas are required to attain the best possible theoretical guarantees. Moreover, Zhang et al. (2026) show that on-policy distillation algorithms can produce students competitive with the expert in terms of accuracy while maintaining very different next-token distributions.

Finally, Balcan et al. (2026) show that a separation between the offline and the interactive setting also occurs when considering the problem of learning a process reward model from a limited number of offline or on-policy calls from an oracle, which can be cast as policy-based imitation learning. In particular, Balcan et al. (2026) show that a process reward model with optimal soundness/completeness tradeoff is possible only given on-policy access to the oracle.

To conclude our review of the theoretical literature in IL, we report a comparison of our results with the bounds obtained and the representational conditions required by other algorithms in Table 2.

Table 2: Comparison with related algorithms, assumptions, and expert-interaction requirements. To state the guarantees of BC, we define a policy class Π such that $\pi_{\mathsf{E}} \in \Pi$. Joshi et al. (2026) study finite reward classes and one-step decision problems. $\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q})$ denotes the largest covering number of the stagewise classes $\mathcal{Q}_h$ in $\|\cdot\|_\infty$ with spacing $r_\varepsilon = \mathcal{O}(\varepsilon^2)$. We define the class $\Pi_{\mathcal{Q}} := \{\pi : \exists m \le K, \forall h, \exists Q_h^1, \dots, Q_h^m \in \mathcal{Q}_h, \pi_h(\cdot \mid x) = \mathrm{softmax}(\eta \sum_{k=1}^m Q_h^k(x, \cdot))\}$ for the algorithm's iteration budget K and learning rate η. $Q^{\Pi_{\mathcal{Q}}}$-realizability is the assumption that $Q^\pi \in \mathcal{Q}$ for any $\pi \in \Pi_{\mathcal{Q}}$.

Algorithm	Representational assumptions	Comp. efficient	Interactive expert	Expert traj./queries
LogLossBC (Foster et al., 2024)	π_{E}-realizability	✓	✗	$\mathcal{O}\left(\frac{H^2 \log\lvert\Pi\rvert}{\varepsilon^2}\right)$
MWAL (Syed and Schapire, 2007)	Reward linear in d features Known transitions	✓	✗	$\widetilde{\mathcal{O}}\left(\frac{H^2 \log(d)}{\varepsilon^2}\right)$
MUWU (Joshi et al., 2026)	Reward realizability with class $\mathcal{R}$ Contextual bandits ($H = 1$)	✗	✗	$\widetilde{\mathcal{O}}\left(\frac{\log(\lvert\mathcal{R}\rvert)}{\varepsilon^2}\right)$
SPOIL (Moulin et al., 2025a)	$Q^{\Pi_{\mathcal{Q}}}$-realizability	✓	✗	$\widetilde{\mathcal{O}}\left(\frac{H^5 Q^4_{\max} \log(A\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q}))}{\varepsilon^4}\right)$
Q-OVI (**Ours**, Theorem J.1)	$Q^{\pi_{\mathsf{E}}}$-realizability	✗	✓	$\widetilde{\mathcal{O}}\left(\frac{H^3 Q^2_{\max} \log \mathcal{N}^{\max}_{\varepsilon/4H}(\mathcal{Q})}{\varepsilon^2}\right)$
OVI (**Ours**, Theorem F.1)	$Q^{\pi_{\mathsf{E}}}$-realizability	✓	✓	$\widetilde{\mathcal{O}}\left(\frac{H^5 Q^4_{\max} \log(A\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q}))}{\varepsilon^4}\right)^{\circ}$
OVI+ (**Ours**, Theorem I.1)	$Q^{\pi_{\mathsf{E}}}$-realizability	✓	✓	$\widetilde{\mathcal{O}}\left(\frac{H^3 Q^2_{\max} \log(\mathcal{N}^{\max}_{\varepsilon/24H}(\mathcal{Q}))}{\varepsilon^2}\right)^{\dagger}$
SPOIL (**Ours**, Theorem K.1)	$Q^{\pi_{\mathsf{E}}}$-realizability Bounded C_∞, optimal π_{E}	✓	✗	$\widetilde{\mathcal{O}}\left(\frac{C^4_\infty H^5 Q^4_{\max} \log(A\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q}))}{\varepsilon^4}\right)^{\diamond}$
Lower Bound (**Ours**, Theorem 4.1)	Linear $Q^{\pi_{\mathsf{E}}}$-realizability	–	✗	$\Omega\left(\frac{\lvert\mathcal{X}\rvert}{\varepsilon}\right)^{\triangle}$

$^{\circ}$ If $\mathcal{Q}$ is convex, the bound improves to $\widetilde{\mathcal{O}}(H^3 Q^2_{\max} \log(\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q}))\varepsilon^{-2})$.

† The number of learner rollouts is of order $\widetilde{\mathcal{O}}(\varepsilon^{-4})$, but expert queries are performed only along $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ of them.

$^{\diamond}$ If $\mathcal{Q}$ is convex, the bound improves to $\widetilde{\mathcal{O}}(C^2_\infty H^3 Q^2_{\max} \log(\mathcal{N}^{\max}_{r_\varepsilon}(\mathcal{Q}))\varepsilon^{-2})$.

$^{\triangle}$ The lower bound holds even if the expert is deterministic and the learner is allowed to replay the expert actions at the states logged in the dataset.

B.2 Empirical Value-Based Imitation Learning

Although early empirical IL successes include behavior-cloning systems such as ALVINN (Pomerleau, 1988, 1991), empirical value-based IL largely grew out of IRL and apprenticeship learning. Early reward-learning methods include maximum-margin planning and maximum-entropy or relative-entropy IRL (Ratliff et al., 2006; Ziebart et al., 2008; Boularias et al., 2011). These methods learn a reward or cost function that explains the expert's behavior and then derive a policy by planning or policy optimization under the learned objective. Deep variants, such as guided cost learning and maximum-entropy deep IRL, replace linear rewards with neural cost or reward functions (Finn et al., 2016; Wulfmeier et al., 2015). However, these methods typically inherit the "RL-in-the-loop" structure of Ziebart et al. (2008), requiring repeated planning or policy-optimization steps under a changing learned reward, which can be expensive in high-dimensional environments.

Other methods replace explicit planning with adversarial or occupancy-matching objectives. Ho et al. (2016) developed model-free apprenticeship-learning methods using policy-gradient optimization, while Ho and Ermon (2016) introduced generative adversarial IL (GAIL), which alternates between training a discriminator/cost signal and updating the policy, using Trust Region Policy Optimization (TRPO; Schulman et al., 2015) in the original implementation. GAIL and related adversarial IL methods inspired follow-up work on learned rewards, discriminator-actor-critic methods, off-policy distribution matching, and stationary-distribution correction (Fu et al., 2018; Kostrikov et al., 2019, 2020; Nachum et al., 2019). These methods can be effective, but adversarial IL is often sample-hungry

and can be difficult to stabilize because the policy optimizer must continually adapt to a nonstationary learned reward or discriminator.

More recent non-adversarial value- and Q-based methods aim to avoid some of this instability. For example, `SQIL` converts demonstrations into a sparse-reward RL problem (Reddy et al., 2019), while the `IQ-Learn` method of Garg et al. (2021) learns a single soft Q-function that implicitly represents both the reward and the policy. Such methods are closely aligned with the broad class of value-based IL algorithms considered in our lower bound: they extract a policy by applying a softmax, greedy, or related activation to a learned value function. `IQ-Learn` has been evaluated in simulated robotics tasks (Garg et al., 2021); related Q-learning, IRL, and moment-matching ideas have also been adapted to autoregressive sequence modeling and LM distillation, including `SequenceMatch` (Cundy and Ermon, 2024), scalable inverse soft-Q learning for language imitation (Wulfmeier et al., 2024), and adversarial moment-matching distillation (Jia, 2024). Demonstration-based reward-learning and IRL-style approaches have also been applied to LM supervised fine-tuning (SFT) and reasoning without preference labels or verifiers (Li et al., 2024; Cai et al., 2026). We refer to Sun and van der Schaar (2025) for a broader overview of IRL and IL in LM post-training, and to Sikchi et al. (2024) for a unifying perspective on practical dual RL and IL methods.

Policy-based imitation learning for language model fine-tuning. Policy-based, `DAgger`-like on-policy distillation methods have also been applied to LM fine-tuning. Agarwal et al. (2024) train student models on their own generated outputs using teacher feedback, and Lu and Thinking Machines Lab (2025) explicitly frames on-policy distillation as querying or evaluating a teacher on states visited by the student. These methods provide a policy-based analogue of the interactive protocols studied in this paper. Their strong empirical performance motivates the broader use of interaction in distillation, but our results suggest a complementary hypothesis: when the student is much less expressive than the teacher, interactive value-based distillation may be especially useful because it can target value-equivalent behavior rather than the teacher's full action distribution. The Gym experiments in Figure 1 provide preliminary evidence for this. Testing the same mechanism in LM distillation remains an important empirical direction.

C Value-Based Imitation Learning versus Policy-Based: Motivating Examples

In this section, we argue that Q^{π_E}-realizability can be weaker than π_E-realizability in two important cases.

1. In Appendix C.1, we show that when the expert is optimal, any class Π that realizes the expert policy π_E can be preprocessed into a class $\mathcal{Q}$ with $|\mathcal{Q}| \leq |\Pi|$ that realizes Q^{π_E} in a relaxed sense.
2. In Appendix C.2, we give an example where the environment makes the state-action value functions of the constructed expert policies simple, regardless of how complex the policy class is. In particular, $\mathcal{Q}$ can be easier to specify than Π when the expert behavior may depend on uncontrollable exogenous noise, while Q^{π_E} does not: the reward and transition dynamics are unaffected by this noise, whereas the expert's control policy is.

C.1 Constructing $\mathcal{Q}$ from Π Reduces Complexity

We first show that, given a policy class Π realizing an optimal expert policy π_E, one can construct a function class $\mathcal{Q}$ that realizes Q^{π_E} in a relaxed sense. We then show that OVI guarantees continue to hold under this relaxed realizability condition. These guarantees improve the representational-complexity term over those of BC, DAgger, and other policy-based methods that aim to match the expert trajectory distribution in total variation: indeed, such methods incur a $\log|\Pi|$ dependence, while OVI pays a factor $\log|\mathcal{Q}|$, where $\mathcal{Q}$ is constructed from Π and has no larger cardinality, i.e., $|\mathcal{Q}| \leq |\Pi|$.

Constructing $\mathcal{Q}$ from a policy class that realizes the expert. We preprocess Π by keeping only the action supports induced by its policies. Specifically, for each $\pi \in \Pi$, let $\widetilde{Q}^{\pi} = (\widetilde{Q}^{\pi}_h)_{h=1}^{H}$ with $\widetilde{Q}^{\pi}_h \colon \mathcal{X} \times \mathcal{A} \to [0, Q_{\max}]$ be defined by

$$\widetilde{Q}^{\pi}_h(x,a) = Q_{\max} \mathbb{1}_{\{a \in \mathsf{supp}(\pi_h(\cdot|x))\}},$$

for every $(h,x,a) \in [H] \times \mathcal{X} \times \mathcal{A}$. We then set

$$\mathcal{Q}(\Pi) = \left\{ \widetilde{Q}^{\pi} \colon \pi \in \Pi \right\}.$$

Since infinitely many stochastic policies in Π can share the same support, Π may be infinite even when $\mathcal{Q}(\Pi)$ is finite under the construction above.

Exact Q^{π_E}-realizability may fail. When the expert is (pointwise) optimal, π_E is greedy with respect to Q^{π_E}. The construction also includes the function corresponding to $\pi = \pi_E$,

$$\widetilde{Q}^{\pi_E}_h(x,a) = Q_{\max} \mathbb{1}_{\{a \in \mathsf{supp}(\pi_{E,h}(\cdot|x))\}},$$

with respect to which π_E is greedy. However, this does not imply that $\widetilde{Q}^{\pi_E} = Q^{\pi_E}$; in general, Q^{π_E} may not belong to $\mathcal{Q}$. The key point is that OVI only needs the weaker relationship between $\widetilde{Q}^{\pi_E}$, Q^{π_E}, and the expert's action support formalized as relaxed Q^{π_E}-realizability in Assumption C.1. The support-indicator witness above satisfies this condition because it assigns $Q_{\max}$ to every expert-supported action and zero elsewhere, while Q^{π_E} takes values in $[0, Q_{\max}]$.

C.1.1 Guarantees for OVI under relaxed Q^{π_E}-realizability

In this section, we present guarantees for OVI under a relaxed notion of Q^{π_E}-realizability. At a high level, the condition asks $\mathcal{Q}$ to contain a surrogate for Q^{π_E} that preserves the action comparisons used by OVI. We show that this condition is sufficient when the expert is statewise optimal, and that expert optimality is necessary.

Assumption C.1 (Relaxed Q^{π_E}-realizability)**.** *We say that $\mathcal{Q}$ satisfies relaxed Q^{π_E}-realizability if there exists a sequence of functions $\widetilde{Q}^{\pi_E} = (\widetilde{Q}^{\pi_E}_h)_{h=1}^{H} \in \mathcal{Q}$ such that, for every $(h,x) \in [H] \times \mathcal{X}$,*

$$\mathsf{supp}(\pi_{E,h}(\cdot \,|\, x)) \subseteq \arg\max_{b \in \mathcal{A}} \widetilde{Q}^{\pi_E}_h(x,b),$$

and, for every $(h,x,a) \in [H] \times \mathcal{X} \times \mathcal{A}$,

$$a \in \mathsf{supp}(\pi_{E,h}(\cdot \,|\, x)) \implies \widetilde{Q}^{\pi_E}_h(x,a) \geq Q^{\pi_E}_h(x,a),$$
$$a \notin \mathsf{supp}(\pi_{E,h}(\cdot \,|\, x)) \implies \widetilde{Q}^{\pi_E}_h(x,a) \leq Q^{\pi_E}_h(x,a).$$

Under the relaxed $Q^{\pi_{\mathsf{E}}}$-realizability assumption above, we can prove the same decomposition as in Lemma F.1 for the perfectly realizable case, i.e., $Q^{\pi_{\mathsf{E}}} \in \mathcal{Q}$.

Lemma C.1. *Let Assumption C.1 hold, and assume that the expert policy π_{E} is statewise optimal. Let policies $\pi^1, \dots, \pi^K$ and functions $Q_h^k \in \mathcal{Q}_h$ be computed as in Eq. (17). Define the π-regret at stage h and state $x \in \mathcal{X}$ as*

$$\mathfrak{R}_h^{\pi}(x) = \sum_{k=1}^{K} \big\langle Q_h^k(x, \cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x) \big\rangle,$$

define the output policy as $\pi_{\mathsf{out}} = \frac{1}{K} \sum_{k=1}^{K} \pi^k$, and, for any policy π, define the estimation error as

$$\Delta(\pi) = \max_{h \in [H]} \sup_{Q_h \in \mathcal{Q}_h} \left| \widehat{\mathcal{L}}_h(\pi_h, Q_h) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h, Q_h) \right|.$$

Then, it holds that

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K} \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \mathfrak{R}_h^{\pi}(x) + \frac{2H}{K} \sum_{k=1}^{K} \Delta\big(\pi^k\big).$$

At this point, one can establish the same complexity guarantee as `OVI` in the exactly realizable setting by bounding $\frac{2H}{K} \sum_{k=1}^{K} \Delta(\pi^k)$ with high probability, as in the proof of Theorem F.1. To conclude the section, we prove Lemma C.1.

Proof of Lemma C.1. The proof follows the proof of Lemma F.1, except that we first compare $Q_h^{\pi_{\mathsf{E}}}$ with the relaxed witness $\widetilde{Q}_h^{\pi_{\mathsf{E}}}$. For any policy π, we use the shorthand $\mathcal{L}_h^{\pi} := \mathcal{L}_h^{d^{\pi}}$ and we claim that, for every $h \in [H]$,

$$\mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}) \le \mathcal{L}_h^{\pi_{\mathsf{out}}}\Big(\pi_{\mathsf{out},h}, \widetilde{Q}_h^{\pi_{\mathsf{E}}}\Big). \tag{2}$$

To prove this claim, fix $h \in [H]$ and $x \in \mathcal{X}$, write $\mathcal{A}_{x,h}^{\mathsf{E}} := \mathsf{supp}(\pi_{\mathsf{E},h}(\cdot \,|\, x))$, and set $\delta_h^x(a) := \pi_{\mathsf{E},h}(a \,|\, x) - \pi_{\mathsf{out},h}(a \,|\, x)$. By definition of $\mathcal{L}_h^{\pi_{\mathsf{out}}}$,

$$\mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}) = \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \sum_{a \in \mathcal{A}} \delta_h^x(a) Q_h^{\pi_{\mathsf{E}}}(x, a).$$

Fix a state x. Since π_{E} is statewise optimal, every action in $\mathcal{A}_{x,h}^{\mathsf{E}}$ is greedy with respect to $Q_h^{\pi_{\mathsf{E}}}(x, \cdot)$, that is, $\mathcal{A}_{x,h}^{\mathsf{E}} \subseteq \arg\max_{b \in \mathcal{A}} Q_h^{\pi_{\mathsf{E}}}(x, b)$. Thus, by definition of δ_h^x,

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \delta_h^x(a) Q_h^{\pi_{\mathsf{E}}}(x, a) &= \sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \delta_h^x(a) Q_h^{\pi_{\mathsf{E}}}(x, a) + \sum_{a \notin \mathcal{A}_{x,h}^{\mathsf{E}}} \delta_h^x(a) Q_h^{\pi_{\mathsf{E}}}(x, a) \\
&= \max_{b \in \mathcal{A}} Q_h^{\pi_{\mathsf{E}}}(x, b) \Big(1 - \sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x)\Big) - \sum_{a \notin \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x) Q_h^{\pi_{\mathsf{E}}}(x, a),
\end{aligned}$$

where we used that $\pi_{\mathsf{E},h}(a \,|\, x) = 0$ for $a \notin \mathcal{A}_{x,h}^{\mathsf{E}}$, that $\sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{E},h}(a \,|\, x) = 1$, and that all actions in $\mathcal{A}_{x,h}^{\mathsf{E}}$ maximize $Q_h^{\pi_{\mathsf{E}}}(x, \cdot)$. By the relaxed realizability condition (Assumption C.1), we can replace $Q_h^{\pi_{\mathsf{E}}}$ with $\widetilde{Q}_h^{\pi_{\mathsf{E}}}$ as follows

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \delta_h^x(a) Q_h^{\pi_{\mathsf{E}}}(x, a) &\overset{\text{(a)}}{\le} \max_{b \in \mathcal{A}} \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, b) \Big(1 - \sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x)\Big) - \sum_{a \notin \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x) \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, a) \\
&\overset{\text{(b)}}{=} \sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \delta_h^x(a) \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, a) - \sum_{a \notin \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x) \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, a) \\
&\overset{\text{(c)}}{=} \sum_{a \in \mathcal{A}} \delta_h^x(a) \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, a).
\end{aligned}$$

Step (a) uses that on the support $\mathcal{A}_{x,h}^{\mathsf{E}}$, $\widetilde{Q}_h^{\pi_{\mathsf{E}}}$ is at least $Q_h^{\pi_{\mathsf{E}}}$, so $\max_{b \in \mathcal{A}} \widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, b) \ge \max_{b \in \mathcal{A}} Q_h^{\pi_{\mathsf{E}}}(x, b)$, that $1 - \sum_{a \in \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x) = \sum_{a \notin \mathcal{A}_{x,h}^{\mathsf{E}}} \pi_{\mathsf{out},h}(a \,|\, x) \ge 0$, and that outside the support $\mathcal{A}_{x,h}^{\mathsf{E}}$, $\widetilde{Q}_h^{\pi_{\mathsf{E}}}(x, a) \le Q_h^{\pi_{\mathsf{E}}}(x, a)$ holds, so the inequality is reversed when multiplied by

$-\pi_{\mathsf{out},h}(a \mid x)$. Step (b) uses $\mathcal{A}^{\mathsf{E}}_{x,h} \subseteq \arg\max_{b \in \mathcal{A}} \widetilde{Q}^{\pi_{\mathsf{E}}}_h(x,b)$ and $\sum_{a \in \mathcal{A}^{\mathsf{E}}_{x,h}} \pi_{\mathsf{E},h}(a \mid x) = 1$. Finally, (c) uses the definition of δ^x_h and the fact that $\delta^x_h(a) = -\pi_{\mathsf{out},h}(a \mid x)$ for $a \notin \mathcal{A}^{\mathsf{E}}_{x,h}$. Averaging the last display over $x \sim d^{\pi_{\mathsf{out}}}_h$ gives Eq. (2).

Since $\widetilde{Q}^{\pi_{\mathsf{E}}} \in \mathcal{Q}$ and hence $\widetilde{Q}^{\pi_{\mathsf{E}}}_h \in \mathcal{Q}_h$ for each $h \in [H]$, we can now upper bound the suboptimality as in Lemma F.1. By the performance difference lemma (Lemma E.9), and the inequality we just proved (Eq. (2)), we have

$$\begin{aligned} K(J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}) &= K \sum_{h=1}^{H} \mathcal{L}^{\pi_{\mathsf{out}}}_h(\pi_{\mathsf{out},h}, Q^{\pi_{\mathsf{E}}}_h) \\ &\le K \sum_{h=1}^{H} \mathcal{L}^{\pi_{\mathsf{out}}}_h\Big(\pi_{\mathsf{out},h}, \widetilde{Q}^{\pi_{\mathsf{E}}}_h\Big). \end{aligned}$$

Furthermore, by the definitions of the policy π_{out} and the estimation error Δ, we have

$$\begin{aligned} K(J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}) &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}^{\pi_{\mathsf{out}}}_h\Big(\pi^k_h, \widetilde{Q}^{\pi_{\mathsf{E}}}_h\Big) \\ &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \widehat{\mathcal{L}}_h\Big(\pi^k_h, \widetilde{Q}^{\pi_{\mathsf{E}}}_h\Big) + H\sum_{k=1}^{K}\Delta\big(\pi^k\big) \\ &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \widehat{\mathcal{L}}_h\big(\pi^k_h, Q^k_h\big) + H\sum_{k=1}^{K}\Delta\big(\pi^k\big) \\ &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}^{\pi_{\mathsf{out}}}_h\big(\pi^k_h, Q^k_h\big) + 2H\sum_{k=1}^{K}\Delta\big(\pi^k\big). \end{aligned}$$

Here, the first and third inequalities use the definition of $\Delta(\pi^k)$, respectively with $\widetilde{Q}^{\pi_{\mathsf{E}}}_h \in \mathcal{Q}_h$ and $Q^k_h \in \mathcal{Q}_h$. The second inequality uses $\widetilde{Q}^{\pi_{\mathsf{E}}}_h \in \mathcal{Q}_h$ (Assumption C.1) and the fact that Q^k_h is a best response to $\widehat{\mathcal{L}}_h(\pi^k_h, \cdot)$ over $\mathcal{Q}_h$. Finally, for each $h \in [H]$, expanding $\mathcal{L}^{\pi_{\mathsf{out}}}_h$ gives

$$\sum_{k=1}^{K} \mathcal{L}^{\pi_{\mathsf{out}}}_h\big(\pi^k_h, Q^k_h\big) = \sum_{x \in \mathcal{X}} d^{\pi_{\mathsf{out}}}_h(x) \sum_{k=1}^{K} \big\langle Q^k_h(x,\cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - \pi^k_h(\cdot \mid x)\big\rangle = \sum_{x \in \mathcal{X}} d^{\pi_{\mathsf{out}}}_h(x)\mathfrak{R}^{\pi}_h(x).$$

Dividing by K concludes the proof. □

C.1.2 Expert optimality is necessary for learning under relaxed $Q^{\pi_{\mathsf{E}}}$-realizability

We conclude by showing that expert optimality is necessary. The failure mode for algorithms that output a VI policy under relaxed $Q^{\pi_{\mathsf{E}}}$-realizability with a suboptimal expert is that the relaxed condition forces the algorithm to treat all actions in the expert support *equally*. For stochastic suboptimal experts, this can increase the probability of actions rarely played by the expert.

Theorem C.1. *For any algorithm* Alg *outputting a VI policy (see Definition G.1, even if allowed expert interaction, there exists a 2-action MDP $\mathcal{M}$ with a suboptimal expert policy and a class $\mathcal{Q}$ satisfying Assumption C.1 such that, even if* Alg *performs infinitely many expert queries,* $\mathbb{E}_{\pi_{\mathsf{out}} \sim \mathrm{Alg}}[J^{\pi_{\mathsf{E}}}_{\mathcal{M}} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}}] = \Omega(H)$.

Proof of Theorem C.1. Consider a class of environments $\{\mathcal{M}_1, \mathcal{M}_2\}$ with action space $\mathcal{A} = \{a_1, a_2\}$ and a single absorbing state x, i.e., both actions from x transition back to x with probability 1. For $i \in [2]$, let $r(x, a_i) = 1$ in $\mathcal{M}_i$ and $r(x,a) = 0$ for the other action $a \neq a_i$. In $\mathcal{M}_i$, let the expert policy $\pi_{\mathsf{E},i}$ satisfy $\pi_{\mathsf{E},i}(a_i \mid x) = 0.9$ and $\pi_{\mathsf{E},i}(a \mid x) = 0.1$ for $a \neq a_i$. Assumption C.1 is satisfied by the singleton class $\mathcal{Q} = \{\widetilde{Q}\}$, where $\widetilde{Q}_h(x,a) = H$ for all $(h,a) \in [H] \times \mathcal{A}$. Indeed, for both $\pi_{\mathsf{E},1}$ and $\pi_{\mathsf{E},2}$, the expert support is the full action space $\mathcal{A}$. Under the additional value-derived output restriction that the learner can only output a policy obtained by applying a scalar link function to a linear combination of functions in $\mathcal{Q}$, the output policy is uniform over the action space because

$\widetilde{Q}_h$ assigns the same value to all actions, regardless of which environment the learner faces. Therefore,

$$\max_{i\in[2]} \mathbb{E}_{\pi_{\mathsf{out}}\sim \mathrm{Alg}}\left[J^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}_i}\right] \geq \frac{1}{2}\sum_{i=1}^{2} \mathbb{E}_{\pi_{\mathsf{out}}\sim \mathrm{Alg}}\left[J^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}_i}\right] = 0.9H - 0.5H = 0.4H.$$

This concludes the proof. □

C.2 $Q^{\pi_{\mathsf{E}}}$-Realizability Can Be Strictly Weaker Than π_{E}-Realizability under Exogenous Noise

In this section, we give an example in which expert-policy realizability is more demanding than $Q^{\pi_{\mathsf{E}}}$-realizability. The example is an exogenous block MDP (ExBMDP), a standard rich-observation model in which the endogenous control problem is low-dimensional, but the observation-based expert policy must decode a high-dimensional observation. A typical example is a robot with high-dimensional camera observations that contain the (small) endogenous state and a background that evolves independently of the robot's actions (e.g., a television). We formalize the setting below, and the results are stated and proved in subsequent subsections.

One limitation of this example is that a single policy that always chooses action 0 is optimal in every MDP in the family, so the construction concerns exact representational complexity rather than control hardness. We view this as acceptable because the purpose of the construction is to show that exact expert-policy realizability can be much more demanding than $Q^{\pi_{\mathsf{E}}}$-realizability.

The motivation is directly tied to the complexity terms that appear in sample-complexity guarantees. Policy-based methods such as `BC` and `DAgger` aim to imitate the expert's action distribution. Under an expert-policy realizability assumption, their guarantees typically depend on the complexity of a policy class Π satisfying $\pi_{\mathsf{E}} \in \Pi$, through a term such as $\log|\Pi|$ (or a log-covering number). In contrast, our value-based guarantees depend on the complexity of a value-function class $\mathcal{Q}$ satisfying $Q^{\pi_{\mathsf{E}}} \in \mathcal{Q}$, through a term such as $\log|\mathcal{Q}|$.

The construction below shows that these two quantities can be exponentially different. We construct a family of ExBMDPs $\{\mathcal{M}_b : b \in \mathcal{B}_n\}$ and deterministic, suboptimal experts $\{\pi^b : b \in \mathcal{B}_n\}$ indexed by balanced Boolean decoders $b\colon \{0,1\}^n \to \{0,1\}$, where n is the number of bits in the observation space at stage 1. Any expert-agnostic policy class that contains all experts in the family must have exponentially large metric entropy: for every $\varepsilon \in (0, 1/2)$,

$$\inf_{\Pi:\forall b\in\mathcal{B}_n, \pi^b\in\Pi} \log_2 \mathcal{N}_\varepsilon(\Pi, d_\Pi) \geq 2^n - n - 1\,,$$

where d_Π is the sup-total variation metric. By contrast, a two-dimensional class $\mathcal{Q}$ realizes the expert action-value function $Q^{\pi^b}_{\mathcal{M}_b}$ for every $b \in \mathcal{B}_n$, and

$$\log \mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty) \leq 2\log(3/\varepsilon)\,.$$

Informally, the construction is as follows. We consider an instance where the endogenous state is a single bit $s \in \{0,1\}$, and the expert's endogenous policy simply plays action $a = s$ among the actions $\mathcal{A} = \{0,1,2\}$. However, the learner does not observe s directly. Instead, the learner observes a high-dimensional binary string $w \in \{0,1\}^n$, and recovering s from w requires computing a decoder $b(w)$. Thus, an observation-based policy that exactly imitates the expert must compute $\pi_{\mathsf{E},1}(w) = b(w)$, and the complexity of a policy class realizing the expert must scale with the complexity of the decoder. On the other hand, the Q-function does not need this decoder. At the first layer, actions 0 and 1 are both safe and value-equivalent, while action 2 is unsafe. Hence $Q^{\pi_{\mathsf{E}}}$ only needs to know the value-relevant distinction of $\{0,1\}$ vs. $\{2\}$. The hard decoder determines only the expert's tie-breaking between the two safe actions, which is what creates this separation. Every expert obtains return $1/2$, while the policy that always chooses action 0 obtains the optimal return 1. Thus the separation holds even though every expert is strictly suboptimal.

C.2.1 Preliminaries

Exogenous block MDPs. A finite-horizon exogenous block MDP (ExBMDP) consists of a tuple

$$\left(H, (\mathcal{S}_h)_{h=1}^H, (\Xi_h)_{h=1}^H, (\mathcal{X}_h)_{h=1}^H, \mathcal{A}, \left(T_h^{\mathsf{endo}}\right)_{h=0}^{H-1}, (T_h^{\mathsf{exo}})_{h=0}^{H-1}, \left(g_h^{\mathsf{obs}}\right)_{h=1}^H, (r_h)_{h=1}^H\right),$$

where H is the horizon, and at each layer h, $\mathcal{S}_h$ is the endogenous state space, Ξ_h is the exogenous state space, $\mathcal{X}_h$ is the observation space, $\mathcal{A}$ is the action space, T_0^{endo} and T_0^{exo} are initial distributions,

for $h \in [H-1]$ the map $T_h^{\mathsf{endo}} \colon \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$ is the endogenous transition kernel and $T_h^{\mathsf{exo}} \colon \Xi_h \to \Delta(\Xi_{h+1})$ is the exogenous transition kernel, $g_h^{\mathsf{obs}} \colon \mathcal{S}_h \times \Xi_h \to \mathcal{X}_h$ is the observation map, and $r_h \colon \mathcal{X}_h \times \mathcal{A} \to [0,1]$ is the reward function.

The initial latent state is generated by $\boldsymbol{s}_1 \sim T_0^{\mathsf{endo}}(\cdot \mid \emptyset)$, $\boldsymbol{\xi}_1 \sim T_0^{\mathsf{exo}}(\cdot \mid \emptyset)$ independently, with the convention that $\mathcal{S}_0 = \Xi_0 = \emptyset$. For $h \in [H-1]$, after taking action $\boldsymbol{a}_h \in \mathcal{A}$, the latent state evolves according to $\boldsymbol{s}_{h+1} \sim T_h^{\mathsf{endo}}(\cdot \mid \boldsymbol{s}_h, \boldsymbol{a}_h)$, and $\boldsymbol{\xi}_{h+1} \sim T_h^{\mathsf{exo}}(\cdot \mid \boldsymbol{\xi}_h)$. Thus only the endogenous state is causally affected by the action; the exogenous state evolves independently of the action.

The observation at layer h is generated by a deterministic emission map $\boldsymbol{x}_h = g_h^{\mathsf{obs}}(\boldsymbol{s}_h, \boldsymbol{\xi}_h) \in \mathcal{X}_h$. The model is decodable if for any h, there exists a decoder $\varphi_h^\star \colon \mathcal{X}_h \to \mathcal{S}_h$ such that for any $s \in \mathcal{S}_h$ and $\xi \in \Xi_h$, $\varphi_h^\star(g_h^{\mathsf{obs}}(s, \xi)) = s$. Finally, the rewards are endogenous if there exist functions $\bar{r}_h \colon \mathcal{S}_h \times \mathcal{A} \to [0,1]$ such that $r_h(x,a) = \bar{r}_h(\varphi_h^\star(x), a)$ for all $h \in [H]$, $x \in \mathcal{X}_h$, and $a \in \mathcal{A}$.

Because we work in a finite-horizon nonstationary MDP, policies and value functions are also indexed by the layer h. Thus a policy is a sequence $\pi = (\pi_h \colon \mathcal{X}_h \to \Delta(\mathcal{A}))_{h=1}^H$. In the construction below, we use separate observation spaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Equivalently, one can combine them into a single disjoint union $\mathcal{X} = \mathcal{X}_1 \sqcup \mathcal{X}_2$.

Balanced decoders. For the construction below, we will consider the family of balanced Boolean functions as decoders, defined for any $n \geq 1$ as

$$\mathcal{B}_n = \left\{ b \colon \{0,1\}^n \to \{0,1\} : \left|b^{-1}(0)\right| = \left|b^{-1}(1)\right| = 2^{n-1} \right\}.$$

The number of balanced Boolean functions is $|\mathcal{B}_n| = \binom{2^n}{2^{n-1}}$. By unimodality of the binomial coefficient, and since the largest binomial coefficient is at least the average binomial coefficient, we have

$$\binom{2^n}{2^{n-1}} = \max_{k \in [2^n]} \binom{2^n}{k} \geq \frac{1}{2^n+1} \sum_{k=0}^{2^n} \binom{2^n}{k} = \frac{2^{2^n}}{2^n+1}.$$

Therefore $\log_2 |\mathcal{B}_n| \geq 2^n - \log_2(2^n+1) \geq 2^n - n - 1$.

C.2.2 The construction

Fix an integer $n \geq 1$ and a decoder $b \in \mathcal{B}_n$.

Exogenous block MDP. We define an ExBMDP $\mathcal{M}_b$ with horizon $H = 2$ and action space $\mathcal{A} = \{0,1,2\}$. Actions 0 and 1 are safe first-layer actions, while action 2 is unsafe.

Let $\mathcal{S} = \mathcal{S}_1 \sqcup \mathcal{S}_2$ be a layered endogenous state space with 7 states, where the first layer is $\mathcal{S}_1 = \{0,1\}$ and the second layer is $\mathcal{S}_2 = \mathcal{G} \cup \{B\}$, with $\mathcal{G} = \{G_{s,p} : s \in \{0,1\},\ p \in \{0,1\}\}$. The state $G_{s,p}$ is a good state reached from first-layer endogenous state s after taking safe action p. The state B is a bad state reached after taking action 2.

Let the exogenous state space be time-homogeneous and given by $\Xi_1 = \Xi_2 = \{0,1\}^{n-1}$. Since b is balanced, we have $\left|b^{-1}(0)\right| = \left|b^{-1}(1)\right| = |\Xi_1| = 2^{n-1}$. Therefore, for each $s \in \{0,1\}$, we can fix an arbitrary bijection $\eta_s^b \colon \Xi_1 \to b^{-1}(s)$ that maps an $(n-1)$-bit exogenous variable to an n-bit observation string.

The observation spaces are $\mathcal{X}_1 = \{0,1\}^n$ and $\mathcal{X}_2 = \mathcal{S}_2 \times \Xi_2$.

The endogenous dynamics are defined as follows. The initial endogenous state is sampled uniformly at random, $\boldsymbol{s}_1 \sim \mathrm{Unif}(\{0,1\})$, and the transition from layer 1 to layer 2 is deterministic:

$$\boldsymbol{s}_2 = \begin{cases} G_{\boldsymbol{s}_1, \boldsymbol{a}_1}, & \boldsymbol{a}_1 \in \{0,1\}, \\ B, & \boldsymbol{a}_1 = 2. \end{cases}$$

Thus, the first action genuinely affects the endogenous state: actions 0 and 1 lead to different good branches, while action 2 leads to the bad state. This specifies a valid deterministic transition kernel from $\mathcal{S}_1 \times \mathcal{A}$ to $\mathcal{S}_2$.

For the exogenous dynamics, the initial exogenous state is sampled independently of $\boldsymbol{s}_1$, also uniformly at random, $\boldsymbol{\xi}_1 \sim \mathrm{Unif}(\Xi_1)$, and the exogenous process is persistent, i.e., $\boldsymbol{\xi}_2 = \boldsymbol{\xi}_1$. Equivalently, $T_1^{\mathsf{exo}}(\xi' \mid \xi) = \mathbb{I}\{\xi' = \xi\}$. The exogenous transition is independent of the action.

We define the observation map at stage 1 for any $(s,\xi) \in \mathcal{S}_1 \times \Xi_1$ as $g^{\mathsf{obs}}_{1,b}(s,\xi) = \eta^b_s(\xi)$. Hence, conditional on $s=0$, the observed string is uniformly distributed over $b^{-1}(0)$, and conditional on $s=1$, it is uniformly distributed over $b^{-1}(1)$. Since $\boldsymbol{s}_1$ and $\boldsymbol{\xi}_1$ are both uniform, every observation $w \in \{0,1\}^n$ is reachable and equally likely at layer 1: if $b(w) = s$, then $\mathbb{P}[\boldsymbol{x}_1 = w] = \mathbb{P}[\boldsymbol{s}_1 = s]\mathbb{P}[\eta^b_s(\boldsymbol{\xi}_1) = w] = \frac{1}{2} \cdot 2^{-(n-1)} = 2^{-n}$. We can think of the observation map as hiding the endogenous state within an n-bit observation string via the mapping η^b_s. The observation map at stage 2 is transparent: for any $(y,\xi) \in \mathcal{S}_2 \times \Xi_2$, $g^{\mathsf{obs}}_{2,b}(y,\xi) = (y,\xi)$.

Define the decoder $\varphi^\star_b$ by $\varphi^\star_{1,b}(w) = b(w)$ for any $w \in \{0,1\}^n$, and $\varphi^\star_{2,b}(y,\xi) = y$ for any $y \in \mathcal{S}_2,\ \xi \in \Xi_2$. Then, for every latent state (s_h, ξ_h), $\varphi^\star_{h,b}(g^{\mathsf{obs}}_{h,b}(s_h,\xi_h)) = s_h$. Indeed, at layer 1, we have

$$\varphi^\star_{1,b}\big(g^{\mathsf{obs}}_{1,b}(s,\xi)\big) = \varphi^\star_{1,b}\big(\eta^b_s(\xi)\big) = b\big(\eta^b_s(\xi)\big) = s,$$

because $\eta^b_s(\xi) \in b^{-1}(s)$. At layer 2, the claim follows from the transparent observation map.

Finally, we define the reward function. At stage 1, the reward is zero everywhere: for any observation $x \in \mathcal{X}_1$ and action $a \in \mathcal{A}$, $r_1(x,a) = 0$. At stage 2, for any observation $(y,\xi) \in \mathcal{X}_2$ and $a \in \mathcal{A}$, we define $r_2((y,\xi),a) = \mathbb{I}\{y \in \mathcal{G}\}\big(\mathbb{I}\{a=0\} + \frac{1}{2}\mathbb{I}\{a=1\}\big)$. Thus action 0 gives reward 1 and action 1 gives reward $1/2$ in good second-layer states, while all actions give reward 0 in the bad state B. The reward is endogenous: taking $\bar{r}_1 \equiv 0$ and $\bar{r}_2(y,a) = \mathbb{I}\{y \in \mathcal{G}\}\big(\mathbb{I}\{a=0\} + \frac{1}{2}\mathbb{I}\{a=1\}\big)$ gives $r_h(x,a) = \bar{r}_h(\varphi^\star_{h,b}(x),a)$ for both layers. Together with the decoder identity above and the action-independent exogenous transition, this defines a valid decodable ExBMDP with endogenous rewards.

Expert policy. Define the endogenous expert policy $\bar{\pi}^b \colon \mathcal{S} \to \mathcal{A}$ by

$$\bar{\pi}^b_1(s) = s, \quad s \in \{0,1\}, \quad \text{and} \quad \bar{\pi}^b_2(y) = 1, \quad y \in \mathcal{S}_2.$$

The corresponding observation-based expert policy $\pi^b \colon \mathcal{X} \to \mathcal{A}$ is

$$\pi^b_1(w) = b(w), \quad w \in \{0,1\}^n, \quad \text{and} \quad \pi^b_2(y,\xi) = 1, \quad y \in \mathcal{S}_2,\ \xi \in \Xi_2.$$

Thus π^b is endogenous, since for any layer h and observation $x \in \mathcal{X}_h$, $\pi^b_h(x) = \bar{\pi}^b_h(\varphi^\star_{h,b}(x))$. In endogenous coordinates, the expert is simple. In raw observation coordinates, the first-layer expert action is exactly the decoder b.

The expert always reaches a good state and then plays action 1, so $J^{\pi^b}_{\mathcal{M}_b} = 1/2$. By contrast, the policy that always plays action 0 reaches a good state and receives reward 1, which is the maximum possible return. Hence every expert is strictly suboptimal with suboptimality gap $1/2$.

C.2.3 A two-dimensional class realizes the expert Q-function

Consider the two-parameter time-indexed function class $\mathcal{Q} = \{Q^{\alpha,\beta} : \alpha,\beta \in [0,1]\}$, where, for observations $x \in \mathcal{X}_1$, $(y,\xi) \in \mathcal{X}_2$, and action $a \in \mathcal{A}$, the functions are defined as follows.

$$Q^{\alpha,\beta}_1(x,a) = \frac{\alpha}{2}\,\mathbb{I}\{a \in \{0,1\}\}, \quad \text{and} \quad Q^{\alpha,\beta}_2((y,\xi),a) = \beta\,\mathbb{I}\{y \in \mathcal{G}\}\left(\mathbb{I}\{a=0\} + \frac{1}{2}\mathbb{I}\{a=1\}\right).$$

Every function in $\mathcal{Q}$ is bounded by $Q_{\mathsf{max}} = 1$. Importantly, the class $\mathcal{Q}$ does not depend on b. At layer 1, it does not distinguish the strings $w \in \{0,1\}^n$ at all; it only distinguishes the safe actions $0,1$ from the unsafe action 2. For any function $Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, we will use the norm $\|Q\|_\infty = \max_{h\in\{1,2\}} \max_{x\in\mathcal{X}_h} \max_{a\in\mathcal{A}} |Q_h(x,a)|$. We start by showing that every expert in the family has the same Q-function, which is realizable by $\mathcal{Q}$.

Lemma C.2 ($Q^{\pi^b}_{\mathcal{M}_b}$-realizability)**.** *For every $b \in \mathcal{B}_n$, $Q^{\pi^b}_{\mathcal{M}_b} = Q^{1,1} \in \mathcal{Q}$.*

Proof of Lemma C.2. At layer 2, there is no future reward. Thus, for every $(y,\xi) \in \mathcal{X}_2$ and $a \in \mathcal{A}$,

$$Q^{\pi^b}_{\mathcal{M}_b,2}((y,\xi),a) = r_2((y,\xi),a) = \mathbb{I}\{y \in \mathcal{G}\}\left(\mathbb{I}\{a=0\} + \frac{1}{2}\mathbb{I}\{a=1\}\right) = Q^{1,1}_2((y,\xi),a).$$

At layer 1, the immediate reward is zero. Fix any observation $w \in \{0,1\}^n$. Its decoded endogenous state is $s = \varphi^\star_{1,b}(w) = b(w)$. If $a \in \{0,1\}$, then the next endogenous state is $G_{s,a} \in \mathcal{G}$. At layer 2,

the expert plays action 1, receiving reward $1/2$. Thus, for $a \in \{0,1\}$, $Q^{\pi^b}_{\mathcal{M}_b,1}(w,a) = 1/2$. If $a = 2$, then the next endogenous state is B. From B, every action receives reward 0, so $Q^{\pi^b}_{\mathcal{M}_b,1}(w,2) = 0$. Hence

$$Q^{\pi^b}_{\mathcal{M}_b,1}(w,a) = \frac{1}{2}\mathbb{I}\{a \in \{0,1\}\} = Q^{1,1}_1(w,a).$$

Combining the two layers gives $Q^{\pi^b}_{\mathcal{M}_b} = Q^{1,1} \in \mathcal{Q}$. □

We can then bound the covering number of the class $\mathcal{Q}$ with the following lemma.

Lemma C.3 (Covering number of the Q-class)**.** *For every $\varepsilon \in (0,1]$, $\mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty) \le (\frac{3}{\varepsilon})^2$.*

Proof of Lemma C.3. For any $\alpha, \beta, \alpha', \beta' \in [0,1]$,

$$\left\| Q^{\alpha,\beta} - Q^{\alpha',\beta'} \right\|_\infty \le \max\{|\alpha - \alpha'|, |\beta - \beta'|\}.$$

Thus any ε-grid of $[0,1]^2$ in the ℓ_∞ norm induces an ε-cover of $\mathcal{Q}$. By Lemma E.5, such a grid has cardinality at most $(\lceil \frac{1}{\varepsilon} \rceil + 1)^2$. For $\varepsilon \in (0,1]$, $\lceil \frac{1}{\varepsilon} \rceil + 1 \le \frac{3}{\varepsilon}$. This proves the claim. □

C.2.4 Policy realizability requires an exponentially large class

We show that any policy class Π that contains all experts in the family must have an exponentially large covering number under the sup-total-variation metric d_Π. In particular, if Π is finite, then it must have exponentially large cardinality. This shows that a policy realizability assumption is much stronger than a value-function realizability assumption, since the same two-dimensional class $\mathcal{Q}$ realizes the expert Q-function for every expert in the family, while any single policy class realizing all experts must have exponential complexity.

For any policies π, π', define the sup-total-variation metric

$$d_\Pi(\pi, \pi') = \frac{1}{2} \max_{h \in \{1,2\}} \|\pi_h - \pi'_h\|_{\infty,1},$$

where $\|\cdot\|_{\infty,1}$ is defined in Eq. (3). For deterministic policies, this distance equals 1 whenever the two policies choose different actions at some layer and observation. Let

$$\Pi_n = \{\pi^b : b \in \mathcal{B}_n\}$$

be the natural expert-agnostic policy class containing all possible experts in the hard-decoder family. We have the following result.

Lemma C.4 (Family-level policy lower bound)**.** *Let $\Pi \subset \Delta(\mathcal{A})^{\mathcal{X}}$ be any policy class. If $\Pi_n \subseteq \Pi$, then, for every $\varepsilon \in (0, 1/2)$, $\mathcal{N}_\varepsilon(\Pi, d_\Pi) \ge |\mathcal{B}_n|$. In particular,*

$$\log_2 \mathcal{N}_\varepsilon(\Pi, d_\Pi) \ge 2^n - n - 1.$$

If Π is finite, then also $|\Pi| \ge |\mathcal{B}_n|$.

Proof of Lemma C.4. If $b \neq b'$ are two different decoders in $\mathcal{B}_n$, then there exists an observation $w \in \{0,1\}^n$ such that $b(w) \neq b'(w)$. As a result, $\pi^b_1(w) = b(w) \neq b'(w) = \pi^{b'}_1(w)$. Therefore, $\pi^b \neq \pi^{b'}$. Hence the map $b \mapsto \pi^b$ is injective. Any finite policy class containing all experts $\{\pi^b : b \in \mathcal{B}_n\}$ must therefore have cardinality at least $|\mathcal{B}_n|$.

Moreover, if $b \neq b'$, then the deterministic policies π^b and $\pi^{b'}$ choose different actions at some observation w in stage 1. Hence $d_\Pi(\pi^b, \pi^{b'}) = 1$. Thus $\{\pi^b : b \in \mathcal{B}_n\}$ is a 1-packing under d_Π. A ball of radius $\varepsilon < 1/2$ cannot contain two distinct elements of this packing. Therefore every ε-cover of Π must contain at least $|\mathcal{B}_n|$ elements. Finally, the lower bound on the log-covering number follows from the lower bound on $|\mathcal{B}_n|$ shown in Appendix C.2.1. □

C.2.5 Separation theorem

We now summarize the separation. The result compares the complexity terms that enter policy-based and value-based guarantees. It shows that on the hard-decoder family, (i) any expert-agnostic class satisfying expert-policy realizability uniformly over all decoders b has exponential metric entropy, while (ii) a value-based guarantee under Q^{π_E}-realizability depends on the complexity of a two-dimensional class realizing the expert action-value function.

Theorem C.2 (Hard-decoder ExBMDP separation)**.** *The ExBMDP family* $\{\mathcal{M}_b : b \in \mathcal{B}_n\}$ *defined in Appendix C.2.2 satisfies the following properties.*

1. ***Uniform*** Q^{π_E}***-realizability.*** *For every balanced decoder* $b \in \mathcal{B}_n$*, the expert policy* π^b *is deterministic and endogenous in* $\mathcal{M}_b$*. It is strictly suboptimal, with* $J^{\pi^b}_{\mathcal{M}_b} = 1/2$ *and* $\sup_\pi J^\pi_{\mathcal{M}_b} = 1$*. Moreover, the two-dimensional class* $\mathcal{Q}$ *defined in Appendix C.2.3 realizes every expert action-value function in the family: for every* $b \in \mathcal{B}_n$*,* $Q^{\pi^b}_{\mathcal{M}_b} = Q^{1,1} \in \mathcal{Q}$*. In addition, for every* $\varepsilon \in (0, 1]$*,*

$$\log \mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty) \le 2 \log\left(\frac{3}{\varepsilon}\right).$$

2. ***Family-level policy lower bound.*** *Let* $\Pi \subset \Delta(\mathcal{A})^{\mathcal{X}}$ *be any expert-agnostic policy class. If* $\{\pi^b : b \in \mathcal{B}_n\} \subseteq \Pi$*, then, for every* $\varepsilon \in (0, 1/2)$*,*

$$\log_2 \mathcal{N}_\varepsilon(\Pi, d_\Pi) \ge 2^n - n - 1.$$

Proof of Theorem C.2. The validity of the ExBMDP construction and the fact that each π^b is deterministic and endogenous follow from Appendix C.2.2. The return gap is also established in that section. The identity $Q^{\pi^b}_{\mathcal{M}_b} = Q^{1,1} \in \mathcal{Q}$ for every $b \in \mathcal{B}_n$ follows from Lemma C.2. The covering-number bound for $\mathcal{Q}$ follows from Lemma C.3. The family-level policy lower bound follows from Lemma C.4. □

Why the policy needs the decoder but the Q-function does not. At layer 1, for any observation w, the expert's action is $\pi^b_1(w) = b(w)$. Thus exact behavioral cloning must recover the decoder b. However, the expert Q-function at the same observation is $Q^{\pi^b}_{\mathcal{M}_b,1}(w, 0) = Q^{\pi^b}_{\mathcal{M}_b,1}(w, 1) = 1/2$, and $Q^{\pi^b}_{\mathcal{M}_b,1}(w, 2) = 0$, which is independent of $b(w)$. The decoder determines which safe action the expert chooses, but both safe actions have the same value. This is why it is easier to represent the expert's Q-function than to represent the expert's policy.

Here the expert is strictly suboptimal because it chooses action 1 rather than action 0 at the second layer. This does not affect the tie-breaking mechanism at the first layer: actions 0 and 1 have the same expert Q-value, and recovering the exact expert policy still requires the decoder-dependent rule $w \mapsto b(w)$. Thus a simple Q-function can correspond to an enormous set of expert policies even when those experts are suboptimal. Making the expert's first-layer action uniquely optimal in this construction would force Q^{π^b} to encode the decoder, eliminating the present decoder-independent witness class.

Why the exogenous structure matters. If the endogenous state s were observed directly, the expert would be the constant-complexity policy $\bar\pi^b_1(s) = s$ and $\bar\pi^b_2(y) = 1$ on an endogenous state space of size 7. The controlled dynamics and rewards are also simple in endogenous coordinates.

The complexity appears only because the observation map hides the endogenous bit inside exogenous variation, $w = \eta^b_s(\xi) \in \{0, 1\}^n$. Recovering the endogenous bit from the raw observation requires computing $s = b(w)$. Thus the observation-based expert policy factors as $\pi^b_1(w) = \bar\pi^b_1(\varphi^\star_{1,b}(w)) = b(w)$. Expert-policy realizability therefore requires representing the hard decoder. By contrast, Q^{π^b} is identical across decoders. This is why the same two-dimensional Q-class works uniformly for all decoders b, whereas any policy class that realizes all corresponding experts must be exponentially large.

D Experiment Details

This section includes additional experimental results and details on the experiments presented in the main text. To start with, we include some experiments in the instance we use for our lower bound to showcase the failure of offline value-based IL under Q^{π_E}-realizability and no additional assumptions. As a representative offline algorithm that outputs a VI policy, we use SPOIL (Moulin et al., 2025a). These results are described in Appendix D.1. Next, in Appendix D.2 we provide additional details and comments about the Gymnasium experiments.

D.1 Experiments on the Lower Bound Instance

We start by investigating the performance of an offline VI-policies-based algorithm, such as SPOIL. Our negative result in Theorem G.1 predicts a constant suboptimality in at least one environment. To observe this prediction in practice, we repeat the experiment 500 times, each time letting the adversary choose uniformly between the two environments $\mathcal{M}_1$ and $\mathcal{M}_2$ in the hard family used in the proof of Theorem G.1. We report their average suboptimality in Figure 3, which is a lower bound on the worst-case suboptimality. Indeed, at least one of $\mathcal{M}_1$ and $\mathcal{M}_2$ has suboptimality no smaller than the average. As Figure 3 shows, OVI reaches zero suboptimality, whereas SPOIL remains at the constant suboptimality predicted by our construction.

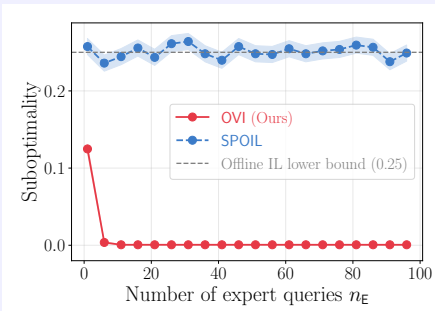


Figure 3: Experiment on the environment class used to prove Theorem G.1.

D.2 Implementation with Neural Networks

For the Gymnasium experiments, we consider a stationary approximation of Algorithm 1. At the beginning of each epoch, we label one trajectory collected with the learner policy output by the previous epoch, π^{ℓ}, rather than from $d_h^{\pi_{\text{out}}}$ as our theory for finite horizon would prescribe. We label such trajectories with expert actions. At this point, we compute the empirical approximation of the objective used by OVI and compute an approximate saddle point by K iterations of the online learning algorithms. The resulting pseudocode is given in Algorithm 2. For the experiments, we set $K = 50$ while we vary the maximum number of queries along the learner trajectories as shown in Figure 4. The exponential weights update for the policy is approximated via Soft DQN (Haarnoja et al., 2017) as in Moulin et al. (2025a). Moreover, we have observed that OVI's performance improves when the state dataset is formed by aggregating both expert and learner trajectories. Such data mixing is common in on-policy distillation (Agarwal et al., 2024) and we prove that it extends the representational benefits of interaction; see Appendix L.

In the easiest environments like CartPole-v1 and Acrobot-v1, we find that 10 trajectories collected offline (for SPOIL and BC) or online and then labeled by the expert (for OVI and DAgger) usually suffice to match the expert performance. In Figure 4, in more complicated environments such as LunarLander-v2 and Pendulum-v1, we increased the number of expert samples, but we did not find it necessary to increase the value of K.

As baselines, we use DAgger (Ross et al., 2011), an interactive method requiring policy realizability. We also use standard BC (Pomerleau, 1991), an offline method requiring expert-policy realizability, and SPOIL (Moulin et al., 2025a), an offline method that does not require interactive expert access but assumes the stronger $Q^{\Pi_{\mathcal{Q}}}$-realizability condition, namely, $Q^{\pi} \in \mathcal{Q}$ for every $\pi \in \Pi_{\mathcal{Q}}$.

D.3 Additional Details and Comments on the Gym Experiments

We ran the above algorithms in some Gymnasium environments and report the results next. In particular, we first trained an expert network with 2 hidden layers with 64 neurons each via DQN (Mnih et al., 2015) in Pendulum-v1 and PPO (Schulman et al., 2017) in CartPole-v1 and Acrobot-v1 and then considered the following two settings. First, we considered a learner parameterized via exactly

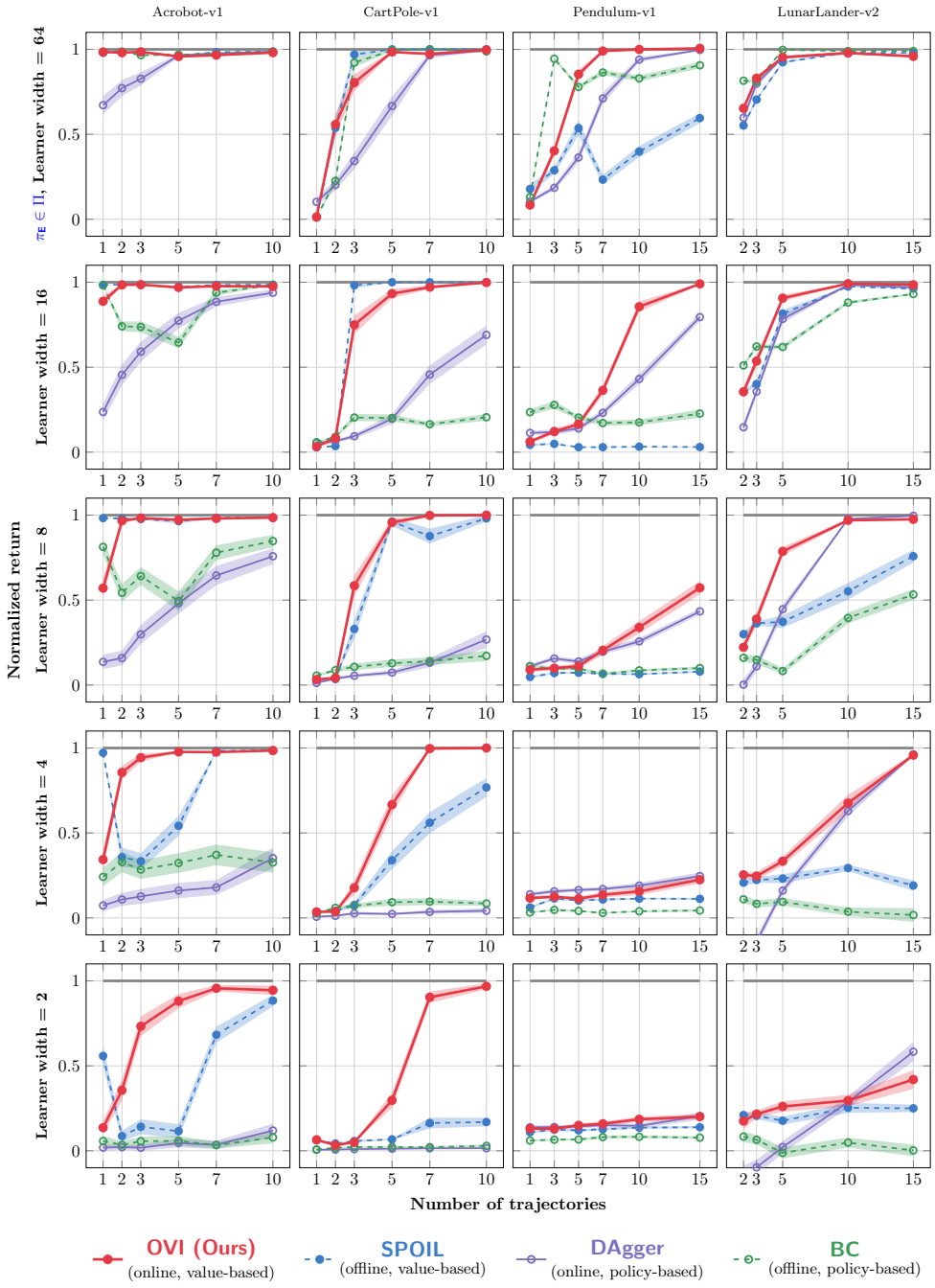


Figure 4: Experiments in the `Gymnasium` library. The top row reports the setting in which the expert and learner have the same network architecture. In the second and subsequent rows, the learner's network width decreases, taking the values $\{16, 8, 4, 2\}$ neurons. The x-axis reports the number of expert trajectories for offline methods (`BC` and `SPOIL`) and the number of learner trajectories labeled with expert actions for `DAgger` and `OVI`.

Algorithm 2 Stationary approximation of `OVI` for neural networks.

Initialize policy weights ψ_0^K and value network weights θ_0^K.
for $\ell = 1, \dots, n_{\mathsf{E}}$ **do**
  Collect a state trajectory $(\boldsymbol{x}_{h,\ell})_{h=1}^H$ with the policy $\pi_{\psi_{\ell-1}^K}$.
  Label the states with expert actions $\boldsymbol{a}_{h,\ell}^{\mathsf{E}} \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_{h,\ell})$ for each $h \in [H]$.
  `# Warm-start initialization of policy and Q network parameters.`
  $\theta_\ell^1 = \theta_{\ell-1}^K$, $\psi_\ell^1 = \psi_{\ell-1}^K$
  for $k = 1, \dots, K-1$ **do**
    `# Update the policy weights.`
    Update ψ_ℓ^k via SoftDQN (Haarnoja et al., 2017) using $Q_{\theta_\ell^k}$ as the Q-network.
    `# The argmax is approximated with several Adam iterations initialized at` θ_ℓ^k.
    $\theta_\ell^{k+1} \approxeq \arg\max_{\theta \in \mathbb{R}^d} \sum_{s=1}^{\ell} \sum_{h=1}^{H} \Big(Q_\theta(\boldsymbol{x}_{h,s}, \boldsymbol{a}_{h,s}^{\mathsf{E}}) - Q_\theta(\boldsymbol{x}_{h,s}, \pi_{\psi_\ell^k}) \Big)$.
Output the policy $\pi_{\psi_{n_{\mathsf{E}}}^K}$.

the same architecture (2 hidden layers with 64 neurons). In this case, expert realizability clearly holds. Therefore, `BC` and `DAgger` are expected to work well. Second, we considered smaller learner networks (2 hidden layers with 16, 8, 4, or 2 neurons per layer). Expert realizability is a poorly motivated assumption in this second case, and we therefore expected methods like `SPOIL` and `OVI` to be more efficient as the expert–learner size gap increases.

The aforementioned hypotheses are indeed confirmed by our experiments in `CartPole-v1` and `Acrobot-v1`. When expert realizability holds (top row of Figure 4), we see that all four methods are essentially equivalent and all eventually manage to match the performance of the expert policy. By contrast, when the learner network becomes smaller, we see that the methods that avoid the expert-realizability assumption achieve better performance. Moreover, the benefits of interaction become evident for the largest size gap we tested (in the last row of Figure 4) in which `OVI` outperforms `SPOIL`. This suggests, therefore, that the network parametrization considered here is not powerful enough to represent the Q-functions of the policy sequence generated by the algorithm. An interesting additional observation is that in `CartPole-v1` and `Acrobot-v1`, value-based methods clearly outperform policy-based ones. In contrast, in `Pendulum-v1` and `LunarLander-v2`, the interactive methods, i.e., `OVI` and `DAgger`, are the best ones. Therefore, for some environments it seems that interaction is key for performance, while in others value-based representation is more important. We think that such *instance-dependent* phenomena raise interesting open questions.

The `Pendulum-v1` environment naturally comes with continuous actions, but we adapted it to our setting via action discretization, a technique used in practice (Dadashi et al., 2022) and recently analyzed by Cao and Xie (2026). Notably, in this case, we notice that `OVI` outperforms the standard `SPOIL`, potentially because `OVI` imposes a weaker requirement on the representational power of the $\mathcal{Q}$-class. The experts achieved the following average returns: 500 in `CartPole-v1`, -77.1 in `Acrobot-v1`, 270 in `LunarLander-v2` and -176.6 in `Pendulum-v1`. The zero point on the y-axis is set to the following returns achieved by very suboptimal policies: -500 in `Acrobot-v1`, 8 in `CartPole-v1`, -1650.0 in `Pendulum-v1` and -600 in `LunarLander-v2`. For each environment, we ran each IL algorithm 50 times with different random seeds from 0 to 49. As is common practice in Gymnasium environments (Garg et al., 2021), we subsampled the expert trajectories for offline methods and retained expert-action labels along learner trajectories only at a fixed frequency for interactive methods. For sample accounting, each retained label counts as one expert sample for offline methods and one expert query for interactive methods. Thus, a trajectory of length T contributes $\lceil T/f \rceil$ samples or queries at subsampling frequency f.

For all algorithms, we tuned their hyperparameters to optimize the performance with no expert–learner gap and then kept the same hyperparameters for the experiments with smaller learner network sizes. The detailed list of hyperparameters is in the `README` file accompanying our code. For `SPOIL` and `OVI`, we approximate the policy update via `Soft DQN` learning over the Q-values, not from rewards, but with the loss motivated by our theoretical analysis. The experiments can be run on CPUs. We ran them on a personal laptop. Note, however, that running all the experiments can take several hours.

E Technical Tools

E.1 Concentration

Lemma E.1 (Azuma–Hoeffding inequality; Boucheron et al., 2013, Section 1.1)**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_i)_{i=1}^n$, i.e., $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0$ and $|X_i| \le M$ almost surely for all $i \in [n]$. Then, for any $t > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > t\right] \le 2e^{-\frac{nt^2}{2M^2}}.$$

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

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le M\sqrt{\frac{2\log(2/\delta)}{n}}.$$

We recap the definition of covering numbers.

Definition E.1 (Covering number)**.** *Let (M, d) be a metric space, S be a subset of M, and $\epsilon > 0$. A set $\mathcal{C}_\epsilon(S, \mathrm{d})$ is an ϵ-covering of S if for any $x \in S$, there exists $y \in \mathcal{C}_\epsilon(S, \mathrm{d})$ such that $\mathrm{d}(x, y) \le \epsilon$. The covering number of S, $\mathcal{N}_\epsilon(S, \mathrm{d})$, is the minimum cardinality of any such covering of S. Moreover, we denote $\mathcal{C}_\epsilon(S) := \mathcal{C}_\epsilon(S, \|\cdot\|_\infty)$.*

We state the finite-horizon concentration lemma that will be used for the uniform covering argument. For decision rules $p, p' \colon \mathcal{X} \to \Delta(\mathcal{A})$, define the policy metric

$$\|p - p'\|_{\infty,1} := \max_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} |p(a \mid x) - p'(a \mid x)|. \tag{3}$$

Given two functions $Q, Q' \in \{f \colon \mathcal{X} \times \mathcal{A} \to [0, Q_{\mathsf{max}}]\}$, and two decision rules $p, p' \in \Delta(\mathcal{A})^{\mathcal{X}}$, define the scaled product metric

$$\rho((Q,p),(Q',p')) := \|Q - Q'\|_\infty + Q_{\mathsf{max}}\|p - p'\|_{\infty,1}. \tag{4}$$

Lemma E.2 (Finite-horizon uniform concentration; analogue of Moulin et al., 2025a, Lemma 5)**.** *Fix nonempty deterministic classes $\Pi_h \subseteq \Delta(\mathcal{A})^{\mathcal{X}}$ and $\mathcal{Q}_h \subseteq \{Q_h \colon \mathcal{X} \times \mathcal{A} \to [0, Q_{\mathsf{max}}]\}$ for every $h \in [H]$. For each h, let $\mathcal{F}_{h-1}$ denote the history available before the samples for stage h are drawn. Let $d_h \in \Delta(\mathcal{X})$ be $\mathcal{F}_{h-1}$-measurable, and suppose that, conditionally on $\mathcal{F}_{h-1}$,*

$$\boldsymbol{x}_h^i \sim d_h, \qquad \boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}\big(\cdot \,\big|\, \boldsymbol{x}_h^i\big), \qquad i \in [n_{\mathsf{E}}],$$

are independent. For any $p \in \Pi_h$ and $Q_h \in \mathcal{Q}_h$, define

$$\mathcal{L}_h^d(p, Q_h) := \sum_{x \in \mathcal{X}} d_h(x)\langle Q_h(x,\cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - p(\cdot \mid x)\rangle,$$

$$\widehat{\mathcal{L}}_h^d(p, Q_h) := \frac{1}{n_{\mathsf{E}}}\sum_{i=1}^{n_{\mathsf{E}}}\big(Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - Q_h\big(\boldsymbol{x}_h^i, p\big)\big).$$

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

$$\max_{h \in [H]} \sup_{p \in \Pi_h} \sup_{Q_h \in \mathcal{Q}_h} \left|\widehat{\mathcal{L}}_h^d(p, Q_h) - \mathcal{L}_h^d(p, Q_h)\right| \le 4r + Q_{\mathsf{max}}\sqrt{\frac{8\log\Big(2\sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h \times \Pi_h, \rho)/\delta\Big)}{n_{\mathsf{E}}}}. \tag{5}$$

If $\sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h \times \Pi_h, \rho) = \infty$, the bound is interpreted as vacuous.

Proof of Lemma E.2. The result is vacuous when the total covering number is infinite. Thus, assume that

$$\sum_{h=1}^{H} \mathcal{N}_r(\mathcal{Q}_h \times \Pi_h, \rho) < \infty.$$

Fix h and condition on $\mathcal{F}_{h-1}$. For a fixed pair $(Q_h, p) \in \mathcal{Q}_h \times \Pi_h$, define

$$Z_i(Q_h, p) = Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - Q_h\big(\boldsymbol{x}_h^i, p\big) - \mathcal{L}_h^d(p, Q_h).$$

Then $(Z_i(Q_h, p))_{i=1}^{n_\mathsf{E}}$ is a conditionally independent mean-zero sequence and $|Z_i(Q_h, p)| \le 2Q_{\mathsf{max}}$ almost surely. By the Azuma–Hoeffding inequality (Lemma E.1), for every $t > 0$,

$$\mathbb{P}\Big[\Big|\widehat{\mathcal{L}}_h^d(p, Q_h) - \mathcal{L}_h^d(p, Q_h)\Big| \ge t \,\Big|\, \mathcal{F}_{h-1}\Big] \le 2\exp\Big(-\frac{n_\mathsf{E} t^2}{8Q_{\mathsf{max}}^2}\Big). \tag{6}$$

For each h, choose a finite set $\mathcal{C}_{h,r} \subseteq \mathcal{Q}_h \times \Pi_h$ with cardinality $\mathcal{N}_r(\mathcal{Q}_h \times \Pi_h, \rho)$ such that every pair in $\mathcal{Q}_h \times \Pi_h$ lies within distance r of some element of $\mathcal{C}_{h,r}$ under ρ. Applying Eq. (6) to all elements of these finite sets, then taking expectations and a union bound over $h \in [H]$, gives an event of probability at least $1 - \delta$ on which, for all h and all $(\widetilde{Q}_h, \widetilde{p}) \in \mathcal{C}_{h,r}$,

$$\Big|\widehat{\mathcal{L}}_h^d\Big(\widetilde{p}, \widetilde{Q}_h\Big) - \mathcal{L}_h^d\Big(\widetilde{p}, \widetilde{Q}_h\Big)\Big| \le Q_{\mathsf{max}}\sqrt{\frac{8\log\Big(2\sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h \times \Pi_h, \rho)/\delta\Big)}{n_\mathsf{E}}}. \tag{7}$$

We work on this event and pass from the cover to the full class. Fix h, $Q_h \in \mathcal{Q}_h$, and $p \in \Pi_h$, and choose $(\widetilde{Q}_h, \widetilde{p}) \in \mathcal{C}_{h,r}$ such that

$$\Big\|Q_h - \widetilde{Q}_h\Big\|_\infty + Q_{\mathsf{max}}\|p - \widetilde{p}\|_{\infty,1} \le r.$$

By the triangle inequality,

$$\begin{aligned}\Big|\widehat{\mathcal{L}}_h^d(p, Q_h) - \widehat{\mathcal{L}}_h^d\Big(\widetilde{p}, \widetilde{Q}_h\Big)\Big| \le{}& \frac{1}{n_\mathsf{E}}\sum_{i=1}^{n_\mathsf{E}}\Big|Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - \widetilde{Q}_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big)\Big| \\ &+ \frac{1}{n_\mathsf{E}}\sum_{i=1}^{n_\mathsf{E}}\Big|Q_h\big(\boldsymbol{x}_h^i, p\big) - \widetilde{Q}_h\big(\boldsymbol{x}_h^i, \widetilde{p}\big)\Big|.\end{aligned}$$

For every state x,

$$\begin{aligned}\Big|Q_h(x,p) - \widetilde{Q}_h(x,\widetilde{p})\Big| &\le \Big|Q_h(x,p) - \widetilde{Q}_h(x,p)\Big| + \Big|\widetilde{Q}_h(x,p) - \widetilde{Q}_h(x,\widetilde{p})\Big| \\ &\le \Big\|Q_h - \widetilde{Q}_h\Big\|_\infty + Q_{\mathsf{max}}\|p - \widetilde{p}\|_{\infty,1}.\end{aligned}$$

Therefore,

$$\Big|\widehat{\mathcal{L}}_h^d(p, Q_h) - \widehat{\mathcal{L}}_h^d\Big(\widetilde{p}, \widetilde{Q}_h\Big)\Big| \le 2\Big\|Q_h - \widetilde{Q}_h\Big\|_\infty + Q_{\mathsf{max}}\|p - \widetilde{p}\|_{\infty,1} \le 2r. \tag{8}$$

The same deterministic calculation, with empirical averages replaced by expectation under $x \sim d_h$ and $a \sim \pi_{\mathsf{E},h}(\cdot \mid x)$, gives

$$\Big|\mathcal{L}_h^d(p, Q_h) - \mathcal{L}_h^d\Big(\widetilde{p}, \widetilde{Q}_h\Big)\Big| \le 2r. \tag{9}$$

Combining Eqs. (7) to (9) yields Eq. (5). Since h, p, and Q_h were arbitrary, the proof is complete. □

For later use, we also record the covering bounds for the decision-rule classes induced by exponentiated-gradient updates. For function classes $\mathcal{Q}_h \subseteq \{Q_h \colon \mathcal{X} \times \mathcal{A} \to [0, Q_{\mathsf{max}}]\}$, an integer $K \ge 0$, and $\eta > 0$, define

$$\Pi_\mathcal{Q} := \Big\{\pi : \exists m \in \{0, \dots, K\}, \forall h, \exists Q_h^1, \dots, Q_h^m \in \mathcal{Q}_h, \quad \pi_h(\cdot \mid x) = \mathrm{softmax}\Big(\eta \textstyle\sum_j Q_h^j(x, \cdot)\Big)\Big\}.$$

For each h, let $\Pi_{\mathcal{Q},h} := \{\pi_h : \pi \in \Pi_\mathcal{Q}\}$. When $m = 0$, the empty sum is the zero vector, so the corresponding decision rule is uniform.

Lemma E.3 (Nonconvex product covering; analogue of Moulin et al., 2025a, Lemma 7)**.** *For every $h \in [H]$ and every $\varepsilon > 0$,*

$$\mathcal{N}_\varepsilon\Big(\Pi_{\mathcal{Q},h}, \|\cdot\|_{\infty,1}\Big) \le 1 + \sum_{m=1}^K \mathcal{N}_{\varepsilon/(\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty)^m. \tag{10}$$

Consequently, for the scaled product metric ρ in Eq. (4), for every $r > 0$,

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le \inf_{\alpha \in (0,r)} \mathcal{N}_\alpha(\mathcal{Q}_h, \|\cdot\|_\infty) \left(1 + \sum_{m=1}^{K} \mathcal{N}_{(r-\alpha)/(Q_{\max}\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty)^m\right). \tag{11}$$

In particular, the simpler bound

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le (K+1)\mathcal{N}_{\frac{r}{2\max\{1, Q_{\max}\eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty)^{K+1} \tag{12}$$

also holds.

Proof of Lemma E.3. Fix h and $m \in \{0, \dots, K\}$. If $m = 0$, the only decision rule represented with m functions is the uniform decision rule. Suppose now that $m \ge 1$, and let $\mathcal{C}_s$ be an s-covering of $\mathcal{Q}_h$ under $\|\cdot\|_\infty$. Consider any $p \in \Pi_{\mathcal{Q},h}$ represented with this value of m by $Q_h^1, \dots, Q_h^m \in \mathcal{Q}_h$, and choose $\widetilde{Q}_h^j \in \mathcal{C}_s$ such that $\|Q_h^j - \widetilde{Q}_h^j\|_\infty \le s$ for all $j \in [m]$. Define

$$\widetilde{p}(\cdot \mid x) = \mathrm{softmax}\left(\eta \sum_{j=1}^{m} \widetilde{Q}_h^j(x, \cdot)\right).$$

By 1-Lipschitz continuity of the softmax (Lemma E.8) and the triangle inequality, for every $x \in \mathcal{X}$,

$$\|p(\cdot \mid x) - \widetilde{p}(\cdot \mid x)\|_1 \le \eta \sum_{j=1}^{m} \left\|Q_h^j - \widetilde{Q}_h^j\right\|_\infty \le \eta m s,$$

where we used the definitions of $\widetilde{Q}_h^1, \dots, \widetilde{Q}_h^m$ in the last inequality. Taking $s = \varepsilon/(\eta m)$ gives a ε-covering for the decision rules represented with this value of m, of cardinality at most $\mathcal{N}_{\varepsilon/(\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty)^m$. Taking the union over $m = 0, \dots, K$ proves Eq. (10).

For the product bound, fix $\alpha \in (0, r)$. Take an α-covering of $\mathcal{Q}_h$ under $\|\cdot\|_\infty$ and an $(r-\alpha)/Q_{\max}$-covering of $\Pi_{\mathcal{Q},h}$ under $\|\cdot\|_{\infty,1}$. By definition of ρ, their Cartesian product is an r-covering of $\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}$ under ρ. Applying Eq. (10) with $\varepsilon = (r-\alpha)/Q_{\max}$, and then taking the infimum over α, gives Eq. (11).

To obtain Eq. (12), take $\alpha = r/2$ in Eq. (11). If $K = 0$, the claim follows immediately. Suppose then that $K \ge 1$. The radius $r/(2\max\{1, Q_{\max}\eta K\})$ is at most $r/2$ and at most $r/(2Q_{\max}\eta m)$ for every $1 \le m \le K$. Thus, by monotonicity of covering numbers,

$$\mathcal{N}_{r/2}(\mathcal{Q}_h, \|\cdot\|_\infty) \le \mathcal{N}_{\frac{r}{2\max\{1, Q_{\max}\eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty),$$
$$\mathcal{N}_{\frac{r}{2Q_{\max}\eta m}}(\mathcal{Q}_h, \|\cdot\|_\infty)^m \le \mathcal{N}_{\frac{r}{2\max\{1, Q_{\max}\eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty)^K.$$

Since covering numbers of nonempty classes are at least one, Eq. (12) follows. □

Lemma E.4 (Convex product covering; analogue of Moulin et al., 2025a, Lemma 8)**.** *Assume that each $\mathcal{Q}_h$ is convex. Then, for every $h \in [H]$ and every $\varepsilon > 0$,*

$$\mathcal{N}_\varepsilon\left(\Pi_{\mathcal{Q},h}, \|\cdot\|_{\infty,1}\right) \le 1 + \sum_{m=1}^{K} \mathcal{N}_{\varepsilon/(\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty). \tag{13}$$

Consequently, for the scaled product metric ρ in Eq. (4), for every $r > 0$,

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le \inf_{\alpha \in (0,r)} \mathcal{N}_\alpha(\mathcal{Q}_h, \|\cdot\|_\infty) \left(1 + \sum_{m=1}^{K} \mathcal{N}_{(r-\alpha)/(Q_{\max}\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty)\right). \tag{14}$$

In particular, the simpler bound

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le (K+1)\mathcal{N}_{\frac{r}{2\max\{1, Q_{\max}\eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty)^2 \tag{15}$$

also holds.

Proof of Lemma E.4. Fix h and $m \in \{0, \dots, K\}$. The case $m = 0$ again gives the uniform decision rule. Suppose that $m \geq 1$, and let $p \in \Pi_{\mathcal{Q},h}$ be represented with this value of m by $Q_h^1, \dots, Q_h^m \in \mathcal{Q}_h$. Define

$$\overline{Q}_h = \frac{1}{m} \sum_{j=1}^{m} Q_h^j.$$

By convexity, $\overline{Q}_h \in \mathcal{Q}_h$, and $p(\cdot \,|\, x) = \mathrm{softmax}(\eta m \overline{Q}_h(x, \cdot))$ for every $x \in \mathcal{X}$. Let $\mathcal{C}_s$ be an s-covering of $\mathcal{Q}_h$ under $\|\cdot\|_\infty$, choose $\widetilde{Q}_h \in \mathcal{C}_s$ such that $\|\overline{Q}_h - \widetilde{Q}_h\|_\infty \leq s$, and define

$$\widetilde{p}(\cdot \,|\, x) = \mathrm{softmax}\Big(\eta m \widetilde{Q}_h(x, \cdot)\Big).$$

By 1-Lipschitz continuity of the softmax (Lemma E.8) and the definition of $\widetilde{Q}_h$, for every $x \in \mathcal{X}$,

$$\|p(\cdot \,|\, x) - \widetilde{p}(\cdot \,|\, x)\|_1 \leq \eta m \Big\|\overline{Q}_h - \widetilde{Q}_h\Big\|_\infty \leq \eta m s.$$

Taking $s = \varepsilon/(\eta m)$ gives a ε-covering for the decision rules represented with this value of m, of cardinality at most $\mathcal{N}_{\varepsilon/(\eta m)}(\mathcal{Q}_h, \|\cdot\|_\infty)$. Taking the union over $m = 0, \dots, K$ proves Eq. (13). The product bound follows from the same Cartesian-product argument used in the proof of Lemma E.3.

To obtain Eq. (15), take $\alpha = r/2$ in Eq. (14). If $K = 0$, the claim follows immediately. Suppose then that $K \geq 1$. The radius $r/(2\max\{1, Q_{\mathtt{max}} \eta K\})$ is at most $r/2$ and at most $r/(2Q_{\mathtt{max}} \eta m)$ for every $1 \leq m \leq K$. Thus, by monotonicity of covering numbers,

$$\mathcal{N}_{r/2}(\mathcal{Q}_h, \|\cdot\|_\infty) \leq \mathcal{N}_{\frac{r}{2\max\{1, Q_{\mathtt{max}} \eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty),$$
$$\mathcal{N}_{\frac{r}{2Q_{\mathtt{max}} \eta m}}(\mathcal{Q}_h, \|\cdot\|_\infty) \leq \mathcal{N}_{\frac{r}{2\max\{1, Q_{\mathtt{max}} \eta K\}}}(\mathcal{Q}_h, \|\cdot\|_\infty).$$

Since covering numbers of nonempty classes are at least one, Eq. (15) follows. □

We also recall the following standard bound.

Lemma E.5 (Covering number of an ℓ_∞ ball; Vershynin, 2018, Section 4.2)**.** *Let $d \in \mathbb{N}$, $R > 0$, and let $\mathfrak{B}_\infty^d(R)$ denote the ℓ_∞ ball of radius R in $\mathbb{R}^d$. Then, for every $\varepsilon > 0$,*

$$\mathcal{N}_\varepsilon\big(\mathfrak{B}_\infty^d(R), \|\cdot\|_\infty\big) \leq \left(\left\lceil \frac{2R}{\varepsilon} \right\rceil + 1\right)^d.$$

E.2 Optimization

We will need the following regret bounds for mirror descent and follow-the-regularized-leader (FTRL), which are standard in the literature. We report them here for completeness.

Lemma E.6 (Simplified version of Orabona, 2026, Theorem 6.11)**.** *Let us consider a non-empty closed convex set V, an arbitrary sequence of adaptively chosen loss vectors $(\ell_k)_{k=1}^K$ such that $\|\ell_k\|_\infty \leq \ell_{\max}$, and let $D \colon V \times \mathrm{int}(V) \to \mathbb{R}$ be a Bregman divergence induced by a λ-strongly convex function in the ℓ_1 norm. Then, for all $u \in V$, the sequence $(x_k)_{k=1}^K$ generated for each k by*

$$x_{k+1} = \arg\min_{v \in V} \left\{ \langle \ell_k, v \rangle + \frac{1}{\eta} D(v, x_k) \right\}$$

for an initial $x_1 \in \mathrm{int}(V)$ and under the assumption that $x_k \in \mathrm{int}(V)$ for every $k \in [K+1]$ satisfies

$$\sum_{k=1}^{K} \langle \ell_k, x_k - u \rangle \leq \frac{D(u, x_1)}{\eta} + \frac{\eta K \ell_{\max}^2}{2\lambda}.$$

Lemma E.7 (FTRL over the simplex)**.** *Let $\Delta(\mathcal{A})$ be the simplex over a discrete action space $\mathcal{A}$, and for any $x \in \Delta(\mathcal{A})$, define the Shannon entropy as $H(x) = -\sum_{a \in \mathcal{A}} x(a) \log x(a)$. Consider the sequence $(x_k)_{k=1}^K$ generated via FTRL using the potential*

$$\psi_k(x) = \frac{-H(x) - \min_{x' \in \Delta(\mathcal{A})}(-H(x'))}{\eta_k}.$$

That is, for a certain sequence of cost vectors $\ell_1, \dots, \ell_K$ we have

$$x_k = \arg\min_{x \in \Delta(\mathcal{A})} \left\langle x, \sum_{k'=1}^{k-1} \ell_{k'} \right\rangle + \psi_k(x) = \frac{e^{-\eta_k \sum_{k'=1}^{k-1} \ell_{k'}}}{\sum_{a \in \mathcal{A}} e^{-\eta_k \sum_{k'=1}^{k-1} \ell_{k'}(a)}} \,.$$

Then, if $\|\ell_k\|_\infty \le \ell_{\max}$ and $\eta_k = \sqrt{\frac{\log(A)}{\ell_{\max}^2 k}}$, it holds that for any $K \in \mathbb{N}$ and any $x^\star \in \Delta(\mathcal{A})$,

$$\sum_{k=1}^{K} \langle \ell_k, x_k - x^\star \rangle \le 3\sqrt{\ell_{\max}^2 \log|\mathcal{A}| K}.$$

Proof of Lemma E.7. Let us define the function $F_k(x) = \sum_{k'=1}^{k-1} \langle \ell_{k'}, x \rangle + \psi_k(x)$. Then, by definition, we have

$$-\sum_{k=1}^{K} \langle \ell_k, x^\star \rangle = \psi_{K+1}(x^\star) - F_{K+1}(x^\star) \,.$$

Adding and subtracting $F_{K+1}(x_{K+1})$ and $F_1(x_1)$ and noticing that x_1 minimizes $F_1 = \psi_1$ over the simplex, we obtain

$$-\sum_{k=1}^{K} \langle \ell_k, x^\star \rangle = \psi_{K+1}(x^\star) - F_{K+1}(x^\star) - F_{K+1}(x_{K+1}) + F_{K+1}(x_{K+1}) - \min_{x \in \Delta(\mathcal{A})} \psi_1(x) + F_1(x_1).$$

Now, writing the difference $F_1(x_1) - F_{K+1}(x_{K+1})$ as a telescoping sum, we have

$$\begin{aligned} -\sum_{k=1}^{K} \langle \ell_k, x^\star \rangle = \psi_{K+1}(x^\star) - F_{K+1}(x^\star) + \sum_{k=1}^{K} (F_k(x_k) - F_{k+1}(x_{k+1})) \\ + F_{K+1}(x_{K+1}) - \min_{x \in \Delta(\mathcal{A})} \psi_1(x) \,, \end{aligned}$$

Finally, adding $\sum_{k=1}^{K} \langle \ell_k, x_k \rangle$ to both sides, we obtain

$$\begin{aligned} \sum_{k=1}^{K} \langle \ell_k, x_k - x^\star \rangle = \psi_{K+1}(x^\star) - F_{K+1}(x^\star) + \sum_{k=1}^{K} (F_k(x_k) - F_{k+1}(x_{k+1}) + \langle \ell_k, x_k \rangle) \\ + F_{K+1}(x_{K+1}) - \min_{x \in \Delta(\mathcal{A})} \psi_1(x). \end{aligned}$$

Using that $\psi_k(x) \ge 0$ for all k and $x \in \Delta(\mathcal{A})$, and $F_{K+1}(x^\star) \ge F_{K+1}(x_{K+1})$ (by definition of x_{K+1}), we obtain

$$\sum_{k=1}^{K} \langle \ell_k, x_k - x^\star \rangle \le \psi_{K+1}(x^\star) + \sum_{k=1}^{K} (F_k(x_k) - F_{k+1}(x_{k+1}) + \langle \ell_k, x_k \rangle).$$

At this point, note

$$\begin{aligned} &F_k(x_k) - F_{k+1}(x_{k+1}) + \langle \ell_k, x_k \rangle \\ &\overset{(a)}{=} F_k(x_k) + \langle \ell_k, x_k \rangle - (F_{k+1}(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle) + \langle \ell_k, x_{k+1} \rangle \\ &\overset{(b)}{=} F_k(x_k) + \langle \ell_k, x_k \rangle - (F_k(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle) + \langle \ell_k, x_{k+1} \rangle + F_k(x_{k+1}) - F_{k+1}(x_{k+1}) \\ &\overset{(c)}{=} F_k(x_k) + \langle \ell_k, x_k \rangle - (F_k(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle) \\ &\quad + F_{k+1}(x_{k+1}) - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1}) - F_{k+1}(x_{k+1}) \\ &\overset{(d)}{=} F_k(x_k) + \langle \ell_k, x_k \rangle - (F_k(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle) - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1}) \,, \end{aligned}$$

where we (a) added and removed a term $\langle \ell_k, x_{k+1} \rangle$, (b) added and removed a term $F_k(x_{k+1})$, (c) used the fact that $\langle \ell_k, x_{k+1} \rangle + F_k(x_{k+1}) = F_{k+1}(x_{k+1}) - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1})$ by definition of F_k, and (d) rearranged terms.

For any $x \in \mathbb{R}^A$, we define $\iota_{\Delta(\mathcal{A})}(x) = 0$ if $x \in \Delta(\mathcal{A})$ and $\iota_{\Delta(\mathcal{A})}(x) = \infty$ otherwise. Then, since both x_{k+1} and x_k are in $\Delta(\mathcal{A})$, it holds that

$$\begin{aligned}
&F_k(x_k) - F_{k+1}(x_{k+1}) + \langle \ell_k, x_k \rangle \\
&\quad = F_k(x_k) + \langle \ell_k, x_k \rangle + \iota_{\Delta(\mathcal{A})}(x_k) - \big(F_k(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle + \iota_{\Delta(\mathcal{A})}(x_{k+1})\big) \\
&\qquad - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1})\,.
\end{aligned}$$

Since ψ_k is η_k^{-1}-strongly convex, the function $F_k + \langle \ell_k, \cdot \rangle + \iota_{\Delta(\mathcal{A})}$ is also η_k^{-1}-strongly convex. Hence, writing ∂ for the subdifferential, for any subgradient $g_k \in \{\ell_k\} + \partial(F_k + \iota_{\Delta(\mathcal{A})})(x_k)$, we have

$$\begin{aligned}
&F_k(x_k) + \langle \ell_k, x_k \rangle + \iota_{\Delta(\mathcal{A})}(x_k) - \big(F_k(x_{k+1}) + \langle \ell_k, x_{k+1} \rangle + \iota_{\Delta(\mathcal{A})}(x_{k+1})\big) \\
&\qquad \le -\langle g_k, x_{k+1} - x_k \rangle - \frac{1}{2\eta_k} \|x_k - x_{k+1}\|_1^2\,.
\end{aligned}$$

Since $x_k = \arg\min_{x \in \Delta(\mathcal{A})}(F_k + \iota_{\Delta(\mathcal{A})})(x)$, we obtain that $0 \in \partial(F_k + \iota_{\Delta(\mathcal{A})})(x_k)$, which implies that we can choose $g_k = \ell_k$. Plugging this back into the previous display, we obtain that

$$\begin{aligned}
&F_k(x_k) - F_{k+1}(x_{k+1}) + \langle \ell_k, x_k \rangle \\
&\qquad \le \langle \ell_k, x_k - x_{k+1} \rangle - \frac{1}{2\eta_k} \|x_k - x_{k+1}\|_1^2 - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1}) \\
&\qquad \overset{(a)}{\le} \|\ell_k\|_\infty \|x_k - x_{k+1}\|_1 - \frac{1}{2\eta_k} \|x_k - x_{k+1}\|_1^2 - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1}) \\
&\qquad \overset{(b)}{\le} \frac{\eta_k}{2} \|\ell_k\|_\infty^2 - \psi_{k+1}(x_{k+1}) + \psi_k(x_{k+1}) \\
&\qquad \overset{(c)}{\le} \eta_k \|\ell_k\|_\infty^2\,,
\end{aligned}$$

where we also used (a) Hölder's inequality, (b) Young's inequality, and (c) the fact that the potential is nondecreasing. Therefore, the regret bound becomes

$$\sum_{k=1}^K \langle \ell_k, x_k - x^\star \rangle \le \psi_{K+1}(x^\star) + \sum_{k=1}^K \eta_k \|\ell_k\|_\infty^2.$$

By definition of the potential ψ_{K+1} and the facts that $-H(x) \le 0$ and $H(x) \le \log(A)$ for all $x \in \Delta(\mathcal{A})$, we obtain

$$\psi_{K+1}(x) = \frac{-H(x) - \min_{x \in \Delta(\mathcal{A})}(-H(x))}{\eta_{K+1}} \le \frac{\max_{x \in \Delta(\mathcal{A})} H(x)}{\eta_{K+1}} \le \frac{\log(A)}{\eta_{K+1}}\,.$$

Therefore, we obtain

$$\sum_{k=1}^K \langle \ell_k, x_k - x^\star \rangle \le \frac{\log(A)}{\eta_{K+1}} + \sum_{k=1}^K \eta_k \|\ell_k\|_\infty^2.$$

Then, choosing $\eta_k = \sqrt{\frac{\log(A)}{\ell_{\max}^2 k}}$ and using the fact that $\|\ell_k\|_\infty^2 \le \ell_{\max}^2$, we obtain that

$$\sum_{k=1}^K \langle \ell_k, x_k - x^\star \rangle \le \sqrt{\ell_{\max}^2 \log(A)(K+1)} + \sum_{k=1}^K \sqrt{\frac{\ell_{\max}^2 \log(A)}{k}} \le 3\sqrt{\ell_{\max}^2 \log(A) K}.$$

Here, the last inequality follows by factoring out $\sqrt{\ell_{\max}^2 \log(A)}$ and using $\sum_{k=1}^K \frac{1}{\sqrt{k}} \le 1 + \int_1^K \frac{1}{\sqrt{x}}\,\mathrm{d}x = 2\sqrt{K} - 1$, and $\sqrt{K+1} \le \sqrt{K} + 1$. This concludes the proof. □

Lemma E.8 (Softmax Lipschitzness). *For any finite action set $\mathcal{A}$ and any $u, v \in \mathbb{R}^{\mathcal{A}}$,*

$$\|\mathrm{softmax}(u) - \mathrm{softmax}(v)\|_1 \le \|u - v\|_\infty.$$

Proof of Lemma E.8. Since the gradient of the log-sum-exp function is the softmax function, the desired inequality is exactly the 1-smoothness of log-sum-exp with respect to the ℓ_∞ norm, whose dual norm is ℓ_1. By Orabona (2026, Theorem 6.26), this smoothness follows from the 1-strong convexity of the convex conjugate of log-sum-exp with respect to the dual ℓ_1 norm. This conjugate is the negative entropy on the simplex, and its 1-strong convexity in the ℓ_1 norm follows from Orabona (2026, Lemma 6.33). □

E.3 Performance Difference Lemma

We use the following version of the performance difference lemma (Howard, 1960; Kakade and Langford, 2002): it expresses the performance gap between two policies as a sum of advantages of one policy under the occupancy of the other.

Lemma E.9 (Performance difference lemma)**.** *Let π and π' be arbitrary policies. Then,*

$$J^{\pi'} - J^{\pi} = \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi}(x) \sum_{a \in \mathcal{A}} (\pi'_h(a \,|\, x) - \pi_h(a \,|\, x)) Q_h^{\pi'}(x, a). \tag{16}$$

Proof of Lemma E.9. By considering the Bellman equations for the policy π', for any state-action pair (x, a) and any $h \in [H]$ we have that

$$Q_h^{\pi'}(x, a) = r_h(x, a) + \sum_{x' \in \mathcal{X}} P_h(x' \,|\, x, a) V_{h+1}^{\pi'}(x').$$

Then, taking the expectation with respect to the state-action occupancy measure of the policy π, we have

$$\sum_{x,a} d_h^{\pi}(x, a) Q_h^{\pi'}(x, a) = \sum_{x,a} d_h^{\pi}(x, a) r_h(x, a) + \sum_{x,a} d_h^{\pi}(x, a) \sum_{x'} P_h(x' \,|\, x, a) V_{h+1}^{\pi'}(x').$$

Then, by swapping the sums and using the flow conditions satisfied by the occupancy measure of π, i.e., $d_{h+1}^{\pi}(x) = \sum_{x',a'} P_h(x \,|\, x', a') d_h^{\pi}(x', a')$, observe that

$$\begin{aligned}
\sum_{x,a} d_h^{\pi}(x, a) \sum_{x'} P_h(x' \,|\, x, a) V_{h+1}^{\pi'}(x') &= \sum_{x'} \left(\sum_{x,a} P_h(x' \,|\, x, a) d_h^{\pi}(x, a) \right) V_{h+1}^{\pi'}(x') \\
&= \sum_{x'} d_{h+1}^{\pi}(x') V_{h+1}^{\pi'}(x').
\end{aligned}$$

This implies that

$$\left\langle d_h^{\pi}, Q_h^{\pi'} \right\rangle - \left\langle d_h^{\pi}, V_h^{\pi'} \right\rangle = \langle d_h^{\pi}, r_h \rangle + \left\langle d_{h+1}^{\pi}, V_{h+1}^{\pi'} \right\rangle - \left\langle d_h^{\pi}, V_h^{\pi'} \right\rangle.$$

Then, summing over $h \in [H]$ and recalling that $V_{H+1}^{\pi'} = 0$ and that $\nu_0 = d_1^{\pi}$, we obtain that

$$\sum_{h=1}^{H} \left(\left\langle d_h^{\pi}, Q_h^{\pi'} \right\rangle - \left\langle d_h^{\pi}, V_h^{\pi'} \right\rangle \right) = \sum_{h=1}^{H} \langle d_h^{\pi}, r_h \rangle - \left\langle \nu_0, V_1^{\pi'} \right\rangle = J^{\pi} - J^{\pi'}.$$

Finally, rearranging, we have

$$\begin{aligned}
J^{\pi} - J^{\pi'} &= \sum_{h=1}^{H} \left(\left\langle d_h^{\pi}, Q_h^{\pi'} \right\rangle - \left\langle d_h^{\pi}, V_h^{\pi'} \right\rangle \right) \\
&= \sum_{h=1}^{H} \sum_{x} d_h^{\pi}(x) \left(\sum_{a} \pi_h(a \,|\, x) Q_h^{\pi'}(x, a) - V_h^{\pi'}(x) \right) \\
&= \sum_{h=1}^{H} \sum_{x} d_h^{\pi}(x) \sum_{a} (\pi_h(a \,|\, x) - \pi'_h(a \,|\, x)) Q_h^{\pi'}(x, a),
\end{aligned}$$

where we used $d_h^{\pi}(x, a) = d_h^{\pi}(x) \pi_h(a \,|\, x)$ and the definition of $V_h^{\pi'}$. Multiplying by -1 concludes the proof. □

Part I

Proofs of Main Results

In this section, we include the proofs omitted from the main text. Appendix E provides technical results in concentration and optimization that will be used to prove the sample complexity guarantees of Algorithm 1 in Appendix F, and to prove the lower bound in Appendix G.

F Proofs from Section 3

As discussed in Section 3.1, OVI reduces the suboptimality gap to a sequence of saddle-point problems involving the objective $\mathcal{L}_h^d$ and learns the output policy in a layer-wise manner. We recall here only the additional notation needed for the proofs. For any (partial) policy $\pi = (\pi_1, \dots, \pi_{h-1})$, we denote $\mathcal{L}_h^\pi = \mathcal{L}_h^{d^\pi}$ for convenience. When learning layer h, the previously computed decision rules $\pi_{\mathsf{out},1}, \dots, \pi_{\mathsf{out},h-1}$ allow us to sample from $d_h^{\pi_{\mathsf{out}}}$. Thus, for each $i \in [n_{\mathsf{E}}]$, OVI independently samples

$$\boldsymbol{x}_h^i \sim d_h^{\pi_{\mathsf{out}}} \quad \text{and} \quad \boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}\big(\cdot \,\big|\, \boldsymbol{x}_h^i\big).$$

For $h = 1$, we can sample from $d_1^{\pi_{\mathsf{out}}} = \nu_0$. For subsequent stages h, we can sample from $d_h^{\pi_{\mathsf{out}}}$ by rolling in with the previously computed policies $\pi_{\mathsf{out},1}, \dots, \pi_{\mathsf{out},h-1}$. Given these samples, for any decision rule p and function $f \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, we define the empirical estimate of $\mathcal{L}_h^{\pi_{\mathsf{out}}}$ as

$$\widehat{\mathcal{L}}_h(p, f) = \frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} \big(f\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - f\big(\boldsymbol{x}_h^i, p\big)\big).$$

For the version of OVI analyzed in Theorem F.1, the empirical game is solved by having the Q-player best respond to $\widehat{\mathcal{L}}_h(\pi_h^k, \cdot)$, while the π-player uses online mirror ascent (Beck and Teboulle, 2003) with negative entropy as a regularizer. Fixing a stage h, for each k,

$$Q_h^k \in \operatorname*{arg\,max}_{Q_h \in \mathcal{Q}_h} \widehat{\mathcal{L}}_h\big(\pi_h^k, Q_h\big), \qquad \text{and} \qquad \pi_h^{k+1} \propto \pi_h^k \odot \exp\big(\eta Q_h^k\big), \tag{17}$$

where $\eta > 0$ is the learning-rate parameter and $\odot$ denotes the element-wise product.

Having recalled the algorithm structure, we can now move towards proving our main result: interaction enables computationally and statistically efficient value-based imitation learning algorithms such as our OVI.

Before proving Theorem F.1, we derive Lemma F.1, which motivates the π-player update that will be used in the proof of Theorem F.1. It shows that if the Q-player best responds to the π-player, then the suboptimality gap can be bounded by the regret of the π-player plus an estimation error term that can be controlled with a large enough number of expert queries n_{E}.

Lemma F.1. *Let Assumption 2.1 hold. Let policies $\pi_h^1, \dots, \pi_h^K$ and functions $Q_h^1, \dots, Q_h^K \in \mathcal{Q}$ be computed as in Eq. (17). We define the π-regret at stage h and state x as $\mathfrak{R}_h^\pi(x) = \sum_{k=1}^K \langle Q_h^k(x, \cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x)\rangle$. Furthermore, for any policy π, we define the estimation error as $\Delta(\pi) = \max_{h\in[H]} \sup_{Q_h\in\mathcal{Q}_h} |\widehat{\mathcal{L}}_h(\pi_h, Q_h) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h, Q_h)|$, where $\pi_{\mathsf{out},h} = \frac{1}{K}\sum_{k=1}^K \pi_h^k$. Then,*

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K} \sum_{h=1}^H \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \mathfrak{R}_h^\pi(x) + \frac{2H}{K} \sum_{k=1}^K \Delta\big(\pi^k\big).$$

Proof of Lemma F.1. By the performance difference lemma (Lemma E.9) with $\pi = \pi_{\mathsf{out}}$ and $\pi' = \pi_{\mathsf{E}}$, and by the definition of $\mathcal{L}_h^{\pi_{\mathsf{out}}}$, we have $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} = \sum_{h=1}^H \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}})$. Since $\mathcal{L}_h^{\pi_{\mathsf{out}}}(p, Q)$ is affine in the decision rule p and $\pi_{\mathsf{out},h} = \frac{1}{K}\sum_{k=1}^K \pi_h^k$, this gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} = \sum_{h=1}^H \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}) = \frac{1}{K} \sum_{h=1}^H \sum_{k=1}^K \mathcal{L}_h^{\pi_{\mathsf{out}}}\big(\pi_h^k, Q_h^{\pi_{\mathsf{E}}}\big).$$

Then, we can upper bound the suboptimality as follows

$$K(J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}) = \sum_{h,k} \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h^k, Q_h^{\pi_{\mathsf{E}}}) \overset{\text{(a)}}{\leq} \sum_{h,k} \widehat{\mathcal{L}}_h(\pi_h^k, Q_h^{\pi_{\mathsf{E}}}) + H \sum_k \Delta(\pi^k)$$
$$\overset{\text{(b)}}{\leq} \sum_{h,k} \widehat{\mathcal{L}}_h(\pi_h^k, Q_h^k) + H \sum_k \Delta(\pi^k) \overset{\text{(a)}}{\leq} \sum_{h,k} \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h^k, Q_h^k) + 2H \sum_k \Delta(\pi^k),$$

where step (a) uses the definition of $\Delta(\pi^k)$ together with $Q_h^{\pi_{\mathsf{E}}} \in \mathcal{Q}_h$ in the first application and $Q_h^k \in \mathcal{Q}_h$ in the second, and step (b) uses that Q_h^k is a best response to $\widehat{\mathcal{L}}_h(\pi_h^k, \cdot)$ over $\mathcal{Q}_h$. Finally, by expanding the definition of $\mathcal{L}_h^{\pi_{\mathsf{out}}}$ and exchanging the finite sums over k, x, and a,

$$\sum_{k=1}^K \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h^k, Q_h^k) = \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \sum_{k=1}^K \langle Q_h^k(x, \cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x) \rangle = \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \mathfrak{R}_h^{\pi}(x).$$

This concludes the proof. □

Next, we use Lemma F.1 to derive a sample complexity guarantee for OVI. At a high level, the proof intuition is to notice that Lemma F.1 reduces the control of the suboptimality gap to two terms: the average regret of the policy player and the average estimation error of the empirical saddle-point objective. Thus, it suffices to ensure that $\mathfrak{R}_h^{\pi}(x)$ grows sublinearly in K for each stage h and state $x \in \mathcal{X}$, which is achieved by the no-regret update in Algorithm 1. The estimation error term is controlled by a uniform concentration argument over the Q-class and over the policies that can be played by our algorithm. In particular, each iterate belongs to the policy class $\Pi_{\mathcal{Q}} := \{\pi : \exists m \leq K, \forall h, \exists Q_h^1, \ldots, Q_h^m \in \mathcal{Q}_h, \quad \pi_h(a \,|\, x) = \mathrm{softmax}(\sum_k Q_h^k(x, \cdot))_a\}$ for some $K = \mathrm{poly}(H, \log A, Q_{\mathsf{max}}, \varepsilon^{-1})$, and concentration uniformly over $\Pi_{\mathcal{Q}} \times \mathcal{Q}$ controls $\frac{1}{K} \sum_{k=1}^K \Delta(\pi^k)$ with a sufficiently large number of expert queries n_{E}. We make the proof formal in the next subsection.

F.1 Proof of Theorem 3.1 (Sample Complexity Guarantee for OVI)

In this section, we present the proof of our main upper bound.

Theorem 3.1 (Sample complexity of OVI)**.** *Let Assumption 2.1 hold for a finite class $\mathcal{Q}$. Then, for any $\varepsilon, \delta \in (0, 1)$, OVI (Algorithm 1) with parameters $n_{\mathsf{E}} = \widetilde{\mathcal{O}}(H^4 Q_{\mathsf{max}}^4 \log(A) \log(|\mathcal{Q}|/\delta) \varepsilon^{-4})$, $\eta = (\log(A)/(K Q_{\mathsf{max}}^2))^{1/2}$, and $K = \mathcal{O}(H^2 Q_{\mathsf{max}}^2 \log(A) \varepsilon^{-2})$, outputs a policy π_{out} such that with probability at least $1 - \delta$, $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon$ after*

$$\widetilde{\mathcal{O}}\left(\frac{H^5 Q_{\mathsf{max}}^4 \log(A) \log(|\mathcal{Q}|/\delta)}{\varepsilon^4} \right) \qquad \textit{expert queries.}$$

Proof of Theorem 3.1. The proof is a direct consequence of Theorem F.1 below, where the covering number argument simplifies to a union bound over the function class $\mathcal{Q}$. □

Below, we present our general result that extends the sample complexity analysis to more general function classes. In particular, Theorem F.1 shows that the sample complexity of OVI is controlled by the largest ℓ_∞-covering number of $\mathcal{Q}_h$ over $h \in [H]$, and that when $\mathcal{Q}$ is convex, the sample complexity improves to $\widetilde{\mathcal{O}}(H^3 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta) \varepsilon^{-2})$ for some $r_\varepsilon = \mathcal{O}(\varepsilon^2)$. When the classes are infinite, the best responses in the Q-player update need not exist automatically. Throughout this result, we consider classes for which the maximizers in Algorithm 1 are well defined. This holds, for instance, when each $\mathcal{Q}_h$ is compact and the empirical objective optimized by the Q-player is continuous on $\mathcal{Q}_h$.

Theorem F.1 (Sample complexity of OVI)**.** *Let Assumption 2.1 hold. For every $r > 0$, let $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q})$ be the largest ℓ_∞-covering number (see Definition E.1) of $\mathcal{Q}_h$ over $h \in [H]$, and choose $r_\varepsilon = \mathcal{O}(\varepsilon^2/(H^2 Q_{\mathsf{max}} \log A))$. Then, for any $\varepsilon, \delta \in (0, 1)$, OVI (Algorithm 1) with parameters*

$$\eta = \sqrt{\frac{\log(A)}{K Q_{\mathsf{max}}^2}}, K = \mathcal{O}\left(\frac{H^2 Q_{\mathsf{max}}^2 \log(A)}{\varepsilon^2} \right), \text{ and } n_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\frac{H^4 Q_{\mathsf{max}}^4 \log(A) \log\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\big)}{\varepsilon^4} \right),$$

outputs a policy π_{out} such that, with probability at least $1 - \delta$, $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon$ after

$$\widetilde{\mathcal{O}}\left(\frac{H^5 Q_{\mathsf{max}}^4 \log(A) \log\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\big)}{\varepsilon^4} \right) \qquad \textit{expert queries.}$$

Moreover, when $\mathcal{Q}$ is convex, the same guarantee holds with $n_{\mathsf{E}} = \widetilde{\mathcal{O}}(H^2 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta)\varepsilon^{-2})$, and the sample complexity improves to $\widetilde{\mathcal{O}}(H^3 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta)\varepsilon^{-2})$.

Proof of Theorem F.1. To prove the sample complexity guarantees of Algorithm 1, we start from the decomposition given in Lemma F.1.

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K}\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x)\mathfrak{R}_h^{\pi}(x) + \frac{2H}{K}\sum_{k=1}^{K}\Delta\big(\pi^k\big).$$

We need to control the regret of the policy player $\mathfrak{R}_h^{\pi}$, defined for any state x as

$$\mathfrak{R}_h^{\pi}(x) = \sum_{k=1}^{K}\big\langle Q_h^k(x,\cdot), \pi_{\mathsf{E},h}(\cdot\,|\,x) - \pi_h^k(\cdot\,|\,x)\big\rangle,$$

and we need to control the average estimation error $\frac{1}{K}\sum_{k=1}^{K}\Delta(\pi^k)$, where, for any policy π, we define

$$\Delta(\pi) = \max_{h\in[H]}\sup_{Q_h\in\mathcal{Q}_h}\Big|\widehat{\mathcal{L}}_h(\pi_h, Q_h) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h, Q_h)\Big|.$$

Controlling the regret $\mathfrak{R}_h^{\pi}$. In order to control $\mathfrak{R}_h^{\pi}$, we use online mirror ascent with negative-entropy regularizer, which corresponds to the following update rule for each stage $h\in[H]$, iteration $k\in[K]$, state $x\in\mathcal{X}$ and action $a\in\mathcal{A}$,

$$\pi_h^{k+1}(a\,|\,x) \propto \pi_h^k(a\,|\,x)e^{\eta Q_h^k(x,a)}.$$

Fix h and x. Applying Lemma E.6 to the simplex with losses $\ell_k = -Q_h^k(x,\cdot)$ and comparator $u = \pi_{\mathsf{E},h}(\cdot\,|\,x)$ gives

$$\mathfrak{R}_h^{\pi}(x) \le \frac{\log|\mathcal{A}|}{\eta} + \eta K Q_{\mathsf{max}}^2\,,$$

where we used that the initial decision rule is uniform, so the negative-entropy Bregman divergence to any comparator is at most $\log|\mathcal{A}|$, and that each $Q_h\in\mathcal{Q}_h$ satisfies $\|Q_h\|_\infty \le Q_{\mathsf{max}}$, as assumed in Assumption 2.1. Plugging the value of η from the statement of Theorem F.1, $\eta = (\log(A)/K)^{1/2}Q_{\mathsf{max}}^{-1}$, we get that $\mathfrak{R}_h^{\pi}(x) \le 2\sqrt{KQ_{\mathsf{max}}^2\log|\mathcal{A}|}$. Since $d_h^{\pi_{\mathsf{out}}}$ is a probability distribution for every h, setting $K = 36H^2Q_{\mathsf{max}}^2\log(A)\varepsilon^{-2}$ implies

$$\frac{1}{K}\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x)\mathfrak{R}_h^{\pi}(x) \le 2\sqrt{\frac{H^2Q_{\mathsf{max}}^2\log|\mathcal{A}|}{K}} = \frac{\varepsilon}{3}.$$

Controlling the estimation error. It remains to control the average estimation error $\frac{1}{K}\sum_{k=1}^{K}\Delta(\pi^k)$. To this end, recall that, in Algorithm 1, we build the empirical estimator $\widehat{\mathcal{L}}_h$ by sampling n_{E} states from $d_h^{\pi_{\mathsf{out}}}$. This sampling is possible because $d_h^{\pi_{\mathsf{out}}}$ depends only on $\pi_{\mathsf{out},1}, \ldots, \pi_{\mathsf{out},h-1}$, which have already been computed. We then query the expert for an action at each sampled state. For each $i\in[n_{\mathsf{E}}]$, we sample

$$\boldsymbol{x}_h^i \sim d_h^{\pi_{\mathsf{out}}} \quad \text{and} \quad \boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}\big(\cdot\,|\,\boldsymbol{x}_h^i\big).$$

For every $k\in[K]$, the policy π^k belongs to the class $\Pi_{\mathcal{Q}}\subset\Delta(\mathcal{A})^{\mathcal{X}}$ defined below.

$$\Pi_{\mathcal{Q}} := \Big\{\pi : \exists m\in\{0,\ldots,K\}, \forall h, \exists Q_h^1,\ldots,Q_h^m\in\mathcal{Q}_h,\quad \pi_h(\cdot\,|\,x) = \mathrm{softmax}\Big(\eta\textstyle\sum_j Q_h^j(x,\cdot)\Big)\Big\}.$$

For each h, write $\Pi_{\mathcal{Q},h} := \{\pi_h : \pi\in\Pi_{\mathcal{Q}}\}$ for the corresponding class of decision rules. We upper bound the average estimation error by

$$\frac{1}{K}\sum_{k=1}^{K}\Delta\big(\pi^k\big) \le \sup_{\pi\in\Pi_{\mathcal{Q}}}\Delta(\pi) \le \max_{h\in[H]}\sup_{p\in\Pi_{\mathcal{Q},h}}\sup_{Q_h\in\mathcal{Q}_h}\Big|\widehat{\mathcal{L}}_h(p,Q_h) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(p,Q_h)\Big|.$$

To control the upper bound, we verify that we can apply the concentration argument from Lemma E.2. Fix h, $p\in\Pi_{\mathcal{Q},h}$, and $Q_h\in\mathcal{Q}_h$, and define

$$Y_i(Q_h,p) := Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - Q_h\big(\boldsymbol{x}_h^i, p\big) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(p,Q_h).$$

Conditionally on the history before the samples for stage h are drawn, $d_h^{\pi_{\mathsf{out}}}$ is fixed and $(Y_i(Q_h,p))_{i=1}^{n_{\mathsf{E}}}$ is a martingale difference sequence. Moreover, since $Q_h \in [0,Q_{\mathsf{max}}]^{\mathcal{X}\times\mathcal{A}}$, both $Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)$ and $Q_h(\boldsymbol{x}_h^i,p)$ belong to $[0,Q_{\mathsf{max}}]$, while $\mathcal{L}_h^{\pi_{\mathsf{out}}}(p,Q_h)$ belongs to $[-Q_{\mathsf{max}},Q_{\mathsf{max}}]$. Hence

$$|Y_i(Q_h,p)| \le 2Q_{\mathsf{max}} \qquad \text{almost surely.}$$

Applying Lemma E.2 with $d_h = d_h^{\pi_{\mathsf{out}}}$ and $\Pi_h = \Pi_{\mathcal{Q},h}$ therefore gives, for any radius $r>0$, with probability at least $1-\delta$,

$$\sup_{\pi\in\Pi_{\mathcal{Q}}} \Delta(\pi) \le 4r + Q_{\mathsf{max}}\sqrt{\frac{8\log\left(\frac{2}{\delta}\sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)\right)}{n_{\mathsf{E}}}}. \tag{18}$$

Taking $r=\varepsilon/(24H)$ and

$$n_{\mathsf{E}} \ge \frac{288H^2Q_{\mathsf{max}}^2}{\varepsilon^2}\log\left(\frac{2}{\delta}\sum_{h=1}^H \mathcal{N}_{\varepsilon/(24H)}(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)\right)$$

yields $H\sup_{\pi\in\Pi_{\mathcal{Q}}}\Delta(\pi)\le\frac{\varepsilon}{3}$ with probability at least $1-\delta$.

Putting everything together. Plugging the two previous bounds in the decomposition shown in Lemma F.1, and using $r=\varepsilon/(24H)$ in Eq. (18), with probability at least $1-\delta$, we have

$$\begin{aligned}
J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &\le \frac{1}{K}\sum_{h=1}^H\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x)\mathfrak{R}_h^{\pi}(x) + \frac{2H}{K}\sum_{k=1}^K \Delta(\pi^k)\\
&\le 2\sqrt{\frac{H^2Q_{\mathsf{max}}^2\log|\mathcal{A}|}{K}} + \frac{\varepsilon}{3} + 2HQ_{\mathsf{max}}\sqrt{\frac{8\log\left(\frac{2}{\delta}\sum_{h=1}^H \mathcal{N}_{\varepsilon/(24H)}(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)\right)}{n_{\mathsf{E}}}}\\
&\le \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3},
\end{aligned}$$

where the last line uses $K = 36H^2Q_{\mathsf{max}}^2\log(A)\varepsilon^{-2}$ and

$$n_{\mathsf{E}} \ge \frac{288H^2Q_{\mathsf{max}}^2}{\varepsilon^2}\log\left(\frac{2\sum_{h=1}^H \mathcal{N}_{\varepsilon/(24H)}(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)}{\delta}\right).$$

Next, we consider separately the convex and nonconvex cases for the classes $\mathcal{Q}_h$, and show that the sample complexity improves in the convex case.

Convex case. Assume first that each $\mathcal{Q}_h$ is convex. Fix $h\in[H]$. By Lemma E.4, more precisely by Eq. (15), for every radius $r>0$,

$$\mathcal{N}_r(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho) \le (K+1)\mathcal{N}_{\frac{r}{2\max\{1,Q_{\mathsf{max}}\eta K\}}}(\mathcal{Q}_h,\|\cdot\|_\infty)^2.$$

If $\varepsilon \ge Q_{\mathsf{max}}$, reward normalization gives $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le Q_{\mathsf{max}} \le \varepsilon$, so the performance claim is immediate. We may therefore assume that $\varepsilon < Q_{\mathsf{max}}$. For the parameter choice in the theorem,

$$Q_{\mathsf{max}}\eta K = \sqrt{K\log|\mathcal{A}|} = \frac{6HQ_{\mathsf{max}}\log|\mathcal{A}|}{\varepsilon}.$$

Since $H\ge 1$ and $A\ge 2$, we then have $Q_{\mathsf{max}}\eta K > 6H\log 2 > 1$. Hence $\max\{1,Q_{\mathsf{max}}\eta K\} = Q_{\mathsf{max}}\eta K$ in the remaining covering estimates. We now plug this stagewise bound into the estimation guarantee. With the radius $r=\varepsilon/(24H)$ used in Eq. (18), the product bound above implies

$$\sum_{h=1}^H \mathcal{N}_{\varepsilon/(24H)}(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho) \le H(K+1)\mathcal{N}^{\mathsf{max}}_{\frac{\varepsilon}{48HQ_{\mathsf{max}}\eta K}}(\mathcal{Q})^2.$$

Combining the bound above with the sufficient condition on n_{E} in the preceding paragraph, it is enough to take

$$n_{\mathsf{E}} \ge \frac{288H^2Q_{\mathsf{max}}^2}{\varepsilon^2}\log\left(\frac{2H(K+1)}{\delta}\mathcal{N}^{\mathsf{max}}_{\frac{\varepsilon}{48HQ_{\mathsf{max}}\eta K}}(\mathcal{Q})^2\right).$$

Substituting the values of K and $Q_{\max}\eta K$, we define

$$r_\varepsilon := \frac{\varepsilon}{48HQ_{\max}\eta K} = \frac{\varepsilon^2}{288H^2Q_{\max}\log|\mathcal{A}|}.$$

The following order of expert queries per stage then suffices.

$$\begin{aligned} n_{\mathsf{E}} &= \mathcal{O}\left(\frac{H^2Q_{\max}^2}{\varepsilon^2}\log\left(\frac{H\left(1+36H^2Q_{\max}^2\log|\mathcal{A}|/\varepsilon^2\right)}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})^2\right)\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{H^2Q_{\max}^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})\right)\right). \end{aligned}$$

Since the algorithm queries the expert at H stages, the total number of expert queries is

$$Hn_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\frac{H^3Q_{\max}^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})\right)\right).$$

Nonconvex classes. We now turn to the general case. Fix $h \in [H]$. By Lemma E.3, more precisely by Eq. (12), and since the case split above gives $Q_{\max}\eta K > 1$, for every radius $r > 0$,

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le (K+1)\mathcal{N}_{\frac{r}{2Q_{\max}\eta K}}(\mathcal{Q}_h, \|\cdot\|_\infty)^{K+1}.$$

With the radius $r = \varepsilon/(24H)$ used in Eq. (18), the product bound above implies

$$\sum_{h=1}^{H}\mathcal{N}_{\varepsilon/(24H)}(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le H(K+1)\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})^{K+1}.$$

Combining the bound above with the sufficient condition on n_{E} in the preceding paragraph, it is enough to take

$$n_{\mathsf{E}} \ge \frac{288H^2Q_{\max}^2}{\varepsilon^2}\log\left(\frac{2H(K+1)}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})^{K+1}\right).$$

Substituting the values of K and $Q_{\max}\eta K$, the following order of expert queries per stage suffices:

$$\begin{aligned} n_{\mathsf{E}} &= \mathcal{O}\left(\frac{H^2Q_{\max}^2}{\varepsilon^2}\log\left(\frac{H\left(1+36H^2Q_{\max}^2\log|\mathcal{A}|/\varepsilon^2\right)}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})^{1+36H^2Q_{\max}^2\log|\mathcal{A}|/\varepsilon^2}\right)\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{H^4Q_{\max}^4\log|\mathcal{A}|}{\varepsilon^4}\log\left(\frac{1}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})\right)\right). \end{aligned}$$

Since the algorithm queries the expert at H stages, the total number of expert queries is

$$Hn_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\frac{H^5Q_{\max}^4\log|\mathcal{A}|}{\varepsilon^4}\log\left(\frac{1}{\delta}\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})\right)\right).$$

This concludes the proof. □

G Proofs from Section 4 ($Q^{\pi_{\mathsf{E}}}$-Realizability Is Insufficient to Learn Offline)

Before presenting the proof of Theorem 4.1 presented in the main text, we show a stronger negative result for any algorithm which only uses *value-induced* policies, defined below.

Definition G.1 (Value-induced policy class)**.** *An offline IL algorithm* Alg *(Definition 4.1) uses a* value-induced *(VI) policy class (for value function class $\mathcal{Q}$) if it outputs policies of the form $\pi_{\mathsf{out},h}(a \mid x) \propto f(w^\top(Q_h^{1:K})(x,a))$, for some parameter $K \in \mathbb{N}$, sequence $Q^{1:K} \in \mathcal{Q}^K$, linear parameter $w \in \mathbb{R}^K$, and function $f : \mathbb{R} \to \mathbb{R}_+$.*

Notation wise, we will also use $\mathrm{LC}((Q_k)_{k=1}^K) := w^\top(Q_h^{1:K})(x,a)$ for some $w \in \mathbb{R}^K$. The VI policy class encompasses the policies used by a broad class of algorithms that extract policies from value functions by applying activation functions (e.g., softmax, ReLU, greedy) to linear combinations of value functions from $\mathcal{Q}$.[10] In particular, it includes the implicit policy classes used by prior theoretical and empirical value-based IL methods (Moulin et al., 2025a; Joshi et al., 2026; Garg et al., 2021; Watson et al., 2023) as well as by `OVI`. See Appendix G for details. The following lower bound shows that, under $Q^{\pi_{\mathsf{E}}}$-realizability alone, offline learners with VI policy classes fail to compete with the expert's performance, even in the limit of infinite data.

Theorem G.1 (Unidentifiability under VI policies)**.** *There exists a family $\mathcal{F}$ of MDPs with: i) deterministic initial states, rewards, and transitions, ii) horizon $H = 2$, iii) state space size $|\mathcal{X}| = 3$, and iv) family size $|\mathcal{F}| = 2$, as well as a value function class $\mathcal{Q}$ with $|\mathcal{Q}| = 2$ such that for every $\mathcal{M} \in \mathcal{F}$, Assumption 2.1 is satisfied for the expert $\pi_{\mathsf{E}} = \pi^\star_{\mathcal{M}}$. Moreover, for any algorithm* `Alg` *(Definition 4.1) with a VI policy class (Definition G.1), for any $n_{\mathsf{E}} \in \mathbb{N}^\star$, there exists an MDP $\mathcal{M} \in \mathcal{F}$ such that, given n_{E} expert trajectories sampled from π_{E},* `Alg` *outputs policies π_{out} such that $\mathbb{E}[J_{\mathcal{M}}^{\pi_{\mathsf{E}}} - J_{\mathcal{M}}^{\pi_{\mathsf{out}}}] \geq \frac{1}{4}$.*[11]

The intuition for Theorem G.1 is that the value-induced policies must randomize uniformly among actions with equally high value, even though the expert itself may only play a subset of those actions. In the construction of Theorem G.1 (illustrated in Figure 5), this implies that the learner's output policy goes out of distribution with constant probability, thus incurring a constant error rate for all dataset sizes. We expect that unidentifiability continues to hold under broader definitions of the VI policy class which assign probabilities to actions based on their *value profiles* (Sun et al., 2019a).

G.1 Examples of offline algorithms using VI policies

Next, we show that the class of algorithms that output a VI policy is general. It includes, for instance, `SPOIL` (Moulin et al., 2025a), while the pointwise-limit convention also represents greedy policies under suitable conditions.

`SPOIL` uses VI policies. Choosing $f(\cdot) = \exp(\cdot)$ and taking LC to be the sum multiplied by a scalar $\eta \geq 0$ gives

$$\pi_{\mathsf{out}}(a \mid x) = \mathrm{softmax}\Big(\eta \textstyle\sum_{k=1}^K Q_k(x,\cdot)\Big)_a,$$

which captures the output strategy of `SPOIL` and Algorithm 1.

Greedy policies as limits of VI policies. In the contextual-bandit setting of Joshi et al. (2026), we have $H = 1$, so we can omit the subscript h and consider a reward class $\mathcal{R}$ rather than a value function class $\mathcal{Q}$. Their learner forms a reward $r(x,a) = \sum_{\tilde{r} \in \mathcal{R}} w(\tilde{r})\tilde{r}(x,a)$ using learned weights $w \in \Delta(\mathcal{R})$ and uses a predictor that is greedy with respect to r. Such a predictor is not a finite-temperature VI policy: for any finite η, choosing $f_\eta(z) = \exp(\eta z)$ gives the softmax VI policy $\pi_\eta(\cdot \mid x) = \mathrm{softmax}(\eta r(x,\cdot))$, which generally assigns positive mass to nongreedy actions. Under our standing assumptions that $\mathcal{A}$ is finite and r is finite-valued, however, the softmax normalizer is strictly positive and finite, and these policies converge pointwise as $\eta \to \infty$.

$$\lim_{\eta \to \infty} \pi_\eta(a \mid x) = \frac{\mathbb{1}_{\{a \in \arg\max_{b \in \mathcal{A}} r(x,b)\}}}{|\arg\max_{b \in \mathcal{A}} r(x,b)|}.$$

Consequently, a uniformly tie-broken greedy policy belongs to the VI class through its pointwise-limit convention. A deterministic greedy policy does as well when the maximizing action is unique. Other

[10] We note that greedy policies are only included as pointwise limits of other VI policies, as shown in Appendix G.

[11] The expectation is over the training data and the internal randomness of `Alg`.

tie-breaking rules are covered only if they can be realized by an admissible sequence of VI scores. This gives a limiting connection to the roundwise greedy extraction step of Joshi et al. (2026); it does not assert that their final mixed output policy is itself a VI policy.

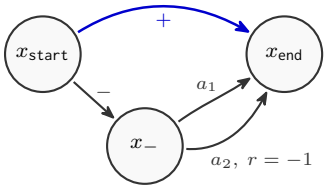
Algorithms used in practice output VI policies. Beyond prior work in IL theory (Moulin et al., 2025b), algorithms used in practice in IL, such as `IQ-LEARN` (Garg et al., 2021) and `CSIL` (Watson et al., 2023), are also value-based imitation algorithms. Indeed, these algorithms choose a single Q-function from the class $\mathcal{Q}$ and output $\pi_{\mathtt{out}} = \mathrm{softmax}(\alpha Q(x, a))$. This is recovered by the definition of algorithms outputting VI policies, choosing α to be the weight of the linear combination and taking $f(\cdot) = \exp(\cdot)$.

G.2 Proof of Theorem G.1 (Lower Bound Against Offline Algorithms that Output VI Policies)

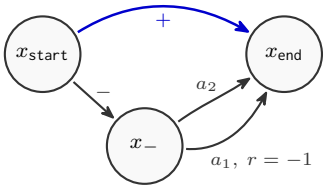
We prove Theorem G.1, which shows nonidentifiability within the class of algorithms outputting VI policies: even with infinite offline data, no such algorithm can guarantee error below $1/4$.

Theorem G.1 (Unidentifiability under VI policies)**.** *There exists a family $\mathcal{F}$ of MDPs with: i) deterministic initial states, rewards, and transitions, ii) horizon $H = 2$, iii) state space size $|\mathcal{X}| = 3$, and iv) family size $|\mathcal{F}| = 2$, as well as a value function class $\mathcal{Q}$ with $|\mathcal{Q}| = 2$ such that for every $\mathcal{M} \in \mathcal{F}$, Assumption 2.1 is satisfied for the expert $\pi_{\mathtt{E}} = \pi^\star_{\mathcal{M}}$. Moreover, for any algorithm* `Alg` *(Definition 4.1) with a VI policy class (Definition G.1), for any $n_{\mathtt{E}} \in \mathbb{N}^\star$, there exists an MDP $\mathcal{M} \in \mathcal{F}$ such that, given $n_{\mathtt{E}}$ expert trajectories sampled from $\pi_{\mathtt{E}}$,* `Alg` *outputs policies $\pi_{\mathtt{out}}$ such that $\mathbb{E}[J^{\pi_{\mathtt{E}}}_{\mathcal{M}} - J^{\pi_{\mathtt{out}}}_{\mathcal{M}}] \geq \frac{1}{4}$.*[12]

Proof of Theorem G.1. For clarity, we carry out the proof using centered rewards $r_h \in [-1, 0]$. Adding 1 to every reward yields an equivalent MDP with rewards in $[0, 1]$. This shift preserves optimal policies, suboptimality gaps, and all action comparisons. Shifting the value class accordingly yields a class of the same cardinality that lies in $[0, 2]$ and satisfies $Q^{\pi_{\mathtt{E}}}$-realizability with $Q_{\mathtt{max}} = 2$. Although an arbitrary link function need not be translation invariant, every function in the shifted class still assigns equal values to the two actions at $x_{\mathtt{start}}$, and the proof treats the action distribution at x_- as arbitrary. Hence the argument below applies unchanged. Formally, $a_1 \equiv +$ and $a_2 \equiv -$ are the two global actions; for every unspecified state-action-stage tuple, we assign zero centered reward and a transition to the absorbing state $x_{\mathtt{end}}$, whose features are zero.



(a) MDP $\mathcal{M}_1$: Action a_1 is optimal at x_-.



(b) MDP $\mathcal{M}_2$: Action a_2 is optimal at x_-.

Figure 5: Family of hard tasks. Unless otherwise specified, the reward is 0. The blue path is the one taken by the deterministic expert. $x_{\mathtt{end}}$ is an absorbing state.

We consider the family of hard MDPs $\mathcal{F} = \{\mathcal{M}_1, \mathcal{M}_2\}$ depicted in Figure 5. We consider the following 2-dimensional features that satisfy the $Q^{\pi_{\mathtt{E}}}$-realizability assumption.

- $\varphi(x_{\mathtt{start}}, +) = \varphi(x_{\mathtt{start}}, -) = [0, 0]^\intercal$,
- $\varphi(x_-, a_1) = [0, 1]^\intercal$,
- $\varphi(x_-, a_2) = [1, 0]^\intercal$.

If the expert policy satisfies $\pi_{\mathtt{E},1}(+ \,|\, x_{\mathtt{start}}) = 1$ and $\pi_{\mathtt{E},1}(a_1 \,|\, x_-) = 1$, then we have

- $Q^{\pi_{\mathtt{E},1}}_{\mathcal{M}_1}(x_{\mathtt{start}}, +) = Q^{\pi_{\mathtt{E},1}}_{\mathcal{M}_1}(x_{\mathtt{start}}, -) = 0$,
- $Q^{\pi_{\mathtt{E},1}}_{\mathcal{M}_1}(x_-, a_1) = 0$,
- $Q^{\pi_{\mathtt{E},1}}_{\mathcal{M}_1}(x_-, a_2) = -1$.

[12] The expectation is over the training data and the internal randomness of `Alg`.

Thus, by choosing $\theta_1 = [-1, 0]$, we can realize the expert action-value function as $Q^{\pi_{\mathsf{E},1}}_{\mathcal{M}_1}(x, a) = \varphi(x, a)^\intercal \theta_1$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. For the second MDP in $\mathcal{F}$, $\mathcal{M}_2$, the optimal expert policy satisfies $\pi_{\mathsf{E},2}(+ \,|\, x_{\mathsf{start}}) = 1$ and $\pi_{\mathsf{E},2}(a_2 \,|\, x_-) = 1$. Taking $\theta_2 = [0, -1]$, we can similarly write $Q^{\pi_{\mathsf{E},2}}_{\mathcal{M}_2}(x, a) = \varphi(x, a)^\intercal \theta_2$, so $\mathcal{M}_2$ is also $Q^{\pi_{\mathsf{E}}}$-realizable with the same features. Therefore, the two MDPs differ only in which action is good at x_-: a_1 in $\mathcal{M}_1$ and a_2 in $\mathcal{M}_2$.

Fix any dataset size $n_{\mathsf{E}} \in \mathbb{N}^\star$. We now start proving our result. For the moment, we consider the case of deterministic value-based offline IL algorithms that choose a sequence of state-action value functions without randomization. Moreover, we consider the class

$$\mathcal{Q} = \{Q^{\pi_{\mathsf{E},1}}_{\mathcal{M}_1}, Q^{\pi_{\mathsf{E},2}}_{\mathcal{M}_2}\},$$

where $Q^{\pi_{\mathsf{E},1}}_{\mathcal{M}_1}(x, a) = \varphi(x, a)^\intercal \theta_1$ and $Q^{\pi_{\mathsf{E},2}}_{\mathcal{M}_2}(x, a) = \varphi(x, a)^\intercal \theta_2$ for all x, a. Notice that in both $\mathcal{M}_1$ and $\mathcal{M}_2$, we have that $Q^{\pi_{\mathsf{E},1}}_{\mathcal{M}_1}(x_{\mathsf{start}}, a) = Q^{\pi_{\mathsf{E},2}}_{\mathcal{M}_2}(x_{\mathsf{start}}, a) = 0$ for all $a \in \{+, -\}$.

Under this setting, no matter how Alg selects the sequence $Q_1, \dots, Q_K$, the output policy at the starting state is uniform: $\pi_{\mathsf{out}}(\cdot \,|\, x_{\mathsf{start}}) = \mathrm{Unif}(\{+, -\})$. Indeed, every $Q \in \mathcal{Q}$ assigns the same value to the two actions $+$ and $-$ at x_{start}. Hence any well-defined linear combination $\mathrm{LC}(Q^{1:K})$ also assigns the same scalar to the two start actions, including when $K = \infty$. Since f is a function and the normalizing denominator is positive by definition of a valid output policy, the two numerator terms are equal and nonzero. Therefore, the policy π_{out} induced by any linear combination LC and function f in Definition G.1 guarantees $\pi_{\mathsf{out}}(+ \,|\, x_{\mathsf{start}}) = \pi_{\mathsf{out}}(- \,|\, x_{\mathsf{start}}) = \frac{1}{2}$. Moreover, taking pointwise limits or averaging policies obtained from different choices of sequences in $\mathcal{Q}$ still gives a uniform policy.

It follows that, with probability $1/2$, the next state is x_-. Suppose that the sequence $Q_1, \dots, Q_K$ chosen by Alg induces probabilities $\pi_{\mathsf{out}}(a_1 \,|\, x_-) = p$ and $\pi_{\mathsf{out}}(a_2 \,|\, x_-) = 1 - p$. The choice of $Q_1, \dots, Q_K$, and hence of p, cannot depend on whether the underlying MDP is $\mathcal{M}_1$ or $\mathcal{M}_2$, because the expert dataset is identical in the two cases, even with infinite data. Therefore, the suboptimality at x_-, averaged over the MDP class $\mathcal{F}$, equals $1/2$ as shown next.

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{2}\big(Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, a_i) - Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, \pi_{\mathsf{out}})\big) &= \frac{1}{2}\sum_{i=1}^{2}\big(-Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, \pi_{\mathsf{out}})\big) \\
&= \frac{p}{2}\sum_{i=1}^{2}(-r_{\mathcal{M}_i}(x_-, a_1)) + \frac{1-p}{2}\sum_{i=1}^{2}(-r_{\mathcal{M}_i}(x_-, a_2)) \\
&= \frac{p}{2} + \frac{1-p}{2} \\
&= \frac{1}{2}.
\end{aligned}$$

Fix $i \in \{1, 2\}$. Applying the performance difference lemma (Lemma E.9) with $\pi = \pi_{\mathsf{out}}$ and $\pi' = \pi_{\mathsf{E},i}$ to the MDP whose initial distribution is a point mass at x_{start} gives the following identity. Suppressing the stage index as above, the only state-occupancy terms that can contribute are $d^{\pi_{\mathsf{out}}}_1(x_{\mathsf{start}}) = 1$ and $d^{\pi_{\mathsf{out}}}_2(x_-) = \pi_{\mathsf{out}}(- \,|\, x_{\mathsf{start}}) = 1/2$. All terms at x_{end} vanish because the expert Q-values there are zero. Therefore,

$$\begin{aligned}
V^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}_i}(x_{\mathsf{start}}) &= 1 \sum_{a \in \{+,-\}} (\pi_{\mathsf{E},i}(a \,|\, x_{\mathsf{start}}) - \pi_{\mathsf{out}}(a \,|\, x_{\mathsf{start}}))Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_{\mathsf{start}}, a) \\
&\quad + \frac{1}{2}\sum_{a \in \{a_1, a_2\}} (\pi_{\mathsf{E},i}(a \,|\, x_-) - \pi_{\mathsf{out}}(a \,|\, x_-))Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, a) \\
&= 0 + \frac{1}{2}\left(Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, a_i) - \sum_{a \in \{a_1, a_2\}} \pi_{\mathsf{out}}(a \,|\, x_-)Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, a)\right) \\
&= \frac{1}{2}\big(Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, a_i) - Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-, \pi_{\mathsf{out}})\big).
\end{aligned}$$

Separately, the worst-case suboptimality gap is at least the average of the two suboptimality gaps over the two instances, so combining the two derivations above gives

$$\begin{aligned}\max_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}} V^{\pi_{\mathsf{E}}}_{\mathcal{M}}(x_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}}(x_{\mathsf{start}}) &\geq \frac{1}{2}\sum_{i=1}^{2}\big(V^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}_i}(x_{\mathsf{start}})\big)\\ &= \frac{1}{2}\sum_{i=1}^{2}\frac{1}{2}\big(Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-,a_i) - Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-,\pi_{\mathsf{out}})\big)\\ &= \frac{1}{2}\cdot\underbrace{\frac{1}{2}\sum_{i=1}^{2}\big(Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-,a_i) - Q^{\pi_{\mathsf{E},i}}_{\mathcal{M}_i}(x_-,\pi_{\mathsf{out}})\big)}_{=1/2 \text{ by the preceding display}} = \frac{1}{4}.\end{aligned}$$

The preceding display holds for every deterministic offline algorithm that outputs a VI policy under the uniform distribution over the two instances in $\mathcal{F}$. Therefore, by Yao's minimax principle (see, e.g., Bubeck and Cesa-Bianchi, 2012), the same lower bound holds against possibly randomized algorithms:

$$\max_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}} \mathbb{E}_{\pi_{\mathsf{out}}\sim\mathrm{Alg}}[V^{\star}_{\mathcal{M}}(x_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}}(x_{\mathsf{start}})] \geq \frac{1}{4},$$

where we used the equalities $V^{\pi_{\mathsf{E},1}}_{\mathcal{M}_1} = V^{\star}_{\mathcal{M}_1}$ and $V^{\pi_{\mathsf{E},2}}_{\mathcal{M}_2} = V^{\star}_{\mathcal{M}_2}$, which follow by optimality of the experts. □

G.3 Proof of Theorem 4.1 (Lower Bound Against Any Offline IL Algorithm)

In the above construction, we could learn by memorizing one expert action in x_{start} and playing it at deployment time. However, we can build a harder extended MDP by repeating the hard instance from Figure 5 several times, as shown in Figure 6. This lets us derive an information-theoretic lower bound against any offline IL algorithm. For algorithms outside the class of algorithms that output a VI policy, the problem is no longer nonidentifiable. However, the required dataset size scales linearly with the number of states, while the class $\mathcal{Q}$ realizing $Q^{\pi_{\mathsf{E}}}$ has cardinality independent of $|\mathcal{X}|$.

Theorem 4.1 (Main lower bound for offline IL)**.** *For any $X \in \mathbb{N}^{\star}$ and any $\varepsilon \in (0, \frac{1}{8}]$, there exists a family $\mathcal{F}$ of MDPs with: i) stochastic initial states but deterministic rewards and transitions, ii) horizon $H = 2$, iii) state space size $|\mathcal{X}| = \mathcal{O}(X)$, and iv) family size $\log(|\mathcal{F}|) = \mathcal{O}(X)$, as well as a value function class $\mathcal{Q}$ with $|\mathcal{Q}| = 2$ such that for every $\mathcal{M} \in \mathcal{F}$, Assumption 2.1 is satisfied for expert policy $\pi_{\mathsf{E}} = \pi^{\star}_{\mathcal{M}}$. Let $\mathtt{Alg}$ be any offline IL algorithm (Definition 4.1). Then, there exists an MDP $\mathcal{M} \in \mathcal{F}$ such that Alg needs $n_{\mathsf{E}} \geq \Omega(\frac{X}{\varepsilon})$ trajectories sampled offline with expert policy $\pi^{\star}_{\mathcal{M}}$, to output a policy π_{out} such that $\mathbb{E}[J^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}}] \leq \varepsilon$.*[13]

Proof of Theorem 4.1. As in the preceding proof of Theorem G.1, we use centered rewards $r_h \in [-1, 0]$ for convenience. Applying the same shifts and conventions yields the required formal construction without affecting the argument.

We consider a family $\mathcal{F} = \{\mathcal{M}^{\mathrm{ext}}_j\}_{j=1}^{2^{X+1}}$ in which each of the 2^{X+1} extended MDPs has $3X + 3$ states and is obtained by considering $X + 1$ copies of the structure in Figure 5, with starting states labeled as x^i_{start} for $i \in \{0, \dots, X\}$. For any $\varepsilon \leq \frac{1}{8}$, denote $q = \frac{8\varepsilon}{X}$ and consider the following initial distribution

$$\nu_0\big(x^0_{\mathsf{start}}\big) = 1 - Xq, \qquad \text{and} \qquad \nu_0\big(x^i_{\mathsf{start}}\big) = q \quad \text{for all } i \neq 0.$$

To construct the MDP family, we consider that each of the X copies indexed by $i \neq 0$ can be in two modes:

- In mode $+$, we consider that the action $+$ goes straight from x^i_{start} to x^i_{end}, without passing through x^i_-.
- In mode $-$, instead the action $-$ goes straight from x^i_{start} to x^i_{end}, without passing through x^i_-.

[13] The expectation is over the training data and the internal randomness of $\mathtt{Alg}$.

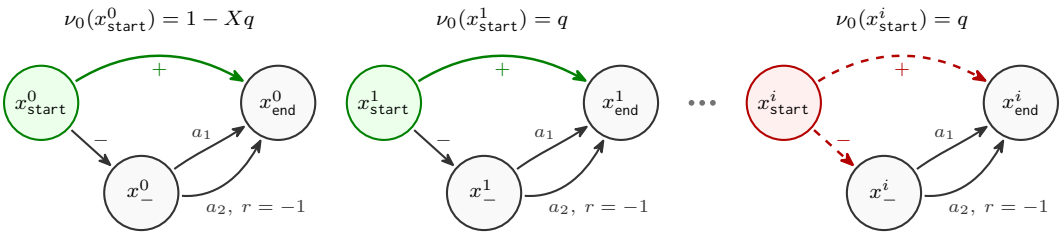


Figure 6: Representation of a MDP in the class $\mathcal{F}$. The states highlighted in green appear in $\mathcal{D}_{\mathsf{E}}$, so the learner can deterministically replay the expert action $+$. The red states are absent from $\mathcal{D}_{\mathsf{E}}$. Thus, no expert guidance is available there, and, conditional on starting at x^i_{start}, the learner reaches x^i_- with probability at least $1/2$ in some environment in $\mathcal{F}$, where it plays an action that is suboptimal under one of the two reward parameter choices θ_1 and θ_2. Here, $q = 8\varepsilon/X$, so ν_0 assigns mass $1 - Xq$ to x^0_{start} and mass q to every other root state. This initial distribution follows the lower-bound construction for tabular offline IL from Rajaraman et al. (2020) and ensures that guaranteeing expected suboptimality below ε uniformly over $\mathcal{F}$ requires $n_{\mathsf{E}} = \Omega(|\mathcal{X}|\varepsilon^{-1})$.

Considering all possible combinations of modes for each copy except the first, i.e., for $i = 1, \dots, X$, we obtain 2^X MDPs. Moreover, for each of these 2^X transition configurations, we can choose, as in the proof of Theorem G.1, the reward parameters to be $\theta_1 = [-1, 0]$ or $\theta_2 = [0, -1]$. Therefore, we obtain a family of 2^{X+1} MDPs. A visual representation of an extended MDP with parameters θ_1 is given in Figure 6.

The family of 2^{X+1} optimal expert policies is defined so that the expert policy in the copy with root state x^i_{start} in the MDP $\mathcal{M}^{\mathrm{ext}}_j$ takes the optimal action that does not pass through x^i_-. That is, the expert chooses action $+$ if the ith copy is in mode $+$ in the MDP $\mathcal{M}^{\mathrm{ext}}_j$ and action $-$ if the copy is in mode $-$. At each x^i_-, the expert takes an action that is optimal for the chosen reward parameter, while at every absorbing state all expert policies take the same fixed action.

Since all the experts are optimal, we denote the expert action-value function in the MDP $\mathcal{M}^{\mathrm{ext}}_j$ by $Q^\star_{\mathcal{M}^{\mathrm{ext}}_j}$ and the state value function by $V^\star_{\mathcal{M}^{\mathrm{ext}}_j}$. At this point, note that for each $j \in [2^{X+1}]$, we have that $Q^\star_{\mathcal{M}^{\mathrm{ext}}_j}$ is affine in the features

$$\varphi(x^i_{\mathsf{start}}, +) = \varphi(x^i_{\mathsf{start}}, -) = [0, 0]^\intercal, \quad \varphi(x^i_-, a_1) = [0, 1]^\intercal, \quad \text{and} \quad \varphi(x^i_-, a_2) = [1, 0]^\intercal,$$

with parameters either θ_1 or θ_2. Denoting $Q_\theta(x, a) = \varphi(x, a)^\intercal\theta$, the class $\mathcal{Q} = \{Q_{\theta_1}, Q_{\theta_2}\}$ satisfies $Q^{\pi_{\mathsf{E}}}$-realizability in $\mathcal{M}^{\mathrm{ext}}_j$ for every $j \in [2^{X+1}]$ and has cardinality independent of X. Formally, we have $Q^\star_{\mathcal{M}^{\mathrm{ext}}_j} \in \mathcal{Q}$ for any $j \in [2^{X+1}]$.

In each MDP, the expert from the root states $x^0_{\mathsf{start}}, \dots, x^X_{\mathsf{start}}$ only takes the action that avoids the states $x^0_-, \dots, x^X_-$, so it does not reveal which action (between a_1 and a_2) is optimal in those states and the learner cannot determine whether the reward parameter vector is θ_1 or θ_2.

For the averaging argument, draw an environment $\mathcal{M}$ uniformly from $\mathcal{F}$, equivalently by drawing its modes $\sigma_1, \dots, \sigma_X$ independently and uniformly from $\{+, -\}$ and its reward parameter Θ independently and uniformly from $\{\theta_1, \theta_2\}$. Here, σ_i specifies the mode of copy i, so action σ_i bypasses x^i_-, while Θ is the reward parameter shared by all copies and determines which action is optimal at each x^i_-. Let ξ be a random seed, independent of the environment and expert data, that encodes all of the learner's internal randomness, and define $\mathsf{T} := (\mathcal{D}_{\mathsf{E}}, \xi)$. Thus, T consists exactly of the expert dataset and the learner's random seed. Fix $i \in [X]$ and consider the event $\mathcal{E}_i := \{x^i_{\mathsf{start}} \notin \mathcal{D}_{\mathsf{E}}\}$. On $\mathcal{E}_i$, the dataset contains no observation from copy i, so it does not depend on σ_i. Moreover, it does not depend on Θ: the expert never visits any of the states $x^0_-, \dots, x^X_-$, rewards are unobserved, and all expert policies take the same fixed action at the absorbing states. Since ξ is independent of the environment and expert data, T is therefore independent of (σ_i, Θ) conditional on $\mathcal{E}_i$, even though it may reveal the modes of other sampled copies. Consequently, conditional on any realization $\mathsf{T} = t$ and $\mathcal{E}_i$, the variables σ_i and Θ remain independent and uniform, while the learner's output policy may depend arbitrarily on t. Fix such a transcript t, write $\pi_{\mathsf{out},t}$ for the resulting output policy, and let $p_t := \pi_{\mathsf{out},t,1}(+ \,|\, x^i_{\mathsf{start}})$ and $u_t := \pi_{\mathsf{out},t,2}(a_1 \,|\, x^i_-)$. Averaging over σ_i, the probability of reaching x^i_- is $(1 - p_t)/2 + p_t/2 = 1/2$. Independently, averaging over Θ, the suboptimality at x^i_-

is $(1-u_t)/2 + u_t/2 = 1/2$. Therefore, for every such t,

$$\mathbb{E}\big[V^\star_{\mathcal{M}}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out},t}}_{\mathcal{M}}(x^i_{\mathsf{start}}) \,\big|\, \mathsf{T} = t, \mathcal{E}_i\big] = \left(\frac{1-p_t}{2} + \frac{p_t}{2}\right)\left(\frac{1-u_t}{2} + \frac{u_t}{2}\right) = \frac{1}{4}.$$

For each j, let $\mathbb{E}_j$ denote expectation under the transcript distribution in $\mathcal{M}^{\mathrm{ext}}_j$ and the learner's internal randomness, and write $\Delta_{j,i} := V^\star_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}})$. Since $\mathcal{E}_i$ depends only on the initial-state draws, it has probability $(1-q)^{n_{\mathsf{E}}}$ in every environment. Averaging the preceding identity over T therefore yields the unconditional bound

$$\frac{1}{2^{X+1}} \sum_{j=1}^{2^{X+1}} \mathbb{E}_j[\Delta_{j,i} \mathbb{1}_{\mathcal{E}_i}] \geq \frac{1}{4}(1-q)^{n_{\mathsf{E}}}.$$

Optimality implies $\Delta_{j,i} \geq 0$ at every root state. Thus, dropping the $i = 0$ term and retaining only the terms on $\mathcal{E}_i$, for any $j \in [2^{X+1}]$ we have

$$\begin{aligned}
\left\langle \nu_0, V^\star_{\mathcal{M}^{\mathrm{ext}}_j} - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j} \right\rangle &= \sum_{i=0}^{X} \nu_0(x^i_{\mathsf{start}}) \Big(V^\star_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) \Big) \\
&\geq \sum_{i=1}^{X} \nu_0(x^i_{\mathsf{start}}) \Big(V^\star_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) \Big) \\
&= \sum_{i=1}^{X} q \Big(V^\star_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) \Big) \\
&\geq \sum_{i=1}^{X} q \Big(V^\star_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j}(x^i_{\mathsf{start}}) \Big) \mathbb{1}_{\mathcal{E}_i}.
\end{aligned}$$

Taking $\mathbb{E}_j$ in the preceding inequality, averaging over the environments, and applying the unconditional bound above gives

$$\begin{aligned}
&\max_{j \in [2^{X+1}]} \mathbb{E}_j \left[\left\langle \nu_0, V^\star_{\mathcal{M}^{\mathrm{ext}}_j} - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j} \right\rangle \right] \\
&\quad\geq \frac{1}{2^{X+1}} \sum_{j=1}^{2^{X+1}} \mathbb{E}_j \left[\left\langle \nu_0, V^\star_{\mathcal{M}^{\mathrm{ext}}_j} - V^{\pi_{\mathsf{out}}}_{\mathcal{M}^{\mathrm{ext}}_j} \right\rangle \right] \\
&\quad\geq \sum_{i=1}^{X} q \frac{1}{2^{X+1}} \sum_{j=1}^{2^{X+1}} \mathbb{E}_j[\Delta_{j,i} \mathbb{1}_{\mathcal{E}_i}] \\
&\quad\geq \frac{q}{4} \sum_{i=1}^{X} (1-q)^{n_{\mathsf{E}}} \\
&\quad= \frac{X}{4} q (1-q)^{n_{\mathsf{E}}}.
\end{aligned}$$

Then, if $n_{\mathsf{E}} \leq X/(32\varepsilon)$, Bernoulli's inequality gives $(1-q)^{n_{\mathsf{E}}} = (1 - \frac{8\varepsilon}{X})^{n_{\mathsf{E}}} \geq 1 - \frac{8\varepsilon n_{\mathsf{E}}}{X} \geq \frac{3}{4}$. Therefore,

$$\max_{\mathcal{M} \in \mathcal{F}} \mathbb{E}[\langle \nu_0, V^\star_{\mathcal{M}} - V^{\pi_{\mathsf{out}}}_{\mathcal{M}} \rangle] \geq \frac{X}{4} \cdot \frac{8\varepsilon}{X} \cdot \frac{3}{4} = \frac{3\varepsilon}{2} > \varepsilon,$$

Thus, recalling $3(X+1) = |\mathcal{X}|$, achieving ε-suboptimality requires $n_{\mathsf{E}} = \Omega(|\mathcal{X}|/\varepsilon)$ for any offline IL algorithm. □

Part II
Additional Results

We collect here additional results omitted from the main text. First, in Appendix H, we show that assuming only reward realizability is not sufficient for efficient learning because the run time of any algorithm can be linear in the number of states in the worst case. Next, we complete Section 3 with a variant of OVI, dubbed OVI+, with better theoretical guarantees at the cost of a slightly more complicated algorithmic design (see Appendix I). In Appendix J, we show that reversing the order in which the players update in OVI yields an algorithm that is less computationally efficient, but provides statistical benefits for nonconvex $\mathcal{Q}$ classes and allows us to learn a stationary solution in discounted MDPs.

In addition to the lower bound from Section 4, we show in Appendix K that it is possible to learn offline if one assumes the (optimal) expert covers the learner's distribution.

Moreover, in Appendix L, we show that interaction enables efficient learning even when the class $\mathcal{Q}$ realizes either Q^{π_E} or the Q-value of any policy that the algorithm might produce, without the learner knowing which case holds. The key algorithmic idea is to sample states from a mixture of the expert and learner occupancy measures, a technique proven successful in the context of language models (Agarwal et al., 2024; Li et al., 2026a). This adaptive variant of OVI therefore applies to a strictly larger set of classes $\mathcal{Q}$ than that covered by the offline algorithm of Moulin et al. (2025a), revealing a previously unknown benefit of interaction.

H Reward realizability is insufficient

The expert's value function Q^{π_E} encodes information about the expert policy π_E, the reward function, and the transitions of the MDP via the Bellman equations. A natural question is whether the assumption can be weakened to reward realizability alone while retaining efficient learning. Specifically, assume that the learner has access to a reward function class $\mathcal{R} \subseteq \{(r'_h \colon \mathcal{X} \times \mathcal{A} \to [0,1])_{h=1}^H\}$ satisfying the following expressivity condition.

Assumption H.1 (Reward realizability)**.** *The class $\mathcal{R}$ contains the MDP's reward function, i.e., $r \in \mathcal{R}$.*

The following result shows that mere reward realizability is insufficient for efficient IL. Our lower bound is established against the class of interactive IL algorithms defined below, formalizing the interaction protocol described in Section 2.1.

Definition H.1 (Interactive IL algorithm)**.** *An interactive IL algorithm* Alg *with a fixed interaction budget $n_\mathsf{E} \in \mathbb{N}^\star$ is a (potentially randomized) procedure that, in each episode $k \in [n_\mathsf{E}]$, selects a policy π^k based on the preceding interaction history, rolls it out in the MDP to generate states $\boldsymbol{x}_h^k \sim d_h^{\pi^k}$ for all $h \in [H]$, and **is allowed** to query the expert for actions $\boldsymbol{a}_{\mathsf{E},h}^k \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^k)$ for all $h \in [H]$, without observing rewards. After the n_E episodes, it maps the resulting interaction history to a nonstationary policy $(\pi_{\mathsf{out},h} \colon \mathcal{X} \to \Delta(\mathcal{A}))_{h=1}^H$.*

Our negative result for this setting establishes a lower bound on the *rounds of interactions* used by any algorithm following the protocol defined in Definition H.1. Given an MDP $\mathcal{M}$, we write $J_\mathcal{M}^\pi$ for the return of a policy π, and $\pi_\mathcal{M}^\star$ for one of its optimal deterministic policies.

Theorem H.1 (Lower bound under reward realizability)**.** *For every $X \geq 1$ and $\varepsilon \in (0, 1/4]$, there exists a family $\mathcal{F}$ of MDPs with: i) stochastic initial states but deterministic rewards and transitions, ii) horizon $H = 2$, iii) state space size $|\mathcal{X}| = \mathcal{O}(X)$, and iv) family size $\log(|\mathcal{F}|) = \mathcal{O}(X)$, such that for every $\mathcal{M} \in \mathcal{F}$, Assumption H.1 is satisfied by the same singleton reward class $\mathcal{R} = \{r^\star\}$. Let* Alg *be any interactive IL algorithm (Definition H.1). Then, there exists an MDP $\mathcal{M} \in \mathcal{F}$, with corresponding expert $\pi_\mathsf{E} = \pi_\mathcal{M}^\star$, such that* Alg *needs $\Omega\big(\frac{X}{\varepsilon}\big)$* rounds of interactions *to output a policy π_out such that $\mathbb{E}[J_\mathcal{M}^{\pi_\mathsf{E}} - J_\mathcal{M}^{\pi_\mathsf{out}}] \leq \varepsilon$.*[14]

We recall that the results of Joshi et al. (2026) establish that reward realizability is sufficient in the offline setting when $H = 1$. By sharp contrast, our lower bound establishes that this is insufficient even

[14] The expectation is over all randomness in the protocol, including the trajectories sampled from the MDP, the actions sampled from the expert policy, and the internal randomness of Alg.

when $H = 2$, and even with interaction. As a simple corollary of Theorem H.1, either the number of environment interactions *or* the number of expert queries must be $\Omega(\min\{|\mathcal{X}|, \log(|\mathcal{F}|)\}/\varepsilon)$, and thus it is not possible to be polynomial in the relevant problem parameters $(\log(|\mathcal{R}|), H, \varepsilon^{-1}, \log(\delta^{-1}))$ for both resources simultaneously.

Our results provide some worst-case impossibility results for the "reward" moment matching template of Swamy et al. (2021). In general, to avoid the "tabular" rate $\Omega(|\mathcal{X}|)$, we expect that algorithms for this setting must estimate functionals of the dynamics, and thus additional structural conditions on the environment or additional representational ability, e.g. of the dynamics (Liu et al., 2022; Kidambi et al., 2021; Viano et al., 2024; Moulin et al., 2025b) or temporal differences of candidate Q functions (Garg et al., 2021), may be needed. We leave this as an interesting direction for future work.

Intuitively, the source of hardness leading to Theorem H.1 is that, even though the reward provides information on which states are rewarding, it does not provide information on which actions lead to those states. Consequently, the learner cannot use the reward information to learn how to act at step $h = 1$, and must resort to *either* using an expert query and cloning the received action *or* applying a learner's action in the environment and using the observed next state to learn the dynamics. Overall this takes $\min\{|\mathcal{X}|, \log(|\mathcal{F}|)\}/\varepsilon$ samples. By contrast, Theorem 3.1 shows that this transition-dependent information is encoded in Q^{π_E} and can be exploited by `OVI`. To conclude the section, we summarize in Table 3 various representational assumptions considered in this work and whether they are statistically tractable or intractable in the offline or interactive IL settings.

Representation	**Reward** r	**Value** Q^{π_E}	**Policy** π_E
Offline IL (Definition 4.1)	✗ (Theorem H.1)	✗ (Theorem 4.1)	✓ (e.g., `BC`)
Interactive IL (Definition H.1)	✗ (Theorem H.1)	✓* (`OVI`; Theorem 3.1)	✓ (e.g., `DAgger`)

Table 3: Tractability of imitation learning (for $H > 1$) under different representational assumptions. ✓: A $\mathrm{poly}(\log(|\mathcal{F}|), H, \varepsilon^{-1}, \log(\delta^{-1}))$ sample complexity is possible, where $\mathcal{F}$ is a given finite reward/value/policy class satisfying realizability (Assumptions H.1, 2.1, and 3.1, respectively). ✗: Said polynomial sample complexity is not possible. *: incurs additional $\log(A)$ dependence.

H.1 Proof of Theorem H.1 (Reward Realizability Is Insufficient)

So far, we showed that, under Q^{π_E}-realizability, interactive IL is possible but offline IL is not without additional coverage. It is natural to ask whether one can weaken the representation requirement even further. A tempting candidate is reward realizability: when $H = 1$, Q^{π_E}-realizability reduces to reward realizability, and Joshi et al. (2026) show that reward realizability is enough for statistical learnability in contextual bandits, although their algorithm is computationally inefficient.

The theorem below shows that, in general MDPs, reward realizability alone cannot yield a representation-dependent guarantee even with interaction: the reward class is a singleton, but the learner still needs coverage of the rare decision states. This motivates the stronger assumption used in the main text, Q^{π_E}-realizability, which stores additional information about the transitions that a reward function alone does not provide.

Theorem H.1 (Lower bound under reward realizability)**.** *For every $X \geq 1$ and $\varepsilon \in (0, 1/4]$, there exists a family $\mathcal{F}$ of MDPs with: i) stochastic initial states but deterministic rewards and transitions, ii) horizon $H = 2$, iii) state space size $|\mathcal{X}| = \mathcal{O}(X)$, and iv) family size $\log(|\mathcal{F}|) = \mathcal{O}(X)$, such that for every $\mathcal{M} \in \mathcal{F}$, Assumption H.1 is satisfied by the same singleton reward class $\mathcal{R} = \{r^\star\}$. Let `Alg` be any interactive IL algorithm (Definition H.1). Then, there exists an MDP $\mathcal{M} \in \mathcal{F}$, with corresponding expert $\pi_E = \pi^\star_{\mathcal{M}}$, such that `Alg` needs $\Omega\big(\frac{X}{\varepsilon}\big)$ rounds of interactions to output a policy $\pi_{\mathtt{out}}$ such that $\mathbb{E}[J^{\pi_E}_{\mathcal{M}} - J^{\pi_{\mathtt{out}}}_{\mathcal{M}}] \leq \varepsilon$.*[15]

Comparison with Q^{π_E}-realizability. This lower bound is specific to reward realizability and should not be read as a contradiction to the guarantees under Assumption 2.1. In the construction below, where each MDP is characterized by a bit vector $b \in \{0, 1\}^X$, the common reward only identifies x^i_+ as rewarding and x^i_- as non-rewarding but it does not reveal which action from $x^i_{\mathtt{start}}$ reaches x^i_+.

[15] The expectation is over all randomness in the protocol, including the trajectories sampled from the MDP, the actions sampled from the expert policy, and the internal randomness of `Alg`.

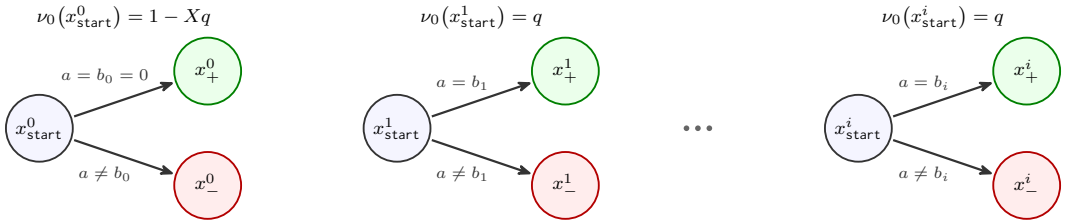


Figure 7: Instance from the family used in Theorem H.1. Every MDP in the family uses the same reward, and, for $i \in [X]$, the hidden bit $b_i \in \{0,1\}$ selects which action reaches x^i_+ from x^i_{start}. It remains hidden unless x^i_{start} is observed during training.

By contrast, for each instance $\mathcal{M}_b$, the expert value function satisfies $Q^{\pi_{\mathsf{E}}^b}_{\mathcal{M}_b,1}(x^i_{\mathsf{start}}, a) = \mathbb{1}\{a = b_i\}$. Thus $Q^{\pi_{\mathsf{E}}}$ stores the information about which action to take at the starting state itself. Equivalently, a value class that $Q^{\pi_{\mathsf{E}}}$-realizes all MDPs in this family must contain one distinct function for each $b \in \{0,1\}^X$, so its complexity already scales with X. Reward realizability avoids this complexity entirely by using a singleton class, but then the learner receives no representation of the control information needed at unobserved states.

Proof of Theorem H.1. Set $q = 4\varepsilon/X$, which satisfies $q \leq 1/X$ because $\varepsilon \leq 1/4$. We first define the hard family $\mathcal{H} = \mathcal{H}_{X,q}$. Let the action space be $\mathcal{A} = \{0,1\}$. For each $i \in \{0,\dots,X\}$, the state space contains a start state x^i_{start} and two terminal states x^i_+ and x^i_-. The initial distribution is

$$\nu_0(x^0_{\mathsf{start}}) = 1 - Xq, \qquad \text{and} \qquad \nu_0(x^i_{\mathsf{start}}) = q \quad \text{for } i \in [X].$$

For $b = (b_1,\dots,b_X) \in \{0,1\}^X$, set $b_0 = 0$. In the MDP $\mathcal{M}_b$, the first transition is deterministic and given by

$$P^b_1(x^i_+ \mid x^i_{\mathsf{start}}, a) = \mathbb{1}\{a = b_i\}, \qquad \text{and} \qquad P^b_1(x^i_- \mid x^i_{\mathsf{start}}, a) = \mathbb{1}\{a \neq b_i\}.$$

All stage-two behavior is fixed independently of b. For concreteness, let the expert's terminal action be 0, and take any fixed terminal transitions. The reward is common to all instances. For any index i and action a,

$$r^\star_1(x^i_{\mathsf{start}}, a) = 0, \qquad r^\star_2(x^i_+, a) = 1, \qquad r^\star_2(x^i_-, a) = 0.$$

Thus the singleton class $\mathcal{R} = \{r^\star\}$ realizes every instance in the family. Let the expert in $\mathcal{M}_b$ be the deterministic policy $\pi^b_{\mathsf{E},1}(x^i_{\mathsf{start}}) = b_i$, with the fixed terminal behavior described above. The expert reaches x^i_+ from every x^i_{start}, so it is optimal and $J^{\pi^b_{\mathsf{E}}}_{\mathcal{M}_b} = 1$. The family is illustrated in Figure 7.

For any policy π, starting from x^i_{start} yields terminal reward one if and only if the first action equals b_i. Hence

$$J^{\pi^b_{\mathsf{E}}}_{\mathcal{M}_b} - J^{\pi}_{\mathcal{M}_b} = 1 - \sum_{i=0}^{X} \nu_0(x^i_{\mathsf{start}})\pi_1(b_i \mid x^i_{\mathsf{start}}) = \sum_{i=0}^{X} \nu_0(x^i_{\mathsf{start}})\pi_1(\mathcal{A} \setminus \{b_i\} \mid x^i_{\mathsf{start}}). \tag{19}$$

We lower bound the average suboptimality gap over a random instance. Let $B = (B_1,\dots,B_X) \sim \mathrm{Unif}(\{0,1\}^X)$, and run $\mathtt{Alg}$ on $\mathcal{M}_B$. Let ξ be an independent random seed encoding all of the algorithm's internal randomness, and let $\mathcal{D}$ denote the complete interaction transcript available to the learner after the fixed n_{E} episodes. For each $i \in [X]$, let V_i be the event that x^i_{start} appears as the initial state in one of the fixed n_{E} interaction rounds.

Although B_i is sampled uniformly initially, we later condition on the dataset used to choose π_{out}. This conditioning can matter because $\mathcal{D}$ is generated in the MDP whose expert policy and transitions depend on B_i. If V_i occurs, then $\mathcal{D}$ contains a transition out of x^i_{start}. No matter whether an expert action is observed or not, the transition out of x^i_{start} reveals which action reaches x^i_+. On V_i^c, however, $\mathcal{D}$ contains no observation from the i-th component of the construction: the state x^i_{start} can only appear as an initial state, and no transition from any other component reaches it. Thus, conditional on V_i^c, $\mathcal{D}$ may depend on the coordinates B_j for $j \neq i$, but the conditional law of $(\mathcal{D}, \xi)$ does not depend on B_i. Since B_i is uniform and independent of B_j for $j \neq i$, even after seeing the dataset and knowing V_i^c, we still have no information about B_i. Whenever $\mathbb{P}[V_i^c] > 0$, this gives

$$\mathbb{P}[B_i = 0 \,|\, \mathcal{D}, V_i^c, \xi] = \mathbb{P}[B_i = 1 \,|\, \mathcal{D}, V_i^c, \xi] = \frac{1}{2}.$$

Conditional on $\mathcal{D}$, V_i^c, and ξ, the output policy is fixed. We define $p_i := \pi_{\mathsf{out},1}(1 \mid x^i_{\mathsf{start}})$. The quantity that appears in the suboptimality gap for the i-th component is $\pi_{\mathsf{out},1}(\mathcal{A} \setminus \{B_i\} \mid x^i_{\mathsf{start}})$, which is the probability that the output policy chooses the wrong first action at x^i_{start}. If $B_i = 0$, action 1 is wrong, and this quantity equals p_i. If $B_i = 1$, action 0 is wrong, and this quantity equals $1 - p_i$.

When $\mathbb{P}[V_i^c] > 0$, averaging over $\mathcal{D}$ and ξ and multiplying by $\mathbb{P}[V_i^c]$ gives the identity below; when $\mathbb{P}[V_i^c] = 0$, the identity is trivial. Hence

$$\mathbb{E}\big[\pi_{\mathsf{out},1}(\mathcal{A} \setminus \{B_i\} \mid x^i_{\mathsf{start}}) \mathbb{1}\{V_i^c\}\big] = \frac{1}{2}\mathbb{P}[V_i^c]. \tag{20}$$

Because n_{E} is a fixed deterministic budget and the initial state in each interaction round is sampled independently from ν_0, the event V_i^c has probability

$$\mathbb{P}[V_i^c] = (1-q)^{n_{\mathsf{E}}}.$$

Combining Eqs. (19) and (20) with this identity and discarding the nonnegative contribution of x^0_{start} gives

$$\mathbb{E}\Big[J^{\pi^B_{\mathsf{E}}}_{\mathcal{M}_B} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}_B}\Big] \geq \sum_{i=1}^{X} q\mathbb{E}\big[\pi_{\mathsf{out},1}(\mathcal{A} \setminus \{B_i\} \mid x^i_{\mathsf{start}}) \mathbb{1}\{V_i^c\}\big] = \frac{Xq}{2}(1-q)^{n_{\mathsf{E}}}. \tag{21}$$

Since the left-hand side averages uniformly over B, there exists a fixed $b \in \{0,1\}^X$ such that

$$\mathbb{E}\Big[J^{\pi^b_{\mathsf{E}}}_{\mathcal{M}_b} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}_b}\Big] \geq \frac{Xq}{2}(1-q)^{n_{\mathsf{E}}}.$$

If $n_{\mathsf{E}} < X/(8\varepsilon)$, Bernoulli's inequality gives

$$(1-q)^{n_{\mathsf{E}}} = (1 - \frac{4\varepsilon}{X})^{n_{\mathsf{E}}} \geq 1 - \frac{4\varepsilon n_{\mathsf{E}}}{X} > \frac{1}{2}.$$

Therefore, $\mathcal{M}_b$ satisfies

$$\mathbb{E}\Big[J^{\pi^b_{\mathsf{E}}}_{\mathcal{M}_b} - J^{\pi_{\mathsf{out}}}_{\mathcal{M}_b}\Big] > \frac{X}{2} \cdot \frac{4\varepsilon}{X} \cdot \frac{1}{2} = \varepsilon.$$

Thus, achieving expected suboptimality at most ε on every instance requires $n_{\mathsf{E}} \geq X/(8\varepsilon)$. Since $|\mathcal{X}| = 3X + 3$, the sample-complexity statement follows.

□

I OVI+: Improved Bound on the Number of Expert Queries

Here, we present OVI+, which reduces the number of expert queries by reusing expert-labeled data while drawing fresh unlabeled states to estimate the learner term. Suppressing logarithmic and value-class complexity factors, the expert-query bound for general value-function classes improves from order ε^{-4} to order ε^{-2}, while the learner-rollout bound retains an ε^{-4} dependence. We first present the pseudocode for OVI+ in Algorithm 3 and explain its design, then establish its guarantees.

I.1 Design of OVI+

All samples within a batch are drawn independently conditional on the history before that batch. As for OVI, we assume that the Q-player's best responses exist.

Algorithm 3 OVI+: **O**n-Policy **V**alue-Based **I**mitation Learning with Expert Sample Buffer

input: Learning rate η, iterations K, expert queries per stage n_{E}, learner batch size per iteration n_{L}.
for $h = 1, \dots, H$ **do**
  Create a dataset at stage h: for $i \in [n_{\mathsf{E}}]$, sample $\boldsymbol{x}_h^i \sim d_h^{\pi_{\mathsf{out}}}$, and query $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_h^i)$.
  Initialize $\pi_h^1 = \mathrm{Unif}(\mathcal{A})$.
  for $k = 1, \dots, K$ **do**
    `# Sample a fresh batch for estimating the learner term`
    For $j \in [n_{\mathsf{L}}]$, sample $\boldsymbol{x}_h^{jk} \sim d_h^{\pi_{\mathsf{out}}}$.
    Set $Q_h^k \in \arg\max_{Q_h \in \mathcal{Q}_h} \underbrace{\frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)}_{\text{Expert term}} - \underbrace{\frac{1}{n_{\mathsf{L}}} \sum_{j=1}^{n_{\mathsf{L}}} Q_h(\boldsymbol{x}_h^{jk}, \pi_h^k)}_{\text{Learner term}}$.
    Set $\pi_h^{k+1}(a \,|\, x) \propto \pi_h^k(a \,|\, x) e^{\eta Q_h^k(x,a)}$, for any state-action pair (x, a).
  Create the output policy at layer h: $\pi_{\mathsf{out},h} = \frac{1}{K} \sum_{k=1}^{K} \pi_h^k$.

The main idea behind OVI+ is to estimate the expert and learner terms in the Q-player's objective using separate samples. At each layer h, the previously learned decision rules fix $d_h^{\pi_{\mathsf{out}}}$. This allows us to collect a batch of expert-labeled data $\{(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)\}_{i=1}^{n_{\mathsf{E}}}$ before starting the inner loop over $k \in [K]$. We reuse this batch to estimate the expert term in the objective of the Q-player, i.e., $\sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h(x, \pi_{\mathsf{E},h})$. For each fixed Q_h, the centered expert-sample contributions form a martingale difference sequence when the batch is revealed sequentially, conditional on the history before that batch. Since this term does not involve the current decision rule π_h^k, uniform concentration over $\mathcal{Q}_h$ suffices to control its estimation error throughout the inner loop. By contrast, the learner term depends on π_h^k, which is updated within the inner loop. When the same states are reused to estimate this term, as in Algorithm 1, the current decision rule depends on those states. Even for a fixed Q_h, the centered sample contributions need not form a martingale difference sequence when evaluated at this data-dependent decision rule. A concentration bound for a fixed decision rule therefore does not directly apply. The analysis of OVI handles this dependence with an estimation bound for $\sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h(x, \pi_h)$ that holds uniformly, i.e., for all $Q_h \in \mathcal{Q}_h$ and $\pi_h \in \Pi_{\mathcal{Q},h}$. In Algorithm 3, we instead collect a fresh batch of n_{L} states $\{\boldsymbol{x}_h^{jk}\}_{j=1}^{n_{\mathsf{L}}}$ from $d_h^{\pi_{\mathsf{out}}}$ at each iteration $k \in [K]$ to estimate $\sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h(x, \pi_h^k)$. Since the batch is sampled after π_h^k is computed, we can condition on the preceding history and treat π_h^k as fixed. For each fixed Q_h, the centered contributions from this fresh batch then form a martingale difference sequence. We therefore only need uniform concentration over $\mathcal{Q}_h$, without uniformity over the learner policy class. Collecting these states only requires rolling in with the previously learned decision rules $\pi_{\mathsf{out},1}, \dots, \pi_{\mathsf{out},h-1}$ and requires no expert queries. These additional samples therefore do not increase the expert-query complexity.

I.2 Guarantees for OVI+

We now state a guarantee that improves upon the expert-query bound for OVI.

Theorem I.1 (Sample complexity of OVI+). *Let Assumption 2.1 hold. For every $r > 0$, let $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q})$ be the largest ℓ_∞-covering number (see Definition E.1) of $\mathcal{Q}_h$ over $h \in [H]$. Then, for any $\varepsilon \in (0, \min\{1, Q_{\mathsf{max}}\})$ and $\delta \in (0,1)$, OVI+ (Algorithm 3) with parameters*

$$\eta = \sqrt{\frac{\log(A)}{KQ_{\mathsf{max}}^2}}, \qquad K = \left\lceil \frac{36H^2Q_{\mathsf{max}}^2 \log(A)}{\varepsilon^2} \right\rceil,$$
$$n_{\mathsf{L}} = n_{\mathsf{E}} = \left\lceil \frac{288H^2Q_{\mathsf{max}}^2 \log\Big(2H(K+1)\mathcal{N}^{\mathsf{max}}_{\varepsilon/24H}(\mathcal{Q})/\delta\Big)}{\varepsilon^2} \right\rceil,$$

outputs a policy π_{out} such that, with probability at least $1-\delta$, $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \varepsilon$ using

$$\widetilde{\mathcal{O}}\left(\frac{H^3Q_{\mathsf{max}}^2 \log\Big(2\mathcal{N}^{\mathsf{max}}_{\varepsilon/24H}(\mathcal{Q})/\delta\Big)}{\varepsilon^2} \right) \qquad \textit{expert queries.}$$

For $\varepsilon \ge Q_{\mathsf{max}}$, any policy already satisfies the performance guarantee, so no expert queries are needed.

Although the expert-query rate improves from $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ to $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ when comparing to Theorem F.1, the number of learner rollouts remains of order $\widetilde{\mathcal{O}}(\varepsilon^{-4})$. Indeed, the total number of learner rollouts is

$$H(n_{\mathsf{E}} + Kn_{\mathsf{L}}).$$

This is of order

$$\widetilde{\mathcal{O}}\left(\frac{H^5Q_{\mathsf{max}}^4 \log(A) \log\Big(2\mathcal{N}^{\mathsf{max}}_{\varepsilon/24H}(\mathcal{Q})/\delta\Big)}{\varepsilon^4} \right).$$

An avenue for reducing this rollout cost is to recursively estimate changes in the learner term between consecutive policies on the same fresh states. The small changes induced by the softmax updates could allow smaller batches, with periodic full estimates or damping to control accumulated error. This could improve the explicit ε^{-4} rollout dependence while preserving the ε^{-2} expert-query dependence and the linear maximization oracle. We leave the development and analysis of this variance-reduced extension to future work.

Proof of Theorem I.1. As in the analysis of OVI, we first separate the performance gap into policy regret and estimation error, then control these two terms. For a decision rule $p\colon \mathcal{X} \to \Delta(\mathcal{A})$ and a function $Q_h \in \mathcal{Q}_h$, define the iteration-dependent empirical objective and recall its population counterpart:

$$\widehat{\mathcal{L}}_{h,k}(p, Q_h) := \frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - \frac{1}{n_{\mathsf{L}}} \sum_{j=1}^{n_{\mathsf{L}}} Q_h\Big(\boldsymbol{x}_h^{jk}, p\Big),$$
$$\mathcal{L}_h^{\pi_{\mathsf{out}}}(p, Q_h) := \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \langle Q_h(x,\cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - p(\cdot \,|\, x) \rangle.$$

For each stage $h \in [H]$, iteration $k \in [K]$, and decision rule p, let

$$\Delta_{h,k}(p) := \sup_{Q_h \in \mathcal{Q}_h} \Big| \widehat{\mathcal{L}}_{h,k}(p, Q_h) - \mathcal{L}_h^{\pi_{\mathsf{out}}}(p, Q_h) \Big|.$$

Unlike in the analysis of OVI, the empirical objective $\widehat{\mathcal{L}}_{h,k}$ and the estimation error $\Delta_{h,k}$ depend explicitly on k because a fresh learner batch is drawn at each iteration. Repeating the argument for the regret and estimation error decomposition (Lemma F.1), with Q_h^k best responding to $\widehat{\mathcal{L}}_{h,k}(\pi_h^k, \cdot)$ over $\mathcal{Q}_h$ at each iteration, gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K} \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \mathfrak{R}_h^{\pi}(x) + \frac{2}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \Delta_{h,k}\big(\pi_h^k\big). \tag{22}$$

Here, for fixed h and x, the policy player's regret is the sum below. The policy update is entropic mirror descent with loss vectors $\ell_k = -Q_h^k(x,\cdot)$ and comparator $\pi_{\mathsf{E},h}(\cdot \mid x)$. The mirror descent regret bound (Lemma E.6) therefore gives

$$\mathfrak{R}_h^\pi(x) = \sum_{k=1}^{K} \big\langle Q_h^k(x,\cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - \pi_h^k(\cdot \mid x) \big\rangle \le \frac{\log|\mathcal{A}|}{\eta} + \eta K Q_{\mathsf{max}}^2.$$

Indeed, the uniform initial decision rule has negative-entropy Bregman divergence at most $\log(A)$ to the comparator, all iterates have full support, and $\|Q_h^k\|_\infty \le Q_{\mathsf{max}}$. This regret bound holds for every realized sequence of critics, including adaptively chosen ones. Substituting η gives $\mathfrak{R}_h^\pi(x) \le 2Q_{\mathsf{max}}\sqrt{K\log(A)}$. Since each $d_h^{\pi_{\mathsf{out}}}$ is a probability distribution, the choice of K implies

$$\frac{1}{K}\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x)\mathfrak{R}_h^\pi(x) \le 2HQ_{\mathsf{max}}\sqrt{\frac{\log(A)}{K}} \le \frac{\varepsilon}{3}. \tag{23}$$

Controlling the estimation error. It remains to control $\Delta_{h,k}(\pi_h^k)$ simultaneously over $h \in [H]$ and $k \in [K]$. By the triangle inequality, we can separate the expert and learner terms:

$$\begin{aligned}\Delta_{h,k}\big(\pi_h^k\big) \le \sup_{Q_h\in\mathcal{Q}_h}\left|\frac{1}{n_{\mathsf{E}}}\sum_{i=1}^{n_{\mathsf{E}}} Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h(x,\pi_{\mathsf{E},h})\right| \\ + \sup_{Q_h\in\mathcal{Q}_h}\left|\frac{1}{n_{\mathsf{L}}}\sum_{j=1}^{n_{\mathsf{L}}} Q_h\Big(\boldsymbol{x}_h^{jk}, \pi_h^k\Big) - \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h\big(x,\pi_h^k\big)\right|.\end{aligned}$$

Fix a stage h. Let $\mathcal{F}_{h,0}$ be the history before collecting the expert buffer at that stage. The distribution $d_h^{\pi_{\mathsf{out}}}$ depends only on the previously learned decision rules $\pi_{\mathsf{out},1},\dots,\pi_{\mathsf{out},h-1}$, so it is $\mathcal{F}_{h,0}$-measurable. For $i \in [n_{\mathsf{E}}]$, let $\mathcal{F}_{h,i}$ be the sigma-algebra generated by $\mathcal{F}_{h,0}$ and the first i expert-labeled pairs at stage h. For any fixed function $Q_h\colon \mathcal{X}\times\mathcal{A}\to[0,Q_{\mathsf{max}}]$, define

$$Y_i(Q_h) := Q_h\big(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i\big) - \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h(x,\pi_{\mathsf{E},h}).$$

Conditional on the preceding history, the sampling rule for the expert buffer gives

$$\mathbb{E}[Y_i(Q_h) \mid \mathcal{F}_{h,i-1}] = 0, \qquad |Y_i(Q_h)| \le Q_{\mathsf{max}} \quad \text{almost surely.}$$

Thus, the expert increments form a martingale difference sequence with respect to $(\mathcal{F}_{h,i})_{i=0}^{n_{\mathsf{E}}}$.

For the learner term, fix $k \in [K]$ and let $\mathcal{G}_{h,k,0}$ be the history immediately before drawing the fresh learner batch at iteration k. This history includes the expert buffer, the preceding learner batches, and the already computed decision rule π_h^k. Both $d_h^{\pi_{\mathsf{out}}}$ and π_h^k are $\mathcal{G}_{h,k,0}$-measurable. For $j \in [n_{\mathsf{L}}]$, let $\mathcal{G}_{h,k,j}$ be the sigma-algebra generated by $\mathcal{G}_{h,k,0}$ and the first j states of this fresh batch. Define

$$\tilde{Y}_j^k(Q_h) := Q_h\Big(\boldsymbol{x}_h^{jk}, \pi_h^k\Big) - \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) Q_h\big(x,\pi_h^k\big).$$

Because the batch is drawn after π_h^k is fixed, its conditional sampling rule gives

$$\mathbb{E}\Big[\tilde{Y}_j^k(Q_h)\,\Big|\,\mathcal{G}_{h,k,j-1}\Big] = 0, \qquad \Big|\tilde{Y}_j^k(Q_h)\Big| \le Q_{\mathsf{max}} \quad \text{almost surely.}$$

The learner increments therefore form a martingale difference sequence with respect to $(\mathcal{G}_{h,k,j})_{j=0}^{n_{\mathsf{L}}}$. In both cases, the bound is Q_{max} because the function evaluation and its conditional mean lie in $[0,Q_{\mathsf{max}}]$. Applying the Azuma–Hoeffding inequality (Lemma E.1) conditional on the corresponding initial history gives, for every $t>0$,

$$\mathbb{P}\left[\left|\frac{1}{n_{\mathsf{L}}}\sum_{j=1}^{n_{\mathsf{L}}}\tilde{Y}_j^k(Q_h)\right| > t \,\middle|\, \mathcal{G}_{h,k,0}\right] \le 2e^{-\frac{n_{\mathsf{L}}t^2}{2Q_{\mathsf{max}}^2}}, \text{ and } \mathbb{P}\left[\left|\frac{1}{n_{\mathsf{E}}}\sum_{i=1}^{n_{\mathsf{E}}} Y_i(Q_h)\right| > t \,\middle|\, \mathcal{F}_{h,0}\right] \le 2e^{-\frac{n_{\mathsf{E}}t^2}{2Q_{\mathsf{max}}^2}}.$$

Taking expectations gives the same unconditional tail bounds. Set $t = \varepsilon/(12H)$ and $r_\varepsilon = \varepsilon/(24H)$. For each h, choose a deterministic r_ε-cover $\mathcal{C}_{h,r_\varepsilon}$ of $\mathcal{Q}_h$ in the ℓ_∞ norm, with $|\mathcal{C}_{h,r_\varepsilon}| \le \mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})$

and all centers taking values in $[0, Q_{\max}]$. Such centers can be obtained by clipping any minimal cover pointwise to this interval, which preserves the covering radius and cannot increase its cardinality. The fixed-function bounds above therefore apply to every center.

A union bound over the cover centers, the H expert buffers, and the HK learner batches gives a total failure probability at most

$$2H(K+1)\mathcal{N}_{r_\varepsilon}^{\max}(\mathcal{Q})e^{-\frac{n_{\mathsf{L}}t^2}{2Q_{\max}^2}} \le \delta,$$

where we used $n_{\mathsf{L}} = n_{\mathsf{E}}$ and the sample size in the theorem statement. Hence, on a single event of probability at least $1-\delta$, the expert and learner deviations are each at most t on their respective covers, simultaneously for all h and k. In particular, for every $h \in [H]$ and $k \in [K]$,

$$\sup_{Q_h \in \mathcal{C}_{h,r_\varepsilon}} \left| \frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} Y_i(Q_h) \right| + \sup_{Q_h \in \mathcal{C}_{h,r_\varepsilon}} \left| \frac{1}{n_{\mathsf{L}}} \sum_{j=1}^{n_{\mathsf{L}}} \tilde{Y}_j^k(Q_h) \right| \le 2t = \frac{\varepsilon}{6H}.$$

To extend this event from the covers to the full classes, fix $Q_h \in \mathcal{Q}_h$ and choose a cover representative $\tilde{Q}_h \in \mathcal{C}_{h,r_\varepsilon}$ with $\|Q_h - \tilde{Q}_h\|_\infty \le r_\varepsilon$. Each empirical average and each population expectation changes by at most r_ε when replacing Q_h by $\tilde{Q}_h$. Consequently,

$$\left| \frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} Y_i(Q_h) - \frac{1}{n_{\mathsf{E}}} \sum_{i=1}^{n_{\mathsf{E}}} Y_i\left(\tilde{Q}_h\right) \right| \le 2r_\varepsilon,$$

and

$$\left| \frac{1}{n_{\mathsf{L}}} \sum_{j=1}^{n_{\mathsf{L}}} \tilde{Y}_j^k(Q_h) - \frac{1}{n_{\mathsf{L}}} \sum_{j=1}^{n_{\mathsf{L}}} \tilde{Y}_j^k\left(\tilde{Q}_h\right) \right| \le 2r_\varepsilon.$$

These approximation bounds hold for every decision rule, including the data-dependent π_h^k. Combining them with the simultaneous cover bounds and the triangle inequality yields

$$\max_{h\in[H],\, k\in[K]} \Delta_{h,k}\left(\pi_h^k\right) \le 2t + 4r_\varepsilon = \frac{\varepsilon}{3H}. \tag{24}$$

Combining the performance-gap decomposition (22), the policy-regret bound (23), and the uniform estimation bound (24), we obtain, on the same event of probability at least $1-\delta$,

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{\varepsilon}{3} + 2H \max_{h\in[H],\, k\in[K]} \Delta_{h,k}\left(\pi_h^k\right) \le \varepsilon.$$

The expert is queried only when collecting the buffers, so the total number of expert queries is Hn_{E}. Since $\varepsilon < Q_{\max}$, rounding the sample counts changes their order only by constant factors. Substituting the stated sample size gives the claimed expert-query bound, with the additional logarithmic dependence on H and K absorbed into $\widetilde{\mathcal{O}}$. □

J Q-OVI: An Inefficient Algorithm with Improved Statistical Guarantees

In this section, we present an alternative algorithmic scheme that first updates the Q-variables via a no-regret algorithm. The resulting method, Q-OVI, is computationally inefficient when $\mathcal{Q}$ is continuous because it requires discretizing $\mathcal{Q}$, but it enjoys better statistical rates when $\mathcal{Q}$ is nonconvex. When the Q-player moves first, the policy player π_h^k can be chosen greedily with respect to Q_h^k. For each $(k,h) \in [K] \times [H]$, define

$$\boldsymbol{x}_h^k \sim d_h^{\pi_{\mathsf{out}}}, \qquad \boldsymbol{a}_{\mathsf{E},h}^k \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_h^k), \qquad \widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q) = Q(\boldsymbol{x}_h^k, \pi_h^k) - Q(\boldsymbol{x}_h^k, \boldsymbol{a}_{\mathsf{E},h}^k).$$

To complete the design of Q-OVI, we use Lemma J.1 below and note that the sequence $(Q_h^k)_{k=1}^K$ can be chosen to minimize a regret with respect to losses $\widehat{\ell}_h^{\pi_{\mathsf{out}},1}, \dots, \widehat{\ell}_h^{\pi_{\mathsf{out}},K}$. In the following, for any $h \in [H]$, $\mathcal{C}_\epsilon(\mathcal{Q}_h)$ denotes an ϵ-cover of $\mathcal{Q}_h$, and we abbreviate $\mathcal{N}_\epsilon^{\mathsf{max}}(\mathcal{Q}) := \max_{h\in[H]} \mathcal{N}_\epsilon(\mathcal{Q}_h)$. Then, for a weight distribution $w_h^k \in \Delta(\mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h))$ over a covering set $\mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$ of $\mathcal{Q}_h$, we define Q_h^k as the corresponding weighted average. For any state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$Q_h^k(x,a) = \sum_{Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)} w_h^k(Q) Q(x,a).$$

The next lemma specifies how to choose the weights.

Algorithm 4 Q-OVI: Q-First On-Policy Value-Based Imitation Learning

input: Learning rate η, iterations K, finite covering sets $\mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h) \subseteq \mathcal{Q}_h$ for $h \in [H]$.
For any $h \in [H]$ and $Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$, set $w_h^1(Q) = 1/\mathcal{N}_{\varepsilon/(4H)}(\mathcal{Q}_h)$.
For any $h \in [H]$ and $(x,a) \in \mathcal{X} \times \mathcal{A}$, set $Q_h^1(x,a) = \sum_{Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)} w_h^1(Q) Q(x,a)$.
For any $h \in [H]$ and $x \in \mathcal{X}$, set $\pi_h^1(\cdot \,|\, x) \in \arg\max_{p \in \Delta(\mathcal{A})} \langle p, Q_h^1(x,\cdot)\rangle$.
for $h = 1, \dots, H$ **do**
  for $k = 1, \dots, K$ **do**
    Sample $\boldsymbol{x}_h^k \sim d_h^{\pi_{\mathsf{out}}}$, $\boldsymbol{a}_{\mathsf{E},h}^k \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_h^k)$.
    For any $Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$, set $w_h^{k+1}(Q) \propto w_h^k(Q) e^{\eta(Q(\boldsymbol{x}_h^k, \boldsymbol{a}_{\mathsf{E},h}^k) - Q(\boldsymbol{x}_h^k, \pi_h^k))}$.
    For any $(x,a) \in \mathcal{X} \times \mathcal{A}$, set $Q_h^{k+1}(x,a) = \sum_{Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)} w_h^{k+1}(Q) Q(x,a)$.
    For any $x \in \mathcal{X}$, set $\pi_h^{k+1}(\cdot \,|\, x) \in \arg\max_{p \in \Delta(\mathcal{A})} \langle p, Q_h^{k+1}(x,\cdot)\rangle$.
  Set $\pi_{\mathsf{out},h}(a \,|\, x) = \frac{1}{K} \sum_{k=1}^K \pi_h^k(a \,|\, x)$ for any $(x,a) \in \mathcal{X} \times \mathcal{A}$.

Lemma J.1. *Let Assumption 2.1 hold, and fix an accuracy parameter $\varepsilon > 0$. For each $(k,h) \in [K] \times [H]$, suppose that $w_h^k \in \Delta(\mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h))$ and that Q_h^k is the corresponding weighted average. For each $h \in [H]$, define $\pi_h^k(\cdot \,|\, x) \in \arg\max_{p \in \Delta(\mathcal{A})} \langle p, Q_h^k(x,\cdot)\rangle$, and set $\pi_{\mathsf{out},h}(a \,|\, x) = \frac{1}{K} \sum_{k=1}^K \pi_h^k(a \,|\, x)$ for any state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$. Let $\widehat{Q}_h^{\pi_{\mathsf{E}}} \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$ be a closest element to $Q_h^{\pi_{\mathsf{E}}}$ for each $h \in [H]$. Then, with probability at least $1-\delta$, we have*

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{H}{K} \max_{h \in [H]} \mathfrak{R}_h^w + 8HQ_{\mathsf{max}} \sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2},$$

where the regret of the sequence $(w_h^k)_{k=1}^K$ is defined by

$$\mathfrak{R}_h^w = \sum_{k=1}^K \left(\sum_{Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)} w_h^k(Q) \widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q) - \widehat{\ell}_h^{\pi_{\mathsf{out}},k}\left(\widehat{Q}_h^{\pi_{\mathsf{E}}}\right) \right).$$

Lemma J.1 says that, with probability at least $1-\delta$, the performance gap $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}$ is upper bounded by $\frac{H}{K} \max_{h\in[H]} \mathfrak{R}_h^w + 8HQ_{\mathsf{max}} \sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2}$. $\mathfrak{R}_h^w$ can be easily controlled by updating $(w_h^k)_{k=1}^K$ via exponential weights. The resulting algorithm is in Algorithm 4. Therefore, using the decomposition in Lemma J.1 and showing that $\mathfrak{R}_h^w = \mathcal{O}(\sqrt{K})$ via a standard online learning bound for the regret of the exponential weights algorithm (see, e.g., Cesa-Bianchi and Lugosi, 2006; Orabona, 2026), we obtain the following guarantees.

Theorem J.1. *Let Assumption 2.1 hold. For any $\varepsilon, \delta \in (0,1)$, recall that $\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q}) := \max_{h\in[H]} \mathcal{N}_{\varepsilon/(4H)}(\mathcal{Q}_h)$. Run Q-OVI (Algorithm 4) with*

$$\eta = \sqrt{\frac{\log \mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})}{K Q^2_{\mathsf{max}}}}, \qquad K = \mathcal{O}\left(\frac{H^2 Q^2_{\mathsf{max}} \log\left(\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q}) H/\delta\right)}{\varepsilon^2}\right),$$

Then, the algorithm outputs a policy π_{out} such that, with probability at least $1-\delta$, we have $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \varepsilon$ after

$$\mathcal{O}\left(\frac{H^3 Q^2_{\mathsf{max}} \log\left(\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q}) H/\delta\right)}{\varepsilon^2}\right) \qquad \textit{expert queries.}$$

Despite having strong statistical guarantees, Algorithm 4 is not computationally attractive when the class is large or continuous because the construction of the covering set is cumbersome. We compare Q-OVI and OVI below.

Comparison of the computational complexity. As mentioned, Algorithm 1 is better suited for a practical implementation because it does not require discretizing the class $\mathcal{Q}$. While we think that the linear case could be made efficient, this approach is clearly not scalable when $\mathcal{Q}$ corresponds to the class of weights for modern architectures. Q-OVI would learn a distribution over a set of exponentially many weight configurations. In contrast, OVI is compatible with the common practice of specifying a loss function and backpropagating through it.

Comparison of the statistical complexity. When the class $\mathcal{Q}$ is convex, both algorithms achieve a rate of order $\mathcal{O}(\varepsilon^{-2})$. A rate difference emerges in the nonconvex case in which Q-OVI attains optimal rates in ε for both expert queries and learner rollouts: Algorithm 4 has an expert query complexity of order $\mathcal{O}(\varepsilon^{-2})$, while Algorithm 1 has a suboptimal $\mathcal{O}(\varepsilon^{-4})$ rate. Moreover, the learner rollouts required by Q-OVI is of order $\mathcal{O}(\varepsilon^{-2})$ while OVI and OVI+ require $\mathcal{O}(\varepsilon^{-4})$.

To summarize, Q-OVI (Algorithm 4) has better statistical properties but is computationally inefficient compared to OVI (Algorithm 1) because it requires discretizing the class $\mathcal{Q}$.

J.1 Proof of Lemma J.1 (Suboptimality Gap Decomposition for Q-OVI)

Proof of Lemma J.1. By the performance difference lemma (Lemma E.9) and by the definition of $\pi_{\mathsf{out},h}$, we have

$$\begin{aligned}
J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &= \sum_{h=1}^{H} \mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}) \\
&= \frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}_h^{\pi_{\mathsf{out}}}\left(\pi_h^k, Q_h^{\pi_{\mathsf{E}}}\right) \\
&= \frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}_h^{\pi_{\mathsf{out}}}\left(\pi_h^k, Q_h^k\right) + \frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}_h^{\pi_{\mathsf{out}}}\left(\pi_h^k, Q_h^{\pi_{\mathsf{E}}} - Q_h^k\right).
\end{aligned}$$

Since $\pi_h^k(\cdot \,|\, x) \in \arg\max_{p\in\Delta(\mathcal{A})} \langle p, Q_h^k(x,\cdot)\rangle$, we have $\mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h^k, Q_h^k) \le 0$. Hence,

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}_h^{\pi_{\mathsf{out}}}\left(\pi_h^k, Q_h^{\pi_{\mathsf{E}}} - Q_h^k\right).$$

Let $r > 0$ and let $\widehat{Q}_h^{\pi_{\mathsf{E}}} \in \mathcal{C}_r(\mathcal{Q}_h)$ be such that $\|\widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^{\pi_{\mathsf{E}}}\|_\infty \le r$ for each $h \in [H]$. For any decision rule p and functions Q_h, Q_h',

$$|\mathcal{L}_h^{\pi_{\mathsf{out}}}(p, Q_h - Q_h')| \le \|Q_h - Q_h'\|_\infty \sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) \|\pi_{\mathsf{E},h}(\cdot \,|\, x) - p(\cdot \,|\, x)\|_1 \le 2\|Q_h - Q_h'\|_\infty.$$

Thus, replacing $Q_h^{\pi_{\mathsf{E}}}$ by $\widehat{Q}_h^{\pi_{\mathsf{E}}}$ costs at most $2r$ for each pair (k,h), hence at most $2Hr$ after averaging over k and summing over h. Replacing $Q_h^{\pi_{\mathsf{E}}}$ by $\widehat{Q}_h^{\pi_{\mathsf{E}}}$ in the previous display gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathcal{L}_h^{\pi_{\mathsf{out}}}\left(\pi_h^k, \widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k\right) + 2Hr. \tag{25}$$

For each $(k,h) \in [K] \times [H]$, define the sampled loss

$$\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q) = Q\big(\boldsymbol{x}_h^k, \pi_h^k\big) - Q\big(\boldsymbol{x}_h^k, \boldsymbol{a}_{\mathsf{E},h}^k\big), \qquad \boldsymbol{x}_h^k \sim d_h^{\pi_{\mathsf{out}}}, \qquad \boldsymbol{a}_{\mathsf{E},h}^k \sim \pi_{\mathsf{E},h}\big(\cdot \,\big|\, \boldsymbol{x}_h^k\big).$$

Then,

$$\mathbb{E}\Big[\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q_h^k) - \widehat{\ell}_h^{\pi_{\mathsf{out}},k}\Big(\widehat{Q}_h^{\pi_{\mathsf{E}}}\Big) \,\Big|\, \mathcal{F}_{k-1,h}\Big] = \mathcal{L}_h^{\pi_{\mathsf{out}}}\Big(\pi_h^k, \widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k\Big),$$

where $\mathcal{F}_{k-1,h}$ contains all randomness from previous layers and from the first $k-1$ samples at layer h. Conditionally on $\mathcal{F}_{k-1,h}$, the distribution $d_h^{\pi_{\mathsf{out}}}$ is fixed because it only depends on the already computed policies $\pi_{\mathsf{out},1}, \dots, \pi_{\mathsf{out},h-1}$, and Q_h^k and π_h^k are also fixed because they are functions of the previous samples at layer h. Therefore, the increments

$$M_h^k = \mathcal{L}_h^{\pi_{\mathsf{out}}}\Big(\pi_h^k, \widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k\Big) - \Big(\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q_h^k) - \widehat{\ell}_h^{\pi_{\mathsf{out}},k}\Big(\widehat{Q}_h^{\pi_{\mathsf{E}}}\Big)\Big)$$

form a martingale difference sequence in k, for each fixed $h \in [H]$. Moreover, since $\widehat{Q}_h^{\pi_{\mathsf{E}}}$ belongs to the cover and Q_h^k is a convex combination of cover elements, both functions are bounded in sup-norm by Q_{max}. Thus, $\|\widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k\|_\infty \le 2Q_{\mathsf{max}}$, and both $|\mathcal{L}_h^{\pi_{\mathsf{out}}}(\pi_h^k, \widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k)|$ and $|\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q_h^k) - \widehat{\ell}_h^{\pi_{\mathsf{out}},k}(\widehat{Q}_h^{\pi_{\mathsf{E}}})|$ are at most $2\|\widehat{Q}_h^{\pi_{\mathsf{E}}} - Q_h^k\|_\infty \le 4Q_{\mathsf{max}}$. Hence $|M_h^k| \le 8Q_{\mathsf{max}}$ almost surely. By the Azuma-Hoeffding inequality and a union bound over $h \in [H]$, with probability at least $1-\delta$, for all $h \in [H]$,

$$\frac{1}{K}\sum_{k=1}^K M_h^k \le 8Q_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}}.$$

Summing over $h \in [H]$ incurs an extra H factor in the upper bound. Finally, using $Q_h^k = \sum_{Q \in \mathcal{C}_r(\mathcal{Q}_h)} w_h^k(Q) Q$ and taking $r = \varepsilon/(4H)$, we obtain from Eq. (25)

$$\begin{aligned}
J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &\le \frac{1}{K}\sum_{h=1}^H \sum_{k=1}^K \left(\sum_{Q \in \mathcal{C}_r(\mathcal{Q}_h)} w_h^k(Q) \widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q) - \widehat{\ell}_h^{\pi_{\mathsf{out}},k}\Big(\widehat{Q}_h^{\pi_{\mathsf{E}}}\Big) \right) \\
&\quad + 8HQ_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}} + 2Hr \\
&= \frac{1}{K}\sum_{h=1}^H \mathfrak{R}_h^w + 8HQ_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2} \\
&\le \frac{H}{K}\max_{h \in [H]} \mathfrak{R}_h^w + 8HQ_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2},
\end{aligned}$$

which gives the stated form. □

J.2 Proof of Theorem J.1 (Sample Complexity Guarantee for Q-OVI)

Proof of Theorem J.1. From the decomposition proven in Lemma J.1, with probability at least $1-\delta$, we have

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \frac{H}{K}\max_{h \in [H]} \mathfrak{R}_h^w + 8HQ_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2}.$$

Thus, it remains to control the regret $\mathfrak{R}_h^w$ for each $h \in [H]$. The update in Algorithm 4 can be written for any $Q \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$ as

$$w_h^{k+1}(Q) \propto w_h^k(Q) e^{\eta\left(Q\left(\boldsymbol{x}_h^k, \boldsymbol{a}_{\mathsf{E},h}^k\right) - Q\left(\boldsymbol{x}_h^k, \pi_h^k\right)\right)}.$$

Equivalently, this is exponential weights with sampled losses $\widehat{\ell}_h^{\pi_{\mathsf{out}},k}$. Moreover, by the definition of the function class, each $Q_h \in \mathcal{Q}_h$ satisfies $\|Q_h\|_\infty \le Q_{\mathsf{max}}$, so $|\widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q_h)| \le 2Q_{\mathsf{max}}$ for all $Q_h \in \mathcal{Q}_h$. If $\mathcal{N}_{\varepsilon/(4H)}^{\mathsf{max}}(\mathcal{Q}) = 1$, every stagewise cover is a singleton, so the prescribed learning rate is $\eta = 0$ but w_h^k assigns unit mass to $\widehat{Q}_h^{\pi_{\mathsf{E}}}$ for every h and k. Hence, $\mathfrak{R}_h^w = 0$ for all h, and the regret bound below holds directly without invoking Lemma E.6. We may therefore suppose that $\mathcal{N}_{\varepsilon/(4H)}^{\mathsf{max}}(\mathcal{Q}) \ge 2$, so $\eta > 0$.

For any stage with a singleton cover, the same zero-regret conclusion holds. Fix a stage $h \in [H]$ with a non-singleton cover and apply Lemma E.6 to the simplex $V = \Delta(\mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h))$ with loss vectors $\ell_k(Q) = \widehat{\ell}_h^{\pi_{\mathsf{out}},k}(Q)$. The update above is mirror descent with the negative entropy regularizer, whose Bregman divergence D is the KL divergence. We use as comparator $\widehat{Q}_h^{\pi_{\mathsf{E}}} \in \mathcal{C}_{\varepsilon/(4H)}(\mathcal{Q}_h)$. Since w_h^1 is uniform over the cover, $D(\boldsymbol{e}_{\widehat{Q}_h^{\pi_{\mathsf{E}}}}, w_h^1) \le \log \mathcal{N}_{\varepsilon/(4H)}(\mathcal{Q}_h)$, where $\boldsymbol{e}_{\widehat{Q}_h^{\pi_{\mathsf{E}}}}$ is the unit vector at $\widehat{Q}_h^{\pi_{\mathsf{E}}}$. Thus, using the bounds $\|\ell_k\|_\infty \le 2Q_{\mathsf{max}}$ and $\lambda \ge 1/2$, Lemma E.6 gives

$$\mathfrak{R}_h^w \le \frac{\log \mathcal{N}_{\varepsilon/(4H)}(\mathcal{Q}_h)}{\eta} + 4\eta K Q_{\mathsf{max}}^2.$$

Using $\mathfrak{R}_h^w = 0$ for singleton stagewise covers and substituting $\eta = \sqrt{\log \mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})/(KQ_{\mathsf{max}}^2)}$ in the preceding bound for non-singleton covers, we obtain

$$\max_{h\in[H]} \mathfrak{R}_h^w \le 5\sqrt{KQ_{\mathsf{max}}^2 \log \mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})}.$$

Combining the decomposition above with the regret bound gives

$$\begin{aligned} J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &\le 5H\sqrt{\frac{Q_{\mathsf{max}}^2 \log \mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})}{K}} + 8HQ_{\mathsf{max}}\sqrt{\frac{2\log(2H/\delta)}{K}} + \frac{\varepsilon}{2} \\ &\le \left(5 + 8\sqrt{2}\right) HQ_{\mathsf{max}} \sqrt{\frac{\log\left(\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})2H/\delta\right)}{K}} + \frac{\varepsilon}{2}. \end{aligned}$$

Choosing K so that

$$K \ge 4\left(5 + 8\sqrt{2}\right)^2 \frac{H^2 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})2H/\delta\right)}{\varepsilon^2}$$

makes the first term in the display above at most $\varepsilon/2$, and therefore $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \varepsilon$ with probability at least $1-\delta$. Since Algorithm 4 makes one expert query for each pair $(k,h) \in [K] \times [H]$, the total number of expert queries is

$$HK = \mathcal{O}\left(H^3 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{\varepsilon/(4H)}(\mathcal{Q})H/\delta\right)\varepsilon^{-2}\right).$$

□

J.3 Learning a Stationary Policy with Q-OVI

In this section, we show that Q-OVI can be used to learn a stationary policy that competes with the expert in the discounted infinite-horizon setting. We formalize the setting next.

Infinite-horizon MDPs. We consider a discounted MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, r, P, \gamma, \nu_0)$, where $\mathcal{X}$ and $\mathcal{A}$ are finite state and action spaces, $r\colon \mathcal{X}\times\mathcal{A} \to [0,1]$ is the reward function, $P\colon \mathcal{X}\times\mathcal{A} \to \Delta(\mathcal{X})$ is the transition kernel, $\gamma \in [0,1)$ is the discount factor, and $\nu_0 \in \Delta(\mathcal{X})$ is the initial distribution. A policy is a mapping $\pi\colon \mathcal{X} \to \Delta(\mathcal{A})$. Rolling out π generates a trajectory $(\boldsymbol{x}_h, \boldsymbol{a}_h)_{h=0}^\infty$ by drawing $\boldsymbol{x}_0 \sim \nu_0$, $\boldsymbol{a}_h \sim \pi(\cdot \mid \boldsymbol{x}_h)$, and $\boldsymbol{x}_{h+1} \sim P(\cdot \mid \boldsymbol{x}_h, \boldsymbol{a}_h)$ for all $h \ge 0$. For any state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, the discounted state-action value function is

$$Q_\gamma^\pi(x,a) = \mathbb{E}^\pi\left[\sum_{h=0}^\infty \gamma^h r(\boldsymbol{x}_h, \boldsymbol{a}_h) \,\middle|\, \boldsymbol{x}_0 = x, \boldsymbol{a}_0 = a\right].$$

For any function $Q\colon \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ and policy π, we use $Q(x,\pi)$ to denote $\mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|x)}[Q(x,\boldsymbol{a})]$. The discounted state occupancy measure of π is denoted by $d_\gamma^\pi \in \Delta(\mathcal{X})$ and defined by

$$d_\gamma^\pi(x) = (1-\gamma)\sum_{h=0}^\infty \gamma^h \mathbb{P}^\pi[\boldsymbol{x}_h = x].$$

Finally, we denote the normalized expected return of π by

$$J_\gamma^\pi = (1-\gamma)\mathbb{E}^\pi\left[\sum_{h=0}^{\infty}\gamma^h r(\boldsymbol{x}_h, \boldsymbol{a}_h)\right].$$

Infinite-horizon Q-OVI and guarantees. For simplicity, we consider finite classes $\mathcal{Q}$ in this section. The result can be extended to infinite classes via covering numbers, as in the finite-horizon case. We study the following infinite-horizon version of Q-OVI. For simplicity, we assume direct access

Algorithm 5 Infinite-horizon Q-OVI

input: Learning rate η, iterations K, finite class $\mathcal{Q}$.
Set $w_1(Q) = 1/|\mathcal{Q}|$ for all $Q \in \mathcal{Q}$.
For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, set $Q_1(x, a) = \sum_{Q\in\mathcal{Q}} w_1(Q)Q(x, a)$.
For any $x \in \mathcal{X}$, set $\pi_1(\cdot \mid x) \in \arg\max_{p\in\Delta(\mathcal{A})}\langle p, Q_1(x, \cdot)\rangle$.
for $k = 1, \dots, K$ **do**
  Sample $\boldsymbol{x}_k \sim d_\gamma^{\pi_k}$, $\boldsymbol{a}_k^{\mathsf{E}} \sim \pi_{\mathsf{E}}(\cdot \mid \boldsymbol{x}_k)$.
  For any $Q \in \mathcal{Q}$, set $w_{k+1}(Q) \propto w_k(Q)e^{\eta(Q(\boldsymbol{x}_k,\boldsymbol{a}_k^{\mathsf{E}})-Q(\boldsymbol{x}_k,\pi_k))}$.
  For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, set $Q_{k+1}(x, a) = \sum_{Q\in\mathcal{Q}} w_{k+1}(Q)Q(x, a)$.
  For any $x \in \mathcal{X}$, set $\pi_{k+1}(\cdot \mid x) \in \arg\max_{p\in\Delta(\mathcal{A})}\langle p, Q_{k+1}(x, \cdot)\rangle$.
Draw $I \sim \mathrm{Unif}([K])$ and output $\pi_{\mathsf{out}} = \pi_I$.

to samples from $d_\gamma^{\pi_k}$. Under ordinary trajectory access, each sample can instead be generated by a geometric rollout, yielding expected environment-interaction complexity $\mathcal{O}(K/(1-\gamma))$, an additional factor of $1/(1-\gamma)$. The algorithm still makes one expert query per round.

The resulting guarantee is as follows.

Theorem J.2. *Assume $Q_\gamma^{\pi_{\mathsf{E}}} \in \mathcal{Q}$, $|\mathcal{Q}| \geq 2$ and $\|Q\|_\infty \leq (1-\gamma)^{-1}$ for all $Q \in \mathcal{Q}$. For any $\varepsilon \in (0, 1)$, run Algorithm 5 with*

$$\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{Q}|}{K}}, \qquad K = \mathcal{O}\left(\frac{\log|\mathcal{Q}|}{(1-\gamma)^2\varepsilon^2}\right).$$

Then, the algorithm outputs a policy π_{out} such that

$$\mathbb{E}\left[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_{\mathsf{out}}}\right] \leq \varepsilon.$$

Proof of Theorem J.2. Since $\pi_{\mathsf{out}} = \pi_I$ with $I \sim \mathrm{Unif}([K])$, it is enough to control the average gap.

$$\mathbb{E}\left[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_{\mathsf{out}}}\right] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_k}\right].$$

For any fixed $k \in [K]$, the performance difference lemma in discounted MDPs (Moulin et al., 2025a, Lemma 1) gives

$$\begin{aligned}
J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_k} &= \sum_{x\in\mathcal{X}} d_\gamma^{\pi_k}(x)\left\langle Q_\gamma^{\pi_{\mathsf{E}}}(x, \cdot), \pi_{\mathsf{E}}(\cdot \mid x) - \pi_k(\cdot \mid x)\right\rangle \\
&= \sum_{x\in\mathcal{X}} d_\gamma^{\pi_k}(x)\langle Q_k(x, \cdot), \pi_{\mathsf{E}}(\cdot \mid x) - \pi_k(\cdot \mid x)\rangle \\
&\quad + \sum_{x\in\mathcal{X}} d_\gamma^{\pi_k}(x)\left\langle Q_\gamma^{\pi_{\mathsf{E}}}(x, \cdot) - Q_k(x, \cdot), \pi_{\mathsf{E}}(\cdot \mid x) - \pi_k(\cdot \mid x)\right\rangle.
\end{aligned}$$

Because $\pi_k(\cdot \mid x)$ is greedy with respect to $Q_k(x, \cdot)$, the first term is nonpositive. Therefore,

$$J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_k} \leq \sum_{x\in\mathcal{X}} d_\gamma^{\pi_k}(x)\left\langle Q_\gamma^{\pi_{\mathsf{E}}}(x, \cdot) - Q_k(x, \cdot), \pi_{\mathsf{E}}(\cdot \mid x) - \pi_k(\cdot \mid x)\right\rangle.$$

Define the sampled gains

$$g_k(Q) = Q\left(\boldsymbol{x}_k, \boldsymbol{a}_k^{\mathsf{E}}\right) - Q(\boldsymbol{x}_k, \pi_k), \qquad \boldsymbol{x}_k \sim d_\gamma^{\pi_k}, \qquad \boldsymbol{a}_k^{\mathsf{E}} \sim \pi_{\mathsf{E}}(\cdot \mid \boldsymbol{x}_k).$$

Let $\mathcal{F}_{k-1}$ be the history before drawing $(\boldsymbol{x}_k, \boldsymbol{a}_k^{\mathsf{E}})$, that is, the σ-field generated by the previous samples $(\boldsymbol{x}_j, \boldsymbol{a}_j^{\mathsf{E}})_{j<k}$ and any additional algorithmic randomness up to round $k-1$. Although the distribution $d_\gamma^{\pi_k}$ changes with k, conditional on $\mathcal{F}_{k-1}$, the weight vector w_k, and hence Q_k, π_k, and $d_\gamma^{\pi_k}$, are fixed. The fresh sample is then drawn conditionally as $\boldsymbol{x}_k \sim d_\gamma^{\pi_k}$ and $\boldsymbol{a}_k^{\mathsf{E}} \sim \pi_{\mathsf{E}}(\cdot \,|\, \boldsymbol{x}_k)$. Hence,

$$\mathbb{E}\big[g_k\big(Q_\gamma^{\pi_{\mathsf{E}}}\big) - g_k(Q_k) \,\big|\, \mathcal{F}_{k-1}\big] = \sum_{x \in \mathcal{X}} d_\gamma^{\pi_k}(x) \big\langle Q_\gamma^{\pi_{\mathsf{E}}}(x, \cdot) - Q_k(x, \cdot), \pi_{\mathsf{E}}(\cdot \,|\, x) - \pi_k(\cdot \,|\, x) \big\rangle.$$

Taking expectations and summing over k gives

$$\sum_{k=1}^{K} \mathbb{E}\big[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_k}\big] \le \mathbb{E}\left[\sum_{k=1}^{K} \big(g_k\big(Q_\gamma^{\pi_{\mathsf{E}}}\big) - g_k(Q_k)\big)\right].$$

Let $w^\star = \boldsymbol{e}_{Q_\gamma^{\pi_{\mathsf{E}}}}$ be the unit vector at $Q_\gamma^{\pi_{\mathsf{E}}}$. Since $Q_k(x,a) = \sum_{Q \in \mathcal{Q}} w_k(Q) Q(x,a)$, the previous display becomes

$$\sum_{k=1}^{K} \mathbb{E}\big[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_k}\big] \le \mathbb{E}\left[\sum_{k=1}^{K} \sum_{Q \in \mathcal{Q}} (w^\star(Q) - w_k(Q)) g_k(Q)\right].$$

The exponential weights update is mirror descent over the simplex $V = \Delta(\mathcal{Q})$ with loss vectors $\ell_k(Q) = -g_k(Q)$. The regularizer is negative entropy, whose Bregman divergence D is the KL divergence. We use the comparator $w^\star$, and since w_1 is uniform over $\mathcal{Q}$, we have $D(w^\star, w_1) \le \log|\mathcal{Q}|$. Moreover, $\|\ell_k\|_\infty \le 2(1-\gamma)^{-1}$, and the negative entropy is 1-strongly convex in the ℓ_1-norm. Thus, Lemma E.6 yields, pathwise,

$$\sum_{k=1}^{K} \sum_{Q \in \mathcal{Q}} (w^\star(Q) - w_k(Q)) g_k(Q) \le \frac{\log|\mathcal{Q}|}{\eta} + \frac{2\eta K}{(1-\gamma)^2}.$$

Assuming $|\mathcal{Q}| \ge 2$, we can set $\eta = (1-\gamma)\sqrt{\log|\mathcal{Q}|/K}$ to obtain

$$\mathbb{E}\big[J_\gamma^{\pi_{\mathsf{E}}} - J_\gamma^{\pi_{\mathsf{out}}}\big] \le 3\sqrt{\frac{\log|\mathcal{Q}|}{K(1-\gamma)^2}}.$$

Choosing $K \ge \frac{9\log|\mathcal{Q}|}{(1-\gamma)^2\varepsilon^2}$ makes the right-hand side at most ε. □

K Breaking the Offline Lower Bound with Coverage

This section shows that the offline lower bound does not apply when the expert is sufficiently exploratory, as measured by a coverage coefficient introduced in Assumptions K.1 and K.2 below. We prove this through a generalized analysis of OVI.

Assumption K.1 (L_∞-coverage)**.** *Assume an algorithm* Alg *outputs a VI policy* π_{out} *in a class* Π_{Alg} *such that, for some finite constant* C_∞*, any stage* h*, and any state* x*,* $\sup_{\pi\in\Pi_{\mathrm{Alg}}} d_h^{\pi}(x)/d_h^{\pi_{\mathsf{E}}}(x) \le C_\infty$.

Assumption K.2 (L_1-coverage)**.** *Assume an algorithm* Alg *that outputs a VI policy* π_{out} *in a class* Π_{Alg} *such that, for some finite constant* C_1*,*

$$\sup_{\pi\in\Pi_{\mathrm{Alg}}} \sum_{h=1}^{H} \sum_{x\in\mathcal{X}} \frac{(d_h^{\pi}(x))^2}{d_h^{\pi_{\mathsf{E}}}(x)} \le C_1.$$

For both cases, we use the convention $0/0 = 0$. We note that Assumption K.1 implies Assumption K.2 with $C_1 \le HC_\infty$, but we keep the two assumptions separate because Theorem K.3 tracks both dependencies. Under this setting our main theorem reads as follows.

Theorem K.1. *Let* $\varepsilon, \delta \in (0,1)$ *and Assumption 2.1 hold. Assume the expert policy* π_{E} *is optimal and that the algorithm* SPOIL *(Moulin et al., 2025a) satisfies Assumption K.1. For every radius* $r > 0$*, write* $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q}) := \max_{h\in[H]} \mathcal{N}_r(\mathcal{Q}_h, \|\cdot\|_\infty)$*, and choose a radius* $r_\varepsilon = \mathcal{O}(\varepsilon^2)$ *small enough. Then,* SPOIL *returns an* ε*-optimal policy with probability at least* $1-\delta$ *using* $\widetilde{\mathcal{O}}(C_\infty^4 H^5 Q_{\mathsf{max}}^4 \log(A) \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\delta^{-1})\varepsilon^{-4})$ *precollected expert data. Moreover, if the class* $\mathcal{Q}$ *is convex the bound improves to* $\widetilde{\mathcal{O}}(C_\infty^2 H^3 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\delta^{-1})\varepsilon^{-2})$.

Theorem K.1 follows from the stronger Theorem K.3 proven below, which also gives a bound under Assumption K.2.

K.1 OVI with Arbitrary Sampling Distributions

We first prove a generalization of Algorithm 1 in which the state sampling distribution at stage h may differ from $d_h^{\pi_{\mathsf{out}}}$. This result will be used several times below.

Algorithm 6 OVI with arbitrary state sampling distribution d

input: Sampling distributions $d = \{d_h\}_{h=1}^H$, learning-rate schedule $\{\eta_k\}_{k=1}^{K+1}$, iterations K, number of expert queries per stage n_{E}.
for $h = 1, \dots, H$ **do**
  Create the stage-h dataset: for $i \in [n_{\mathsf{E}}]$, sample $\boldsymbol{x}_h^i \sim d_h$ and query $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^i)$.
  Initialize $\pi_h^1 = \mathrm{Unif}(\mathcal{A})$.
  for $k = 1, \dots, K$ **do**
    Set $Q_h^k \in \arg\max_{Q_h\in\mathcal{Q}_h} \sum_{i=1}^{n_{\mathsf{E}}} (Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i) - Q_h(\boldsymbol{x}_h^i, \pi_h^k(\boldsymbol{x}_h^i)))$.
    Set $\pi_h^{k+1}(a \mid x) \propto \pi_h^1(a \mid x) \exp(\eta_{k+1} \sum_{k'=1}^{k} Q_h^{k'}(x,a))$.
  Set the output policy at stage h to $\pi_{\mathsf{out},h}(a \mid x) = \frac{1}{K}\sum_{k=1}^{K} \pi_h^k(a \mid x)$.

Theorem K.2 (Sample complexity for OVI with arbitrary sampling distributions)**.** *For each* $h \in [H]$*, let* $\mathcal{F}_{h-1}$ *denote the history available before the stage-*h *samples are drawn. Fix any sequence of state sampling distributions* $d = \{d_h\}_{h=1}^H$*, where each* d_h *may be random but is* $\mathcal{F}_{h-1}$*-measurable, and run Algorithm 6 with* $K, n_{\mathsf{E}} \in \mathbb{N}$ *and the FTRL learning-rate schedule*

$$\eta_k = \sqrt{\frac{\log A}{k Q_{\mathsf{max}}^2}} \qquad \text{for every } k \ge 1.$$

For each $h \in [H]$*, assume that, conditionally on* $\mathcal{F}_{h-1}$*, the pairs* $(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)_{i=1}^{n_{\mathsf{E}}}$ *are independent, with* $\boldsymbol{x}_h^i \sim d_h$ *and* $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^i)$*. For every radius* $r > 0$*, write* $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q}) := \max_{h\in[H]} \mathcal{N}_r(\mathcal{Q}_h, \|\cdot\|_\infty)$*. Let* $\varepsilon, \delta \in (0,1)$ *and choose a radius* $r_\varepsilon = \mathcal{O}(\varepsilon^2)$ *small enough. Fix any comparator policy* π' *such that* $Q_h^{\pi'} \in \mathcal{Q}_h$ *for every* $h \in [H]$*. For the policy* π_{out} *returned by*

Algorithm 6, we recall the definition

$$\mathcal{L}_h^d\Big(\pi_{\mathsf{out},h}, Q_h^{\pi'}\Big) = \sum_{x\in\mathcal{X}} d_h(x)\Big\langle Q_h^{\pi'}(x,\cdot), \pi_{\mathsf{E},h}(\cdot\,|\,x) - \pi_{\mathsf{out},h}(\cdot\,|\,x)\Big\rangle.$$

If $\mathcal{Q}$ is convex, then with probability at least $1-\delta$,

$$\sum_{h=1}^H \mathcal{L}_h^d\Big(\pi_{\mathsf{out},h}, Q_h^{\pi'}\Big) \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\big)}{n_{\mathsf{E}}}} + \varepsilon\right).$$

If $\mathcal{Q}$ is not assumed to be convex, then with probability at least $1-\delta$,

$$\sum_{h=1}^H \mathcal{L}_h^d\Big(\pi_{\mathsf{out},h}, Q_h^{\pi'}\Big) \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{KH^2 Q_{\mathsf{max}}^2 \log\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\big)}{n_{\mathsf{E}}}} + \varepsilon\right).$$

In particular, up to logarithmic factors, the convex bound is of order ε by taking $K \gtrsim H^2 Q_{\mathsf{max}}^2 \log(A)/\varepsilon^2$ and $n_{\mathsf{E}} \gtrsim H^2 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta)/\varepsilon^2$. In the nonconvex case, the same choice of K and $n_{\mathsf{E}} \gtrsim KH^2 Q_{\mathsf{max}}^2 \log(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta)/\varepsilon^2$ suffices.

Proof of Theorem K.2. Recall the definitions of the empirical objective and the uniform estimation error

$$\widehat{\mathcal{L}}_h^d(\pi_h, Q_h) := n_{\mathsf{E}}^{-1} \sum_{i=1}^{n_{\mathsf{E}}} \big(Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i) - Q_h(\boldsymbol{x}_h^i, \pi_h)\big),$$

$$\Delta_d(\pi) := \max_{h\in[H]} \sup_{Q_h\in\mathcal{Q}_h} \Big|\widehat{\mathcal{L}}_h^d(\pi_h, Q_h) - \mathcal{L}_h^d(\pi_h, Q_h)\Big|.$$

Fix $h \in [H]$. Since $Q_h^{\pi'} \in \mathcal{Q}_h$, the first inequality below follows from the definition of $\Delta_d(\pi^k)$. The second inequality follows from the empirical maximization property defining Q_h^k, and the last inequality uses the definition of $\Delta_d(\pi^k)$ again, now with Q_h^k:

$$\begin{aligned}
\sum_{k=1}^K \mathcal{L}_h^d\Big(\pi_h^k, Q_h^{\pi'}\Big) &\le \sum_{k=1}^K \widehat{\mathcal{L}}_h^d\Big(\pi_h^k, Q_h^{\pi'}\Big) + \sum_{k=1}^K \Delta_d\big(\pi^k\big) \\
&\le \sum_{k=1}^K \widehat{\mathcal{L}}_h^d\big(\pi_h^k, Q_h^k\big) + \sum_{k=1}^K \Delta_d\big(\pi^k\big) \\
&\le \sum_{k=1}^K \mathcal{L}_h^d\big(\pi_h^k, Q_h^k\big) + 2\sum_{k=1}^K \Delta_d\big(\pi^k\big).
\end{aligned}$$

We now control the first term on the right-hand side using the FTRL regret bound. Fix $x \in \mathcal{X}$ and define the loss vectors $\ell_{k,h,x}(a) := -Q_h^k(x,a)$. Since π_h^1 is uniform, the update in Algorithm 6 implies that, for every $k \ge 1$,

$$\pi_h^k(a\,|\,x) = \frac{\exp\Big(\eta_k \sum_{j<k} Q_h^j(x,a)\Big)}{\sum_{a'\in\mathcal{A}} \exp\Big(\eta_k \sum_{j<k} Q_h^j(x,a')\Big)} = \frac{\exp\Big(-\eta_k \sum_{j<k} \ell_{j,h,x}(a)\Big)}{\sum_{a'\in\mathcal{A}} \exp\Big(-\eta_k \sum_{j<k} \ell_{j,h,x}(a')\Big)}.$$

Thus, for this fixed h and x, the sequence $(\pi_h^k(\cdot\,|\,x))_{k=1}^K$ is exactly the FTRL sequence of Lemma E.7 with comparator $\pi_{\mathsf{E},h}(\cdot\,|\,x)$. Moreover, $\|\ell_{k,h,x}\|_\infty \le Q_{\mathsf{max}}$ and $\eta_k = \sqrt{\log A/(kQ_{\mathsf{max}}^2)}$. Applying Lemma E.7 pointwise in x and then averaging over d_h, for any $K \in \mathbb{N}$, we obtain

$$\sum_{k=1}^K \mathcal{L}_h^d\big(\pi_h^k, Q_h^k\big) = \sum_{x\in\mathcal{X}} d_h(x) \sum_{k=1}^K \big\langle Q_h^k(x,\cdot), \pi_{\mathsf{E},h}(\cdot\,|\,x) - \pi_h^k(\cdot\,|\,x)\big\rangle \le 3\sqrt{Q_{\mathsf{max}}^2 K \log A},$$

where the last inequality uses that d_h is a probability distribution. Compared with the proof of Theorem F.1, note that we use FTRL rather than online mirror descent.[16]

[16] This change gives anytime bounds that hold for every K, which is needed because we invoke Theorem K.2 with different values of K.

It remains to control the estimation error. Since the learning rate varies with k, define the policy class

$$\Pi_{\mathcal{Q}} := \left\{ \begin{array}{c} \pi : \exists m \in \{0, \dots, K\}, \forall h, \exists Q_h^1, \dots, Q_h^m \in \mathcal{Q}_h, \\ \pi_h(\cdot \mid x) = \mathrm{softmax}\Big(\eta_{m+1} \sum_{j=1}^m Q_h^j(x, \cdot)\Big) \end{array} \right\}.$$

For each h, let $\Pi_{\mathcal{Q},h} := \{\pi_h : \pi \in \Pi_{\mathcal{Q}}\}$. The iterates generated by Algorithm 6 satisfy $\pi^k \in \Pi_{\mathcal{Q}}$ for every $k \in [K]$, by taking $m = k - 1$.

For every h, condition on $\mathcal{F}_{h-1}$. Then d_h is fixed, and the pairs $(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)_{i=1}^{n_{\mathsf{E}}}$ are independent, with $\boldsymbol{x}_h^i \sim d_h$ and $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^i)$. Applying the uniform concentration bound in Lemma E.2 with $\Pi_h = \Pi_{\mathcal{Q},h}$ gives that, for any radius $r > 0$, with probability at least $1 - \delta$,

$$\sum_{k=1}^K \Delta_d(\pi^k) \le K \sup_{\pi \in \Pi_{\mathcal{Q}}} \Delta_d(\pi) \le K \left(4r + Q_{\mathsf{max}} \sqrt{\frac{8 \log\Big(2 \sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho)/\delta\Big)}{n_{\mathsf{E}}}} \right).$$

Therefore, dividing by K and using $\mathcal{L}_h^d(\pi_{\mathsf{out},h}, Q_h^{\pi'}) = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_h^d(\pi_h^k, Q_h^{\pi'})$, we obtain

$$\begin{aligned} \sum_{h=1}^H \mathcal{L}_h^d\Big(\pi_{\mathsf{out},h}, Q_h^{\pi'}\Big) &\le 3\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + 8Hr \\ &\quad + 2HQ_{\mathsf{max}} \sqrt{\frac{8 \log\Big(2 \sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho)/\delta\Big)}{n_{\mathsf{E}}}}. \end{aligned}$$

Next, we bound the covering number of the product class. Choose $r = \varepsilon/(8H)$. The product-cover lemmas in Lemma E.4 and Lemma E.3 are stated for a fixed learning rate η, but their proofs apply almost verbatim to the present class. For policies represented by m functions, the factor ηm is replaced by $\eta_{m+1} m$. Thus, they naturally lead to a covering number with radius

$$\frac{\varepsilon}{16H \max\{1, Q_{\mathsf{max}} \max_{1 \le m \le K} m\eta_{m+1}\}}.$$

Increasing K by an absolute constant if necessary, we may assume $K \log A \ge 1$. Then the learning rate schedule gives

$$\begin{aligned} Q_{\mathsf{max}} \max_{1 \le m \le K} m\eta_{m+1} &= K\sqrt{\frac{\log A}{K+1}} \le \sqrt{K \log A}, \\ \max\left\{1, Q_{\mathsf{max}} \max_{1 \le m \le K} m\eta_{m+1}\right\} &\le \sqrt{K \log A}. \end{aligned}$$

Thus, for the parameter choices below, it is enough to use the smaller radius

$$r_\varepsilon := \frac{\varepsilon}{16H\sqrt{K \log A}},$$

which only enlarges the covering number. After the choices of K below, this radius satisfies $r_\varepsilon = \mathcal{O}(\varepsilon^2)$.

Convex case. If $\mathcal{Q}$ is convex, then each projection $\mathcal{Q}_h$ is convex. Hence, by Lemma E.4, for every $h \in [H]$,

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \le (K+1) \mathcal{N}_{r_\varepsilon}(\mathcal{Q}_h, \|\cdot\|_\infty)^2.$$

Plugging this into the preceding estimation bound gives

$$\sum_{h=1}^H \mathcal{L}_h^d\Big(\pi_{\mathsf{out},h}, Q_h^{\pi'}\Big) \le 3\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \varepsilon + 2HQ_{\mathsf{max}} \sqrt{\frac{8 \log\Big(2H(K+1)\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\big)^2/\delta\Big)}{n_{\mathsf{E}}}}.$$

Thus, the convex bound becomes

$$\widetilde{\mathcal{O}}\left(\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log\big(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\big)}{n_{\mathsf{E}}}} + \varepsilon \right).$$

In particular, choosing

$$K \geq \frac{9H^2Q_{\mathsf{max}}^2 \log A}{\varepsilon^2}, \qquad n_{\mathsf{E}} \geq \frac{32H^2Q_{\mathsf{max}}^2}{\varepsilon^2} \log\left(\frac{2H(K+1)\left(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\right)^2}{\delta}\right)$$

makes the three terms in the previous display of order ε.

Nonconvex case. For arbitrary, not necessarily convex, classes $\mathcal{Q}$, Lemma E.3 instead gives, for every $h \in [H]$,

$$\mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho) \leq (K+1)\mathcal{N}_{r_\varepsilon}(\mathcal{Q}_h, \|\cdot\|_\infty)^{K+1}.$$

Therefore,

$$\sum_{h=1}^{H} \mathcal{L}_h^d\left(\pi_{\mathsf{out},h}, Q_h^{\pi'}\right) \leq 3\sqrt{\frac{H^2Q_{\mathsf{max}}^2 \log A}{K}} + \varepsilon + 2HQ_{\mathsf{max}}\sqrt{\frac{8\log\left(2H(K+1)\left(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\right)^{K+1}/\delta\right)}{n_{\mathsf{E}}}},$$

or, equivalently,

$$\widetilde{\mathcal{O}}\left(\sqrt{\frac{H^2Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{KH^2Q_{\mathsf{max}}^2 \log\left(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{n_{\mathsf{E}}}} + \varepsilon\right).$$

With the same choice of K as above, taking

$$n_{\mathsf{E}} \geq \frac{32H^2Q_{\mathsf{max}}^2}{\varepsilon^2} \log\left(\frac{2H(K+1)\left(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})\right)^{K+1}}{\delta}\right) = \widetilde{\mathcal{O}}\left(\frac{KH^2Q_{\mathsf{max}}^2 \log\left(\mathcal{N}_{r_\varepsilon}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{\varepsilon^2}\right)$$

makes the three terms of the nonconvex bound of order ε. This concludes the proof. □

K.2 Proof of Theorem K.1 (Offline IL with Coverage and Optimal Expert)

We prove Theorem K.1 through the stronger Theorem K.3. The sharper statement also gives a guarantee under Assumption K.2 alone in terms of the L_1-coverage coefficient C_1, which can be much smaller than HC_∞.

Theorem K.3. *Let $\varepsilon, \delta \in (0,1)$. Let Assumption 2.1 hold. Assume the expert policy π_{E} is optimal and that* SPOIL *(Moulin et al., 2025a) satisfies at least one of Assumptions K.1 and K.2. Run* SPOIL *(i.e., Algorithm 6 with $d = d^{\pi_{\mathsf{E}}}$) with the FTRL learning-rate schedule*

$$\eta_k = \sqrt{\frac{\log A}{kQ_{\mathsf{max}}^2}} \qquad \textit{for every } k \geq 1.$$

For every radius $r > 0$, write $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q}) := \max_{h\in[H]} \mathcal{N}_r(\mathcal{Q}_h, \|\cdot\|_\infty)$. Choose radii $r_\varepsilon^{(1)} = \mathcal{O}(\varepsilon^4)$ and $r_\varepsilon^{(\infty)} = \mathcal{O}(\varepsilon^2)$ small enough. If $\mathcal{Q}$ is arbitrary, the C_1-dependent choice below guarantees $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon$ with probability at least $1-\delta$ under Assumption K.2, while the C_∞-dependent choice guarantees the same under Assumption K.1:

$$K^{(1)} = \widetilde{\mathcal{O}}\left(\frac{C_1^2H^2Q_{\mathsf{max}}^4 \log A}{\varepsilon^4}\right), \qquad n_{\mathsf{E}}^{(1)} = \widetilde{\mathcal{O}}\left(\frac{C_1^4H^4Q_{\mathsf{max}}^8 \log(A) \log\left(\mathcal{N}_{r_\varepsilon^{(1)}}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{\varepsilon^8}\right),$$

$$K^{(\infty)} = \widetilde{\mathcal{O}}\left(\frac{C_\infty^2H^2Q_{\mathsf{max}}^2 \log A}{\varepsilon^2}\right), \qquad n_{\mathsf{E}}^{(\infty)} = \widetilde{\mathcal{O}}\left(\frac{C_\infty^4H^4Q_{\mathsf{max}}^4 \log(A) \log\left(\mathcal{N}_{r_\varepsilon^{(\infty)}}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{\varepsilon^4}\right).$$

If both coverage assumptions hold, choosing the better of the two branches gives total expert sample complexity

$$Hn_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\min\left\{\frac{C_1^4H^5Q_{\mathsf{max}}^8 \log(A) \log\left(\mathcal{N}_{r_\varepsilon^{(1)}}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{\varepsilon^8}, \frac{C_\infty^4H^5Q_{\mathsf{max}}^4 \log(A) \log\left(\mathcal{N}_{r_\varepsilon^{(\infty)}}^{\mathsf{max}}(\mathcal{Q})/\delta\right)}{\varepsilon^4}\right\}\right).$$

If $\mathcal{Q}$ is convex, the same branch-wise guarantees hold under their respective coverage assumptions, with the same choices of $K^{(1)}$ and $K^{(\infty)}$, but with

$$n_{\mathsf{E}}^{(1)} = \widetilde{\mathcal{O}}\left(\frac{C_1^2 H^2 Q_{\mathsf{max}}^4 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(1)}}(\mathcal{Q})/\delta\right)}{\varepsilon^4}\right), \quad n_{\mathsf{E}}^{(\infty)} = \widetilde{\mathcal{O}}\left(\frac{C_\infty^2 H^2 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(\infty)}}(\mathcal{Q})/\delta\right)}{\varepsilon^2}\right).$$

If both coverage assumptions hold, choosing the better of the two branches gives total expert sample complexity

$$Hn_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\min\left\{\frac{C_1^2 H^3 Q_{\mathsf{max}}^4 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(1)}}(\mathcal{Q})/\delta\right)}{\varepsilon^4}, \frac{C_\infty^2 H^3 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(\infty)}}(\mathcal{Q})/\delta\right)}{\varepsilon^2}\right\}\right).$$

Here, the $\widetilde{\mathcal{O}}$ notation in each branch hides logarithmic factors in H, Q_{max}, and the corresponding coverage coefficient.

Proof of Theorem K.3. Since the expert is optimal, π_{E} is greedy with respect to $Q^{\pi_{\mathsf{E}}}$ at every stage h and state $x \in \mathrm{supp}(d_h^{\pi_{\mathsf{E}}})$. Hence, for any policy π, stage h, and state $x \in \mathrm{supp}(d_h^{\pi_{\mathsf{E}}})$,

$$\langle Q_h^{\pi_{\mathsf{E}}}(x,\cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h(\cdot \,|\, x)\rangle \geq 0.$$

This nonnegativity is the only point at which expert optimality is used. It lets us remove absolute values in the change-of-measure arguments below. We prove two independent bounds: one under Assumption K.2, in terms of C_1, and one under Assumption K.1, in terms of C_∞. When both assumptions hold, the guarantee takes the better of the two.

Bounds scaling with C_1. Let

$$g_h(x) := \langle Q_h^{\pi_{\mathsf{E}}}(x,\cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_{\mathsf{out},h}(\cdot \,|\, x)\rangle.$$

Finiteness of C_1 implies that, for every h and x, $d_h^{\pi_{\mathsf{out}}}(x) = 0$ whenever $d_h^{\pi_{\mathsf{E}}}(x) = 0$, so the following change of measure is well defined. The expert value function satisfies $Q_h^{\pi_{\mathsf{E}}}(x,a) \in [0, Q_{\mathsf{max}}]$ for every h, x, a. Together with the nonnegativity above, this gives $0 \leq g_h(x) \leq Q_{\mathsf{max}}$, and hence $g_h(x)^2 \leq Q_{\mathsf{max}} g_h(x)$ for every $x \in \mathrm{supp}(d_h^{\pi_{\mathsf{E}}})$. By the performance difference lemma (Lemma E.9), the Cauchy–Schwarz inequality, and Assumption K.2, we have

$$\begin{aligned}
J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &= \sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) g_h(x) \\
&\leq \sqrt{\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} \frac{\left(d_h^{\pi_{\mathsf{out}}}(x)\right)^2}{d_h^{\pi_{\mathsf{E}}}(x)}} \sqrt{\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) g_h(x)^2} \\
&\leq \sqrt{C_1 Q_{\mathsf{max}} \sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) g_h(x)}.
\end{aligned}$$

The last summation is exactly the loss controlled by Theorem K.2 with $d = d^{\pi_{\mathsf{E}}}$ and $\pi' = \pi_{\mathsf{E}}$:

$$\sum_{h=1}^{H}\sum_{x\in\mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) g_h(x) = \sum_{h=1}^{H} \mathcal{L}_h^{\pi_{\mathsf{E}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}).$$

For nonconvex $\mathcal{Q}$, applying Theorem K.2 with accuracy parameter ε_0 and associated covering radius $r_{\varepsilon_0} = \mathcal{O}(\varepsilon_0^2)$ gives, with probability at least $1-\delta$,

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \widetilde{\mathcal{O}}\left(\sqrt{C_1 Q_{\mathsf{max}}\left(\sqrt{\frac{H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{K H^2 Q_{\mathsf{max}}^2}{n_{\mathsf{E}}} \log\left(\mathcal{N}^{\mathsf{max}}_{r_{\varepsilon_0}}(\mathcal{Q})/\delta\right)} + \varepsilon_0\right)}\right).$$

Taking $\varepsilon_0 = \varepsilon^2/(C_1 Q_{\mathsf{max}})$ and choosing $r_\varepsilon^{(1)} \leq r_{\varepsilon_0}$, the monotonicity of covering numbers and the absorption of logarithmic factors in H, Q_{max}, and C_1 into $\widetilde{\mathcal{O}}$ yield

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \widetilde{\mathcal{O}}\left(\sqrt[4]{\frac{C_1^2 H^2 Q_{\mathsf{max}}^4 \log A}{K}} + \sqrt[4]{\frac{K C_1^2 H^2 Q_{\mathsf{max}}^4 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(1)}}(\mathcal{Q})/\delta\right)}{n_{\mathsf{E}}}} + \varepsilon\right).$$

Thus the C_1-tuned nonconvex choice in the theorem makes the right-hand side at most $\mathcal{O}(\varepsilon)$. If $\mathcal{Q}$ is convex, the second term in Theorem K.2 has no factor K. Repeating the same calculation gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \widetilde{\mathcal{O}}\left(\sqrt[4]{\frac{C_1^2 H^2 Q_{\mathsf{max}}^4 \log A}{K}} + \sqrt[4]{\frac{C_1^2 H^2 Q_{\mathsf{max}}^4 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(1)}}(\mathcal{Q})/\delta\right)}{n_{\mathsf{E}}}} + \varepsilon\right),$$

which gives the improved convex choice of $n_{\mathsf{E}}^{(1)}$.

Bounds scaling with C_∞. The C_∞ argument gives a better dependence on $1/\varepsilon$ at the cost of replacing C_1 by the stronger coverage coefficient C_∞. Since $g_h(x) \ge 0$, Assumption K.1 gives

$$\begin{aligned} J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &= \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{out}}}(x) g_h(x) \\ &\le \left(\max_{y \in \mathcal{X}} \max_{h \in [H]} \frac{d_h^{\pi_{\mathsf{out}}}(y)}{d_h^{\pi_{\mathsf{E}}}(y)}\right) \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) |g_h(x)| \\ &\le C_\infty \sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) g_h(x) = C_\infty \sum_{h=1}^{H} \mathcal{L}_h^{\pi_{\mathsf{E}}}(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}). \end{aligned}$$

Applying Theorem K.2 with $d = d^{\pi_{\mathsf{E}}}$, $\pi' = \pi_{\mathsf{E}}$, and accuracy parameter $\varepsilon_0 = \varepsilon / C_\infty$, and choosing $r_\varepsilon^{(\infty)} \le r_{\varepsilon_0}$ for the associated radius $r_{\varepsilon_0} = \mathcal{O}(\varepsilon_0^2)$, the monotonicity of covering numbers gives for nonconvex $\mathcal{Q}$ that

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{C_\infty^2 H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{K C_\infty^2 H^2 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(\infty)}}(\mathcal{Q})/\delta\right)}{n_{\mathsf{E}}}} + \varepsilon\right).$$

This is at most $\mathcal{O}(\varepsilon)$ for the C_∞-tuned nonconvex choice of $K^{(\infty)}$ and $n_{\mathsf{E}}^{(\infty)}$ in the theorem. If $\mathcal{Q}$ is convex, the same calculation with the convex version of Theorem K.2 gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{C_\infty^2 H^2 Q_{\mathsf{max}}^2 \log A}{K}} + \sqrt{\frac{C_\infty^2 H^2 Q_{\mathsf{max}}^2 \log\left(\mathcal{N}^{\mathsf{max}}_{r_\varepsilon^{(\infty)}}(\mathcal{Q})/\delta\right)}{n_{\mathsf{E}}}} + \varepsilon\right),$$

which yields the stated convex choice of $n_{\mathsf{E}}^{(\infty)}$. Finally, SPOIL uses n_{E} expert samples at each of the H stages, so the total number of precollected expert samples is Hn_{E}. Multiplying the branch-wise bounds on n_{E} by H gives the two sample-complexity displays in the theorem. □

L On the Benefits of Mixing Expert and Learner Trajectories

In this section, we analyze the popular technique of mixing expert and learner data to create the empirical objective in LM distillation (Agarwal et al., 2024; Li et al., 2026a) and show that it leads to representational benefits. We develop an algorithm that is statistically efficient when the class $\mathcal{Q}$ satisfies either $Q^{\pi_{\mathsf{E}}}$-realizability or the $Q^{\Pi_{\mathcal{Q}}}$-realizability condition introduced by Moulin et al. (2025a), *without knowing a priori which representational condition holds*.

To obtain the anytime regret bounds needed in this section, we use a slightly different version of the $Q^{\Pi_{\mathcal{Q}}}$-realizability condition. Compared with Moulin et al. (2025a), our version uses FTRL rather than OMD and allows the softmax weights to depend on a fixed learning-rate schedule $\eta_{1:K+1}$.

Assumption L.1 ($Q^{\Pi_{\mathcal{Q}}}$-realizability)**.** *Fix* $K = \mathrm{poly}(H, \log A, Q_{\mathtt{max}}, \varepsilon^{-1})$ *and learning rates* $\eta_{1:K+1} \in \mathbb{R}^{K+1}$*, and define*

$$\Pi_{\mathcal{Q}} := \Big\{\pi : \forall h, \exists m_h \le K, \exists \big(Q_j^h\big)_{j=1}^{m_h} \subset \mathcal{Q}_h, \quad \pi_h(\cdot \,|\, x) = \mathrm{softmax}\Big(\eta_{m_h+1} \textstyle\sum_{j=1}^{m_h} Q_j^h(x, \cdot)\Big)\Big\}.$$

The class $\mathcal{Q}$ *satisfies* $Q^{\Pi_{\mathcal{Q}}}$*-realizability if* $Q^\pi \in \mathcal{Q}$ *for every* $\pi \in \Pi_{\mathcal{Q}}$.

For each stage h, write $\Pi_{\mathcal{Q},h} := \{\pi_h : \pi \in \Pi_{\mathcal{Q}}\}$. This notation is always understood with the same value of K and the same learning-rate schedule $\eta_{1:K+1}$ as the algorithm under consideration. This is a slightly broader policy class than the version with a single common value of m shared by all stages. We use it because the hybrid policies in the proof may take different stages from different FTRL rounds, so the number of accumulated Q-functions can depend on h.

This assumption requires realizing the Q-functions of many policies, whereas $Q^{\pi_{\mathsf{E}}}$-realizability only requires realizing the expert's Q-function. Since the expert need not belong to $\Pi_{\mathcal{Q}}$, $Q^{\Pi_{\mathcal{Q}}}$-realizability does not imply $Q^{\pi_{\mathsf{E}}}$-realizability. Nevertheless, the upper bound of Moulin et al. (2025a) implies that, under $Q^{\Pi_{\mathcal{Q}}}$-realizability, there exists a policy in $\Pi_{\mathcal{Q}}$ that is ε-optimal with respect to π_{E}.

As shown by Moulin et al. (2025a), offline IL is possible under $Q^{\Pi_{\mathcal{Q}}}$-realizability, and therefore interactive IL is possible as well. However, Algorithms 1 and 4 do not directly use this condition. We instead modify the state-sampling rule in Algorithm 1: for each $h \in [H]$, we sample states from a mixture of the expert occupancy $d_h^{\pi_{\mathsf{E}}}$ and the learner occupancy $d_h^{\pi_{\mathsf{out}}}$. Sampling from the former is possible in the interactive setting by rolling in with actions queried from the expert. Thus, Algorithm 7 is an instance of Algorithm 6 with sampling distribution

$$d_h = (1-\alpha) d_h^{\pi_{\mathsf{out}}} + \alpha d_h^{\pi_{\mathsf{E}}},$$

for an appropriate choice of $\alpha \in [0, 1]$. The mixture is the key adaptive device: the $d^{\pi_{\mathsf{out}}}$ component gives coverage of the learner states needed under $Q^{\pi_{\mathsf{E}}}$-realizability, while the $d^{\pi_{\mathsf{E}}}$ component keeps enough expert-state mass for the $Q^{\Pi_{\mathcal{Q}}}$ argument. The change relative to Algorithm 6 is highlighted

Algorithm 7 `RAOVI`: Representation-Adaptive `OVI`

input: Mixture parameter $\alpha \in [0, 1]$, iterations K, learning rates $(\eta_k)_{k=1}^{K+1}$, dataset size per stage n_{E}.
for $h = 1, \dots, H$ **do**
  Create the stage-h dataset: for $i \in [n_{\mathsf{E}}]$, sample

$$\boldsymbol{x}_h^i \sim (1-\alpha) d_h^{\pi_{\mathsf{out}}} + \alpha d_h^{\pi_{\mathsf{E}}}, \qquad \boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \,|\, \boldsymbol{x}_h^i).$$

  Initialize $\pi_h^1 = \mathrm{Unif}(\mathcal{A})$.
  for $k = 1, \dots, K$ **do**
    Set $Q_h^k \in \arg\max_{Q_h \in \mathcal{Q}_h} \sum_{i=1}^{n_{\mathsf{E}}} (Q_h(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i) - Q_h(\boldsymbol{x}_h^i, \pi_h^k))$.
    Set $\pi_h^{k+1}(a \,|\, x) \propto \pi_h^1(a \,|\, x) \exp(\eta_{k+1} \sum_{k'=1}^{k} Q_h^{k'}(x, a))$.
  Set $\pi_{\mathsf{out},h}(a \,|\, x) = \frac{1}{K} \sum_{k=1}^{K} \pi_h^k(a \,|\, x)$.

in orange. We have the following result.

Theorem L.1. *Let the expert policy π_{E} be optimal, $\varepsilon \in (0, \min\{1, 2HQ_{\mathsf{max}}\})$, and $\delta \in (0,1)$. Run Algorithm 7 with*

$$\alpha = 1 - \frac{\varepsilon}{2HQ_{\mathsf{max}}}, \qquad \eta_k = \sqrt{\frac{\log A}{kQ_{\mathsf{max}}^2}} \qquad \text{for every } k \geq 1.$$

Then, with probability at least $1-\delta$, each of the following guarantees holds under the corresponding condition.

1. *Under $Q^{\pi_{\mathsf{E}}}$-realizability (Assumption 2.1), $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon$ is guaranteed by the choices*

$$K = \widetilde{\mathcal{O}}\left(\frac{H^4 Q_{\mathsf{max}}^4 \log A}{\varepsilon^4}\right), \quad n_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\frac{H^4 Q_{\mathsf{max}}^4 \log\left(2H\mathcal{N}^{\mathsf{max}}_{\varepsilon^2/(8H^2Q_{\mathsf{max}})}(\mathcal{Q}\times\Pi_{\mathcal{Q}},\rho)/\delta\right)}{\varepsilon^4}\right).$$

2. *Under $Q^{\Pi_{\mathcal{Q}}}$-realizability (Assumption L.1), $J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} \leq \varepsilon$ is guaranteed by the choices*

$$K = \widetilde{\mathcal{O}}\left(\frac{H^2 Q_{\mathsf{max}}^2 \log A}{\varepsilon^2}\right), \quad n_{\mathsf{E}} = \widetilde{\mathcal{O}}\left(\frac{H^2 Q_{\mathsf{max}}^2 \log\left(2H\mathcal{N}^{\mathsf{max}}_{\varepsilon/H}(\mathcal{Q}\times\Pi_{\mathcal{Q}},\rho)/\delta\right)}{\varepsilon^2}\right).$$

Here, $\mathcal{N}_r^{\mathsf{max}}(\mathcal{Q}\times\Pi_{\mathcal{Q}},\rho) := \max_{h\in[H]}\mathcal{N}_r(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)$. The total number of expert queries is at most $H^2 n_{\mathsf{E}}$. Thus, without knowing which condition holds, one can choose K and n_{E} to be the componentwise maximum of the two displayed requirements, and the corresponding guarantee applies under either condition.

We remark that the rates in the $Q^{\pi_{\mathsf{E}}}$-realizable case are not optimal. It is an interesting open question whether they can be improved while maintaining the faster rates in the $Q^{\Pi_{\mathcal{Q}}}$-realizable case.

Proof of Theorem L.1. We analyze each case separately. Throughout, let

$$d_h = \alpha d_h^{\pi_{\mathsf{E}}} + (1-\alpha) d_h^{\pi_{\mathsf{out}}}.$$

As in the previous cases, d_h is measurable with respect to the history available before sampling at stage h, and conditionally on this history, the pairs $(\boldsymbol{x}_h^i, \boldsymbol{a}_{\mathsf{E},h}^i)_{i=1}^{n_{\mathsf{E}}}$ are independent with $\boldsymbol{x}_h^i \sim d_h$ and $\boldsymbol{a}_{\mathsf{E},h}^i \sim \pi_{\mathsf{E},h}(\cdot \mid \boldsymbol{x}_h^i)$, as required by Lemma E.2. Obtaining one labeled sample at stage h requires at most h expert queries, including the queries used to roll in under the expert. Summing over the n_{E} samples at each stage gives at most $n_{\mathsf{E}}\sum_{h=1}^H h \leq H^2 n_{\mathsf{E}}$ expert queries.

Case 1: $Q^{\pi_{\mathsf{E}}}$-realizability. For every $h\in[H]$ and $x\in\mathcal{X}$, define $g_h(x) := \langle Q_h^{\pi_{\mathsf{E}}}(x,\cdot), \pi_{\mathsf{E},h}(\cdot\mid x) - \pi_{\mathsf{out},h}(\cdot\mid x)\rangle$. A one-step deviation argument using expert optimality gives $g_h(x)\geq 0$ for $d_h^{\pi_{\mathsf{E}}}$-almost every x. Therefore, the definition of d_h and the performance difference lemma (Lemma E.9) give

$$\begin{aligned}\sum_{h=1}^H \mathbb{E}_{x\sim d_h}[g_h(x)] &= (1-\alpha)(J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}) + \alpha\sum_{h=1}^H \mathbb{E}_{x\sim d_h^{\pi_{\mathsf{E}}}}[g_h(x)] \\ &\geq (1-\alpha)(J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}}).\end{aligned}$$

Consequently,

$$\begin{aligned}J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} &\leq \frac{1}{1-\alpha}\sum_{h=1}^H\sum_{x\in\mathcal{X}} d_h(x)\langle Q_h^{\pi_{\mathsf{E}}}(x,\cdot), \pi_{\mathsf{E},h}(\cdot\mid x) - \pi_{\mathsf{out},h}(\cdot\mid x)\rangle \\ &= \frac{1}{1-\alpha}\sum_{h=1}^H \mathcal{L}_h^d(\pi_{\mathsf{out},h}, Q_h^{\pi_{\mathsf{E}}}) \\ &\leq \frac{1}{1-\alpha}\left(3\sqrt{\frac{H^2Q_{\mathsf{max}}^2\log A}{K}} + 8Hr + 2HQ_{\mathsf{max}}\sqrt{\frac{8\log\left(2\sum_{h=1}^H \mathcal{N}_r(\mathcal{Q}_h\times\Pi_{\mathcal{Q},h},\rho)/\delta\right)}{n_{\mathsf{E}}}}\right),\end{aligned}$$

where the last inequality holds with probability at least $1-\delta$ by rederiving the proof of Theorem K.2 with sampling distribution $d_h = \alpha d_h^{\pi_{\mathsf{E}}} + (1-\alpha) d_h^{\pi_{\mathsf{out}}}$ and comparator $\pi' = \pi_{\mathsf{E}}$, stopping at the intermediate product-cover bound, for any $r > 0$. Taking $r = \varepsilon^2/(8H^2 Q_{\mathsf{max}})$ gives $8Hr = \varepsilon^2/(HQ_{\mathsf{max}})$ and covering radius $\varepsilon^2/(8H^2 Q_{\mathsf{max}})$ in the logarithm. With $\alpha = 1 - \varepsilon/(2HQ_{\mathsf{max}})$, the choices of K and n_{E} in the $Q^{\pi_{\mathsf{E}}}$-realizable branch make the right-hand side at most $\mathcal{O}(\varepsilon)$.

Case 2: $Q^{\Pi_{\mathcal{Q}}}$-realizability. The subtle point is that we should not apply the performance difference lemma directly with $Q^{\pi_{\mathsf{out}}}$, because π_{out} is a pointwise average of softmax policies and need not belong to $\Pi_{\mathcal{Q}}$. Instead, we represent the trajectory distribution of π_{out} as a mixture over stagewise hybrids of the iterates. Let $I = (I_1, \dots, I_H)$ be sampled uniformly from $[K]^H$, and define the hybrid policy π^I by

$$\pi_h^I = \pi_h^{I_h} \qquad \text{for every } h \in [H].$$

We now justify carefully why these hybrids are related to the averaged policy. Let $\tau = (\boldsymbol{x}_1, \boldsymbol{a}_1, \dots, \boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{x}_{H+1})$ be a trajectory under policy π_{out}. By the definition of π_{out}, we have

$$\begin{aligned}
\mathbb{P}^{\pi_{\mathsf{out}}}(\tau) &= \nu_0(\boldsymbol{x}_1) \prod_{h=1}^{H} [\pi_{\mathsf{out},h}(\boldsymbol{a}_h \mid \boldsymbol{x}_h) P_h(\boldsymbol{x}_{h+1} \mid \boldsymbol{x}_h, \boldsymbol{a}_h)] \\
&= \nu_0(\boldsymbol{x}_1) \prod_{h=1}^{H} \left[\left(\frac{1}{K} \sum_{k_h=1}^{K} \pi_h^{k_h}(\boldsymbol{a}_h \mid \boldsymbol{x}_h) \right) P_h(\boldsymbol{x}_{h+1} \mid \boldsymbol{x}_h, \boldsymbol{a}_h) \right] \\
&= \frac{1}{K^H} \sum_{i_1, \dots, i_H \in [K]} \nu_0(\boldsymbol{x}_1) \prod_{h=1}^{H} \left[\pi_h^{i_h}(\boldsymbol{a}_h \mid \boldsymbol{x}_h) P_h(\boldsymbol{x}_{h+1} \mid \boldsymbol{x}_h, \boldsymbol{a}_h) \right] \\
&= \mathbb{E}_I \left[\mathbb{P}^{\pi^I}(\tau) \right].
\end{aligned}$$

The third line is where we use that the coordinates $I_1, \dots, I_H$ are sampled independently. If one sampled a single common index for all stages, the resulting average of products would not generally equal the product of averages defining π_{out}. Therefore, for any trajectory-level function F, and in particular for the cumulative reward $F(\tau) = \sum_{h=1}^{H} r_h(\boldsymbol{x}_h, \boldsymbol{a}_h)$, we can average over the finite trajectory space to obtain

$$\mathbb{E}^{\pi_{\mathsf{out}}}[F(\tau)] = \sum_{\tau} \mathbb{P}^{\pi_{\mathsf{out}}}(\tau) F(\tau) = \mathbb{E}_I \left[\sum_{\tau} \mathbb{P}^{\pi^I}(\tau) F(\tau) \right] = \mathbb{E}_I \left[\mathbb{E}^{\pi^I}[F(\tau)] \right].$$

Taking F to be the cumulative reward gives

$$J^{\pi_{\mathsf{out}}} = \mathbb{E}_I \left[J^{\pi^I} \right].$$

For every I, the hybrid policy π^I belongs to $\Pi_{\mathcal{Q}}$. Indeed, at each stage h, choose $m_h = I_h - 1$, the functions $Q_h^1, \dots, Q_h^{I_h - 1} \in \mathcal{Q}_h$ (empty if $I_h = 1$), and $\eta_{I_h} = \eta_{m_h+1}$. By Assumption L.1, we have $Q^{\pi^I} \in \mathcal{Q}$, and hence $Q_h^{\pi^I} \in \mathcal{Q}_h$ for every h. Applying the performance difference lemma (Lemma E.9) to each hybrid policy gives

$$J^{\pi_{\mathsf{E}}} - J^{\pi_{\mathsf{out}}} = \mathbb{E}_I \left[J^{\pi_{\mathsf{E}}} - J^{\pi^I} \right] = \mathbb{E}_I \left[\sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h^{\pi_{\mathsf{E}}}(x) \left\langle Q_h^{\pi^I}(x, \cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - \pi_h^{I_h}(\cdot \mid x) \right\rangle \right].$$

Using $d_h^{\pi_{\mathsf{E}}} = d_h + (1-\alpha)(d_h^{\pi_{\mathsf{E}}} - d_h^{\pi_{\mathsf{out}}})$, the right-hand side is equal to $T_1 + T_2$, where we define

$$T_1 := \mathbb{E}_I \left[\sum_{h=1}^{H} \sum_{x \in \mathcal{X}} d_h(x) \left\langle Q_h^{\pi^I}(x, \cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - \pi_h^{I_h}(\cdot \mid x) \right\rangle \right],$$

$$T_2 := (1-\alpha) \mathbb{E}_I \left[\sum_{h=1}^{H} \sum_{x \in \mathcal{X}} (d_h^{\pi_{\mathsf{E}}}(x) - d_h^{\pi_{\mathsf{out}}}(x)) \left\langle Q_h^{\pi^I}(x, \cdot), \pi_{\mathsf{E},h}(\cdot \mid x) - \pi_h^{I_h}(\cdot \mid x) \right\rangle \right].$$

We first control T_1. For a fixed stage h, write $J \in [K]^{H-1}$ for the collection of indices at all stages except h. For $k \in [K]$, let $I^{h,k,J}$ be the full index vector obtained by setting $I_h = k$ and $I_{-h} = J$, and define

$$Q_h^{k,J} := Q_h^{\pi^{I^{h,k,J}}}.$$

For every h, k, J, we have $Q_h^{k,J} \in \mathcal{Q}_h$. Using the independence and uniformity of I_h, we can rewrite T_1 as

$$T_1 = \sum_{h=1}^{H} \mathbb{E}_{I_h, I_{-h}} \left[\mathcal{L}_h^d \left(\pi_h^{I_h}, Q_h^{\pi^I} \right) \right] = \frac{1}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{E}_J \left[\mathcal{L}_h^d \left(\pi_h^k, Q_h^{k,J} \right) \right].$$

We now compare each realized comparator $Q_h^{k,J}$ to the empirical best response Q_h^k before averaging over J. With a slight abuse of notation, recall that π^k denotes the policy whose decision rule at stage h is π_h^k. Define the same uniform estimation error used in Theorem K.2 by

$$\Delta_d(\pi) = \max_{h \in [H]} \sup_{Q_h \in \mathcal{Q}_h} \left| \widehat{\mathcal{L}}_h^d(\pi_h, Q_h) - \mathcal{L}_h^d(\pi_h, Q_h) \right|.$$

For every fixed h, k, J, on this uniform-concentration event,

$$\mathcal{L}_h^d(\pi_h^k, Q_h^{k,J}) \le \widehat{\mathcal{L}}_h^d(\pi_h^k, Q_h^{k,J}) + \Delta_d(\pi^k) \le \widehat{\mathcal{L}}_h^d(\pi_h^k, Q_h^k) + \Delta_d(\pi^k) \le \mathcal{L}_h^d(\pi_h^k, Q_h^k) + 2\Delta_d(\pi^k).$$

The middle inequality is the empirical best-response property of Q_h^k, which we can use since $Q_h^{k,J}$ is one feasible element of $\mathcal{Q}_h$ by Assumption L.1. Averaging the last display over J, summing over h and k, and dividing by K, we obtain

$$T_1 \le \frac{1}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \mathcal{L}_h^d(\pi_h^k, Q_h^k) + \frac{2H}{K} \sum_{k=1}^{K} \Delta_d(\pi^k).$$

The first term is controlled by the FTRL regret bound stage by stage (Lemma E.7). Indeed, for each fixed h,

$$\sum_{k=1}^{K} \mathcal{L}_h^d(\pi_h^k, Q_h^k) = \sum_{x \in \mathcal{X}} d_h(x) \sum_{k=1}^{K} \left\langle Q_h^k(x, \cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x) \right\rangle \le 3\sqrt{Q_{\mathsf{max}}^2 K \log A}.$$

Since π^k is in $\Pi_{\mathcal{Q}}$, applying Lemma E.2 with $\Pi_h = \Pi_{\mathcal{Q},h}$ and radius $r > 0$ gives, on an event of probability at least $1 - \delta$,

$$\sum_{k=1}^{K} \Delta_d(\pi^k) \le 4Kr + KQ_{\mathsf{max}} \sqrt{\frac{8 \log\left(2 \sum_{h=1}^{H} \mathcal{N}_r(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho)/\delta \right)}{n_{\mathsf{E}}}}.$$

Combining the previous three displays, taking $r = \varepsilon/H$, and bounding the sum by the maximum gives

$$T_1 \le 3H \sqrt{\frac{Q_{\mathsf{max}}^2 \log A}{K}} + 8\varepsilon + 2HQ_{\mathsf{max}} \sqrt{\frac{8 \log\left(2H \max_h \mathcal{N}_{\varepsilon/H}(\mathcal{Q}_h \times \Pi_{\mathcal{Q},h}, \rho)/\delta \right)}{n_{\mathsf{E}}}}.$$

It remains to control T_2. Define, for each h, x, I,

$$g_h^I(x) := \left\langle Q_h^{\pi^I}(x, \cdot), \pi_{\mathsf{E},h}(\cdot \,|\, x) - \pi_h^{I_h}(\cdot \,|\, x) \right\rangle.$$

We have $Q_h^{\pi^I}(x, a) \in [0, Q_{\mathsf{max}}]$ for every x, a, h, I. Therefore

$$\left| g_h^I(x) \right| = \left| Q_h^{\pi^I}(x, \pi_{\mathsf{E},h}) - Q_h^{\pi^I}\left(x, \pi_h^{I_h}\right) \right| \le Q_{\mathsf{max}}.$$

Using the triangle inequality and the previous inequality, we can bound the signed change-of-measure term explicitly:

$$\begin{aligned}
|T_2| &\le (1-\alpha) \mathbb{E}_I \left[\sum_{h=1}^{H} \left| \sum_{x \in \mathcal{X}} (d_h^{\pi_{\mathsf{E}}}(x) - d_h^{\pi_{\mathsf{out}}}(x)) g_h^I(x) \right| \right] \\
&\le (1-\alpha) \sum_{h=1}^{H} \mathbb{E}_I \left[\sum_{x \in \mathcal{X}} |d_h^{\pi_{\mathsf{E}}}(x) - d_h^{\pi_{\mathsf{out}}}(x)| \left| g_h^I(x) \right| \right] \\
&\le (1-\alpha) Q_{\mathsf{max}} \sum_{h=1}^{H} \| d_h^{\pi_{\mathsf{E}}} - d_h^{\pi_{\mathsf{out}}} \|_1 \\
&\le 2(1-\alpha) H Q_{\mathsf{max}}.
\end{aligned}$$

In the last line, we used that both $d_h^{\pi_{\mathsf{E}}}$ and $d_h^{\pi_{\mathsf{out}}}$ are probability distributions, so their ℓ_1 distance is at most 2. With $\alpha = 1 - \varepsilon/(2HQ_{\mathsf{max}})$, the term T_2 is at most ε. The $Q^{\Pi_{\mathcal{Q}}}$-realizable choices of K and n_{E} make T_1 at most $\mathcal{O}(\varepsilon)$, proving the second branch for the same averaged output policy π_{out}. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, all claims in the abstract are reflected in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss the limitations of our set of assumptions and point out in the conclusions future directions of improvement.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, the complete proofs are in Appendices F and G.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we provide code and all needed information for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the code is open access and attached to the submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all training details are specified in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, all our results have error bars computed over 50 independent seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide details in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper respects the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is theoretical. We do not foresee a broader societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not foresee any risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit all creators of existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, all the code is well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects were involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs did not do any original contribution.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.