Imitation Learning in Discounted Linear MDP without exploration assumptions

Luca Viano
LIONS, EPFL
luca.viano@epfl.ch

Stratis Skoulakis
LIONS, EPFL
stratis.skoulakis@epfl.ch

Volkan Cevher
LIONS , EPFL
volkan.cevher@epfl.ch

Abstract

We present a new algorithm for imitation learning in infinite horizon linear MDPs dubbed ILARL which greatly improves the bound on the number of trajectories that the learner needs to sample from the environment. In particular, we remove exploration assumptions required in previous works and we improve the dependence on the desired accuracy ϵ from $\mathcal{O}\left(\epsilon^{-5}\right)$ to $\mathcal{O}\left(\epsilon^{-4}\right)$. Our result relies on a connection between imitation learning and online learning in MDPs with adversarial losses. For the latter setting, we present the first result for infinite horizon linear MDP which may be of independent interest. Moreover, we are able to provide a strengthen result for the finite horizon case where we achieve $\mathcal{O}\left(\epsilon^{-2}\right)$. Numerical experiments with linear function approximation shows that ILARL outperforms other commonly used algorithms.

1 Introduction

Imitation Learning (IL) is of extreme importance for all applications where designing a reward function is cumbersome while collecting demonstrations from an expert policy π_{E} is easy. Examples are autonomous driving [23], robotics [32], and economics/finance [10]. The goal is to learn a policy which competes with the expert policy under the true unknown cost function of the Markov Decision Process (MDP) [34] with discount factor γ. In particular we consider that the cost vector and the transition dynamics are linear in some state action dependent d-dimensional features $\phi(s,a) \in \mathbb{R}^d$ [1].

Imitation learning relies on two data resources: expert demonstrations collected acting with π_{E} and data that can be collected interacting in the MDP with policies chosen by the learning algorithm. The first approach known as behavioural cloning (BC) solves the problem applying supervised learning. That is, it requires no interaction in the MDP but it requires knowledge of a class Π such that $\pi_{\mathrm{E}} \in \Pi$ and $\widetilde{\mathcal{O}}\left(\frac{\log|\Pi|}{(1-\gamma)^4\epsilon_E^2}\right)$ expert demonstrations to ensure with high probability that the output policy is at most ϵ_E-suboptimal.

The quartic dependence on the effective horizon term ($(1-\gamma)^{-1}$) is problematic for long horizon problems. Moreover, the dependence on Π requires to make prior assumption on the expert policy structure to provide bounds which do not scale with the number of states in the function approximation setting. Thankfully, the dependence on the effective horizon can be improved resorting to MDP interaction. There exists an interesting line of works achieving this goal considering an interacting setting where the learner has the possibility to query the expert policy at any state visited during the MDP interaction [39, 38] or that require a generative model to implement efficiently the moment matching procedure [44]. Another recent work requires a generative model to sample the initial state of the trajectory from the expert occupancy measure [45]. In this work, we considered a different scenario which is adopted in most of applied imitation learning [20, 19, 15, 37, 11, 49, 16, 30]. In

[1] That is, we consider Linear MDP [22] whose formal definition is deferred to Section 2.1

17th European Workshop on Reinforcement Learning (ewrl 2024).

this case, the expert policy can not be queried but only a dataset of expert demonstrations collected beforehand is available.

The setting has received scarse theoretical attention so far. The only results we are aware of are: [41] that focus on the tabular, finite horizon case, [26] in the finite horizon linear mixture MDP setting and [48] in the infinite horizon Linear MDP setting. In all these works bound the number of required expert demonstrations scale as $(1-\gamma)^{-2}$ which improves considerably over the quartic depedence attained by BC. However, [48] made the following assumption on the features that greatly simplifies the exploration in the MDP.

Assumption. Persistent excitation *It holds that for any policy π^k in the sequence of policies generated by the algorithm adopted by the learner $\lambda_{\min}\left(\mathbb{E}_{s,a\sim d^{\pi^k}}\left[\phi(s,a)\phi(s,a)^\intercal\right]\right) \geq \beta > 0$.*

Despite being commonly used in infinite horizon function approximation setting (see for example [1, 17, 14, 24, 3, 5]), the persistent excitation assumption is very restrictive as it can be easily violated by deterministic policies with tabular features.

Our contribution We propose a new algorithm that improves the results of [48] in two important aspects: it bypasses the persistent excitation assumption (i.e. $\beta = 0$ does not cause the bound to blow up) and it improves the dependence on ϵ. In particular, the new proposed algorithm Algorithm 3 only requires $\mathcal{O}\left(\frac{d^3}{(1-\gamma)^8\epsilon^4}\right)$ MDP interactions which greatly improves upon the bound $\widetilde{\mathcal{O}}(\frac{d^2}{\beta^6(1-\gamma)^9\epsilon^5})$ proven by [48]. Moreover, it holds that $\beta \leq d^{-1}$. Therefore, the bound for PPIL scales at least as $\widetilde{\mathcal{O}}(\frac{d^8}{(1-\gamma)^9\epsilon^5})$. Therefore, the bound of ILARL is better in all the relevant parameters $d, (1-\gamma)$ and ϵ. The design is different from [48] and it builds on a connection between imitation learning and online learning in MDP with full information. Therefore, we design as a submodule of our algorithm the first algorithm for adversarial infinite horizon linear MDPs which achieves $\mathcal{O}(K^{3/4})$ pseudo-regret. We also consider the finite horizon version of this algorithm which obtains a regret bound $\widetilde{\mathcal{O}}\left(d^{3/4}H^{3/2}K^{3/4}\right)$, where H denotes the horizon. Our bound improves by a factor $H^{1/2}$ the first result in this setting proven in [55]. Concurrently to our work [42] derived a further improvement with optimal dependence on K .

Finally, we provide a stronger result for the finite horizon setting. Key for this result is realizing that in the regret decomposition of [41] one of the two players can in fact play the best response rather than a more conservative no regret strategy. This observations leads to Algorithm 4 which only requires $\mathcal{O}(H^4d^3\epsilon^{-2})$ MDP interactions.

Related Works Early works in behavioural cloning (BC) [33] popularized the framework showing its success in driving problem and [39, 38] show that the problem can be analyzed via a reduction to supervised learning which provides an expert trajectories bound of order $\frac{H^4\log|\Pi|}{\epsilon^2}$. Successive algorithms like Dagger [38] and Logger [25] requires to query the expert interactively. In this case, the expert trajectories improve to $\frac{H^2\max_{s,a}(A^\star(s,a))^2\log|\Pi|}{\epsilon^2}$ where $A^\star$ is the optimal advantage. Recent works [36] showed that in the worst case Dagger does not improve over BC but also that both can use only $\widetilde{\mathcal{O}}\left(\frac{H^2|\mathcal{S}|}{\epsilon}\right)$ in the tabular case. Moreover, when transitions and initial distribution are known and the expert is deterministic, the result can be improved to $\mathcal{O}\left(\frac{H^{3/2}|\mathcal{S}|}{\epsilon}\right)$ using Mimic-MD [36]. Later, [51] introduced MB-TAIL that having trajectory access to the MDP attains the same bound. This shows that the traditional bound obtained matching occupancy measure [46] adopted in [41] is suboptimal in the tabular setting. For the linear function approximation, the works in [44, 35] introduced algorithms that uses $\mathcal{O}\left(\frac{H^{3/2}d}{\epsilon}\right)$ expert trajectories with knowledge of the transitions but those require strong assumptions such as linear expert [35, Definition 4], particular choice of features, linear reward and uniform expert occupancy measure. [35] also proves an improved result for BC but under the linear expert assumption which implies that the expert is deterministic. While one can notice that there exists an optimal policy in a Linear MDP which is a linear expert, in our work we do not impose assumption on the expert policy and we require $\mathcal{O}\left(\frac{H^2d}{\epsilon^2}\right)$ demonstrations. Under the same setting, the best known bound for BC is $\frac{H^2\log|\Pi|}{d}$ times larger which makes our algorithm

Table 1: **Comparison with related algorithms** We report the number of expert trajectories and MDP interactions needed for the various algorithms to be ϵ-suboptimal according to Definition 1. Our algorithms provide guarantees for the number of expert trajectories independent on $\mathcal{S}$ and Π without assumptions on the expert policy. By **Linear Expert**, me mean that the expert policy is $\pi(s) = \max_{a\in\mathcal{A}} \phi(s,a)^\intercal\theta$ for some unknown vector θ.

Algorithm	Setting	Expert Traj.	MDP Traj.
Behavioural Cloning	Function Approximation, Offline [4]	$\mathcal{O}\left(\frac{H^4 \log\lvert\Pi\rvert}{\epsilon^2}\right)$	-
	Tabular, Offline [36]	$\widetilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\epsilon}\right)$	-
	Linear Expert, Offline [35]	$\widetilde{\mathcal{O}}\left(\frac{H^2 d}{\epsilon}\right)$	-
Mimic-MD [36]	Tabular, Known Transitions, Deterministic Expert	$\mathcal{O}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\epsilon}\right)$	-
OAL [41]	Tabular	$\mathcal{O}\left(\frac{H^2\lvert\mathcal{S}\rvert}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{H^4\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\epsilon^2}\right)$
MB-TAIL [51]	Tabular, Deterministic Expert	$\mathcal{O}\left(\frac{H^{3/2}\lvert\mathcal{S}\rvert}{\epsilon}\right)$	$\mathcal{O}\left(\frac{H^3\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\epsilon^2}\right)$
OGAIL [26]	Linear Mixture MDP	$\mathcal{O}\left(\frac{H^3 d^2}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{H^4 d^3}{\epsilon^2}\right)$
PPIL [48]	Linear MDP, Persistent Excitation	$\mathcal{O}\left(\frac{d}{(1-\gamma)^2\epsilon^2}\right)$	$\mathcal{O}\left(\frac{d^2}{\beta^6(1-\gamma)^9\epsilon^5}\right)$
ILARL (Algorithm 3)	Linear MDP	$\mathcal{O}\left(\frac{d}{(1-\gamma)^2\epsilon^2}\right)$	$\mathcal{O}\left(\frac{d^3}{(1-\gamma)^8\epsilon^4}\right)$
BRIG (Algorithm 4)	Episodic Linear MDP	$\mathcal{O}\left(\frac{dH^2}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{d^3H^4}{\epsilon^2}\right)$

preferrable whenever $|\Pi| \geq \exp(dH^{-2})$. We report a comparison with existing IL theory work in Table 1.

2 Background and Notation

In imitation learning [32], the environment is abstracted as Markov Decision Process (MDP) [34] which consists of a tuple $(\mathcal{S}, \mathcal{A}, P, c, \boldsymbol{\nu}_0)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S}\times\mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition kernel, that is, $P(s'|s,a)$ denotes the probability of landing in state s' after choosing action a in state s. Moreover, $\boldsymbol{\nu}_0$ is a distribution over states from which the initial state is sampled. Finally, $c : \mathcal{S}\times\mathcal{A} \to [-1, 1]$ is the cost function. In the infinite horizon setting, we endow the MDP tuple with an additional element called the discount factor $\gamma \in [0, 1)$. Alternatively, in the finite horizon setting we append to the MDP tuple the horizon $H \in \mathbb{N}$ and we consider possibly inhomogenous transitions or costs function. That is, they depend on the stage within the episode. The agent plays action in the environment sampled from a policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$. The learner is allowed to adopt an algorithm to update the policy across episodes given the previously observed history. We will see that imitation learning has a strong connection with MDPs with adversarial costs. The latter setting allows the cost function to change each time the learner samples a new episode in the MDP. For clarity, we include the pseudocode for the interaction in Protocol 1 in Appendix C.

Value functions and occupancy measures We define the state value function at state $s \in \mathcal{S}$ for the policy π under the cost function c as $V^\pi(s;c) \triangleq \mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h c(s_h,a_h)|s_1 = s\right]$. In the finite horizon case, the state value function also depends on the stage index h, that is $V_h^\pi(s;c) \triangleq \mathbb{E}\left[\sum_{\ell=h}^{H} c(s_\ell,a_\ell)|s_h = s\right]$[2]. In both cases, the expectation over both the randomness of the transition dynamics and the one of the learner's policy. Another convenient quantity is the occupancy measure of a policy π denoted as $d^\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and defined as follows $d^\pi(s,a) \triangleq (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathbb{P}\left[s, a \text{ is visited after } h \text{ steps acting with } \pi\right]$. We can also define the state occupancy measure as $d^\pi(s) \triangleq (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathbb{P}\left[s \text{ is visited after } h \text{ steps acting with } \pi\right]$. In the finite horizon setting, the occupancy measure depends on the stage h and its defined simply as $d_h^\pi(s,a) \triangleq \mathbb{P}\left[s, a \text{ is visited after } h \text{ steps acting with } \pi\right]$. The state occupancy measure is defined analogously.

Imitation Learning In imitation learning, the learner is given a dataset $\mathcal{D}_{\pi_{\mathrm{E}}} \triangleq \left\{\boldsymbol{\tau}^k\right\}_{k=1}^{\tau_E}$ containing τ_E trajectories collected in the MDP by an expert policy π_{E} according to Protocol 1. By trajectory $\boldsymbol{\tau}^k$, we mean the sequence of states and actions sampled at the kth iteration of Protocol 1, that is

[2] In the finite horizon case we may use $V^\pi(s;c)$ as a shortcut for $V_1^\pi(s;c)$

$\boldsymbol{\tau}^k = \left\{(s_h^k, a_k^h)\right\}_{h=1}^H$ for finite horizon case. For the infinite horizon case, the trajectories have random lenght sampled from the distribution $\mathrm{Geometric}(1-\gamma)$. Given $\mathcal{D}_{\pi_{\mathrm{E}}}$, the learner adopts an algorithm $\mathcal{A}$ to learn a policy π^{out} such that is ϵ-suboptimal according to the next definition.

Definition 1. *An algorithm $\mathcal{A}$ is said **ϵ-suboptimal** if it outputs a policy π whose value function with respect to the unknown true cost $\mathbf{c}_{\text{true}}$ satisfies $\mathbb{E}_{\mathcal{A}}\mathbb{E}_{s_1\sim\boldsymbol{\nu}_0}\left[V^{\pi}(s_1;\mathbf{c}_{\text{true}}) - V^{\pi_{\mathrm{E}}}(s_1;\mathbf{c}_{\text{true}})\right] \leq \epsilon$ where the first expectation is on the randomness of the algorithm $\mathcal{A}$.*

2.1 Setting

We study imitation learning in the linear MDP setting popularized by [22] and studied in imitation learning in [48]. When studying finite horizon problems we consider possible inhomogeneous transition dynamics and cost function. That is, we work under the following assumptions.

Assumption 1. ***Episodic Linear MDP*** *There exist a feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ known to the learner, an unknown sequence of vectors $\mathbf{w}_h^k \in \mathbb{R}^d$ and an unknown matrix sequences $M_h \in \mathbb{R}^{d\times|\mathcal{S}|}$ such that the transition matrices P_h factorize as $P_h = \boldsymbol{\Phi}M_h$ and the sequence of adversarial costs c_h^k can be written as $c_h^k = \boldsymbol{\Phi}\mathbf{w}_h^k$. Moreover, it holds for all $k \in [K], h \in [H]$ and for all state action pairs $s, a \in \mathcal{S}\times\mathcal{A}$ that $\|\boldsymbol{\Phi}\|_{1,\infty} \leq 1$, $\|M_h\|_{1,\infty} \leq 1$, $\left\|\mathbf{w}_h^k\right\|_2 \leq 1$.*

Assumption 2. ***Linear MDP*** *There exist a feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ known to the learner, an unknown sequence of vectors $\mathbf{w}^k \in \mathbb{R}^d$ and an unknown matrix $M \in \mathbb{R}^{d\times|\mathcal{S}|}$ such that the transition matrices P factorize as $P = \boldsymbol{\Phi}M$ and the sequence of adversarial costs c^k can be written as $c^k = \boldsymbol{\Phi}\mathbf{w}^k$. Moreover, it holds for all $k \in [K]$ and for all state action pairs $s, a \in \mathcal{S}\times\mathcal{A}$ that $\|\boldsymbol{\Phi}\|_{1,\infty} \leq 1$, $\|M\|_{1,\infty} \leq 1$, $\left\|\mathbf{w}^k\right\|_2 \leq 1$.*

In the context of imitation learning, we also need to assume that the true unknown cost is realizable.

Assumption 3. ***Realizable cost*** *The learner has access to a feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ such that $\mathbf{c}_{\text{true}} = \boldsymbol{\Phi}\mathbf{w}_{\text{true}}$.*

3 Main Results and techniques

We provide our main results for the infinite horizon case in Theorem 1 and the stronger result for the finite horizon in Theorem 2.

Theorem 1. *Under Assumptions 2,3, there exists an algorithm, i.e. Algorithm 3, such that after using $\widetilde{\mathcal{O}}\left(\frac{\log|\mathcal{A}|d^3}{(1-\gamma)^8\epsilon^4}\right)$ state action pairs from the MDP and using $\widetilde{\mathcal{O}}\left(\frac{2d\log(2d)}{(1-\gamma)^2\epsilon_E^2}\right)$ expert demonstrations is $\epsilon + \epsilon_E$-suboptimal.*

Theorem 2. *Under Assumptions 1,3,there exists an algorithm, i.e. Algorithm 4, such that after sampling $\mathcal{O}\left(H^4d^3\log(dH/(\epsilon))\epsilon^{-2}\right)$ trajectories and having access to a dataset of $\tau_E = \widetilde{\mathcal{O}}\left(\frac{2H^2d\log(2d)}{\epsilon_E^2}\right)$ expert demonstrations is $\epsilon + \epsilon_E$-suboptimal.*

Remark 1. *The results are proven via the high probability bounds in Theorems 5 and 6 respectively and apply the high probability to expectation conversion lemma in Lemma 7.*

3.1 Technique overview

Online-to-batch conversion The core idea is to extract the policy achieving the sample complexity guarantees above via an online-to-batch conservation. That is the output policy is sampled uniformly from a collection of K policies $\left\{\pi^k\right\}_{k=1}^K$. The sample complexity result is proven, showing that the policies $\left\{\pi^k\right\}_{k=1}^K$ produced by the algorithms under study have sublinear pseudo regret in high probability, that is,

$$\mathrm{Regret}(K) \triangleq \frac{1}{1-\gamma}\sum_{k=1}^K\left\langle \mathbf{c}_{\text{true}}, d^{\pi^k} - d^{\pi_{\mathrm{E}}}\right\rangle \leq \mathcal{O}(K^{3/4}) \quad \text{w.h.p.}$$

for the infinite horizon discounted setting with Algorithm 3 and

$$\mathrm{Regret}(K) \triangleq \sum_{h=1}^H\sum_{k=1}^K\left\langle \mathbf{c}_{\text{true},h}, d_h^{\pi^k} - d_h^{\pi_{\mathrm{E}}}\right\rangle \leq \mathcal{O}(\sqrt{K}) \quad \text{w.h.p.} \tag{1}$$

for the finite horizon setting with Algorithm 4. The next section presents the regret decomposition giving the crucial insights for the design of Algorithms 3 and 4.

Regret decomposition To obtain both regret bounds, we decompose the pseudo regret in 3 terms. We present it for the infinite horizon case, where $(1-\gamma)\mathrm{Regret}(K)$ can be upper bounded by

$$\underbrace{\sum_{k=1}^{K}\left\langle \mathbf{c}^k, d^{\pi^k} - d^{\pi_{\mathrm{E}}}\right\rangle}_{\mathrm{Regret}_{\pi}(K;d^{\pi_{\mathrm{E}}})} + \underbrace{\sum_{k=1}^{K}\left\langle \mathbf{w}_{\mathrm{true}} - \mathbf{w}^k, \mathbf{\Phi}^{\intercal} d^{\pi^k} - \mathbf{\Phi}^{\intercal} d^{\pi_{\mathrm{E}}}\right\rangle}_{\mathrm{Regret}_{\mathbf{w}}(K;\mathbf{w}_{\mathrm{true}})} \tag{2}$$

This decomposition is inspired from [41] but it applies also to the infinite horizon setting and exploits the linear structure using Assumptions 2,3 to write $c^k = \mathbf{\Phi}\mathbf{w}^k$ and $\mathbf{c}_{\text{true}} = \mathbf{\Phi}\mathbf{w}_{\mathrm{true}}$.

$\mathrm{Regret}_{\mathbf{w}}(K;\mathbf{w}_{\mathrm{true}})$ is the pseudo regret of a player updating a sequence of cost functions and having $\mathbf{c}_{\text{true}}$ as comparator while $\mathrm{Regret}_{\pi}(K;d^{\pi_{\mathrm{E}}})$ is the pseudo regret in a Linear MDP with adversarial costs $\left\{c^k\right\}_{k=1}^{K}$ and having the expert occupancy measure as a comparator.

Imitation Learning via no-regret algorithms. The decomposition in Equation (2) suggests that imitation learning algorithm can be designed chaining one algorithm that updates the sequence $\mathbf{w}^k$ to make sure that $\mathrm{Regret}_{\mathbf{w}}(K;\mathbf{w}_{\mathrm{true}})$ grows sublinearly and a second one that updates the policy sequence to control $\mathrm{Regret}_{\pi}(K;d^{\pi_{\mathrm{E}}})$. Controlling $\mathrm{Regret}_{\mathbf{w}}(K;\mathbf{w}_{\mathrm{true}})$ can be easily done via projected online gradient descent [57].

Unfortunately, controlling $\mathrm{Regret}_{\pi}(K;d^{\pi_{\mathrm{E}}})$ is way more challenging because we have no knowledge of the transition dynamics. Therefore, we can not project on the feasible set of occupancy measures. To circumvent this issue we rely on the recent literature [27, 43, 12] that however focuses on bandit feedback. In our case, the π player has full information on the cost vector c^k. Thus, we design a simpler algorithm Algorithm 1 which achieves a better regret bound in the easier full information case. Algorithm 1 improves over the regret bound in [55] and easily extends to the infinite horizon setting (see Algorithm 2).

Improved algorithm for finite horizon The techniques explained so far do not allow to get the better bound of order $\mathcal{O}(\sqrt{K})$ in the finite horizon setting (see Equation (1)). The idea is to let the $\mathbf{w}$ player update first, then the π player can update their policy knowing in advance the loss that they will suffer. This allows to use LSVI-UCB [22] for the π-player which has been originally designed for a fixed cost but we show that it still guarantees $\mathcal{O}(\sqrt{T})$ regret against an arbitrary sequence of costs when the learner knows in advance the cost function at the next episode. On the other hand, LSVI-UCB suffers linear regret if the adversarial loss is not known in advance so letting the $\mathbf{w}$ player update first is crucial. This result is provided in Appendix A.

4 Warm up: Online Learning in Adversarial Linear MDP

We start by presenting our result in full information episodic linear MDP with adversarial costs that improves over [55] by a factor $H^{1/2}$. The algorithm is quite simple. We apply a policy iteration like method with two important twist: (i) in the policy improvement step, we update the policy with a no regret algorithm rather than a greedy step. Moreover, the policy is updated only every τ episodes using as loss vector the average Q value over the last batch of collected episodes, (ii) in the policy evaluation step, we compute an optimistic estimate of the Q function for the current policy using only on-policy data.

The last part is crucial because the use of off-policy data makes the covering argument for Linear MDP problematic. Indeed, one would need to cover the space of stochastic policy when computing the covering number of the value function class but this leads to the undesirable dependence on the number of states and actions for the log covering number (see for example [2]). An alternative bound on the covering number shown in [55] would instead lead to linear regret.

Instead, using data collected on-policy allows to apply the covering argument in [43] avoiding the dependence on the number of states and actions. The first twist is at this point necessary to make the policy updates more rare giving the possibility to collect more on-policy episodes with a fixed policy. The algorithm pseudocode is in Algorithm 1.

Algorithm 1 On-policy MDP-E with unknown transitions and adversarial costs.

Input: Dataset size τ, Exploration parameter β, Step size η, initialize π_0 as uniform distribution over $\mathcal{A}$
for $j = 1, \ldots \lfloor K/\tau \rfloor$ **do**
  Denote the indices interval $T_j \triangleq [(j-1)\lfloor K/\tau \rfloor, j\lfloor K/\tau \rfloor)$.
  // Collect on-policy data
  Collect τ trajectories with policy $\pi^{(j)}$ and store them in the dataset $\mathcal{D}_h^{(j)} = \left\{(s_h^i, a_h^i, \mathbf{c}_h^i, s_{h+1}^i)\right\}_{i \in T_j}$.
  Denote global dataset $\mathcal{D}^{(j)} = \cup_{h=1}^H \mathcal{D}_h^{(j)}$.
  for $k \in T_j$ **do**
    // Optimistic policy evaluation
    Initialize $V_{H+1}^k = 0$
    for $h = H, \ldots, 1$ **do**
      $\Lambda_h^k = \sum_{(s,a)\in\mathcal{D}_h^{(j)}} \phi(s,a)\phi(s,a)^\intercal + I$ // $\phi(s,a)$ is the $(s,a)^{\text{th}}$ row of the matrix $\mathbf{\Phi}$.
      $\mathbf{v}_h^k = (\Lambda_h^k)^{-1} \sum_{(s,a,s')\in\mathcal{D}_h^{(j)}} \phi(s,a) V_{h+1}^k(s')$
      $b_h^k(s,a) = \beta \left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}}$
      $Q_h^k = \left[\mathbf{c}_h^k + \mathbf{\Phi}\mathbf{v}_h^k - b_h^k\right]_{-H+h-1}^{H-h+1}$
      $V_h^k(s) = \left\langle \pi_h^k(s), Q_h^k(s,\cdot)\right\rangle$ (with $\pi^k = \pi^{(j)}$).
    end for
  end for
  // Policy Improvement Step
  Compute average Q value $\bar{Q}_h^{(j)}(s,a) = \frac{1}{\tau}\sum_{k\in T_j} Q_h^k(s,a)$.
  Update policy $\pi_h^{(j+1)}(a|s) \propto \exp\left(-\eta \sum_{i=1}^j \bar{Q}_h^{(i)}(s,a)\right)$
end for

4.1 Analysis

Theorem 3. *Under Assumption 1, run Algorithm 1 with exploration parameter $\beta = \tilde{\mathcal{O}}(dH)$, dataset size $\tau = \frac{5\beta}{2}\sqrt{\frac{Kd}{\log|\mathcal{A}|}}$ and step size $\eta = \sqrt{\frac{\tau \log|A|}{KH^2}}$. Then, it holds with probability $1-\delta$, that*

$$\text{Regret}(K;\pi^\star) = \sum_{k=1}^K V_1^{\pi^k,k}(s_1) - V_1^{\pi^\star,k}(s_1) \leq \widetilde{\mathcal{O}}\left(d^{3/4}H^{3/2}\log^{1/4}|\mathcal{A}|\, K^{3/4}\log\frac{K}{\delta}\right) \quad (3)$$

where we use the compact notation $V_h^{\pi,k}(\cdot) \triangleq V_h^\pi(\cdot; c^k)$.

Proof. Sketch Adding and subtracting the term $\sum_{k=1}^K V_1^k(s_1)$ in the definition of regret, we have that defining $\delta_h^k(s,a) \triangleq \mathbf{c}_h^k(s,a) + P_h V_{h+1}^k(s,a) - Q_h^k(s,a)$

$$\begin{aligned}
\text{Regret}(K;\pi^\star) &= \sum_{k=1}^K V_1^{\pi^k,k}(s_1) - V_1^k(s_1) + V_1^k(s_1) - V_1^{\pi^\star,k}(s_1) \\
&\leq \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\left\langle Q_h^k(s,\cdot), \pi_h^k(s) - \pi_h^\star(s)\right\rangle\right] - \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim d_h^{\pi^\star}}\left[\delta_h^k(s,a)\right] \\
&\quad + \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[\delta_h^k(s,a)\right]
\end{aligned}$$

where the last inequality holds by the extended performance difference lemma [8, 40]. At this point we can invoke Lemma 3 (see Appendix G) to obtain

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - \mathbf{c}_h^k(s,a) - P_h V_{h+1}^k(s,a) \leq 0 \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}, h\in[H], k\in[K] \quad (4)$$

Algorithm 2 Infinite Horizon Linear MDP with adversarial losses.

1: **Input:** Dataset size τ, Exploration parameter β, Step size η, Initial policy π_0 (uniform over $\mathcal{A}$), initialize $V^1 = 0$.
2: **for** $j = 1, \dots \lfloor K/\tau \rfloor$ **do**
3: // Collect on-policy data
4: Denote the indices interval $T_j \triangleq [(j-1)\lfloor K/\tau \rfloor, j \lfloor K/\tau \rfloor)$.
5: Sample $\mathcal{D}^{(j)} = \left\{(s^i, a^i, s'^{,i}, \mathbf{c}^i)\right\}_{i \in T_j} \sim d^{\pi^{(j)}}$ using [6, Algorithm 1].
6: Compute $\Lambda^{(j)} = \sum_{(s,a)\in\mathcal{D}^{(j)}} \phi(s,a)\phi(s,a)^\intercal + I$.
7: Compute $b^{(j)}(s,a) = \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}}$.
8: // Optimistic Policy Evaluation
9: **for** $k \in T_j$ **do**
10: $\mathbf{v}^k = (\Lambda^{(j)})^{-1} \sum_{(s,a,s')\in\mathcal{D}^{(j)}} \phi(s,a) V^k(s')$
11: $Q^{k+1} = \left[\mathbf{c}^k + \gamma \mathbf{\Phi} \mathbf{v}^k - b^{(j)}\right]_{-(1-\gamma)^{-1}}^{(1-\gamma)^{-1}}$
12: $V^{k+1}(s) = \left\langle \pi^{(j)}(a|s), Q^{k+1}(s,a) \right\rangle$
13: **end for**
14: // Policy Improvement Step
15: Compute average Q value $\bar{Q}^{(j)}(s,a) = \frac{1}{\tau}\sum_{k\in T_j} Q^k(s,a)$.
16: Update policy: $\pi^{(j+1)}(a|s) \propto \exp\left(-\eta \sum_{i=1}^{j} \bar{Q}^{(i)}(s,a)\right)$
17: **end for**

with probability $1-\delta$. This implies that with probability $1-\delta$

$$\begin{aligned}\mathrm{Regret}(K;\pi^\star) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\left\langle Q_h^k(s,\cdot), \pi_h^k(s) - \pi_h^\star(s)\right\rangle\right] + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[b_h^k(s,a)\right] \\ &\leq \frac{\tau \log|\mathcal{A}|}{\eta} + \tau H + \eta K H^2 + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[b_h^k(s,a)\right]\end{aligned}$$

Notice that the last inequality follows from the mirror descent with blocking result [43, Lemma F.5]. Then, by [43, Lemma C.5], it holds that with probability $1-2\delta$.

$$\mathrm{Regret}(K;\pi^\star) \leq \frac{\tau \log|\mathcal{A}|}{\eta} + \tau H + \eta K H^2 + \frac{10 K H \sqrt{d}\beta \log(2\tau/\delta)}{\sqrt{\tau}}$$

The proof is concluded plugging in the values specified in the theorem statement. □

4.2 Extension to the Infinite Horizon Setting.

We show our proposed extension to the infinite horizon in Algorithm 2. The main difference in the analysis is to the handle the fact that in the infinite horizon setting we can not run a backward recursion to compute the optimistic value functions as done in Steps 10-16 of Algorithm 1. Instead, we use the optimistic estimate at the previous iterate Q^k to build an approximate optimistic estimate at the next iterate (see Steps 10-12 in Algorithm 2). The error introduced in this way can be controlled thanks to the regularization in the policy improvement step as noticed in [28]. [3]

Theorem 4. *Under Assumption 2, consider K iterations of Algorithm 2 run with $\tau \leq \frac{K}{\sqrt{\tau}}$ and $\beta = \widetilde{\mathcal{O}}(dH)$, then it holds for any comparator policy $\pi^\star$ that $(1-\gamma)\mathrm{Regret}(K;\pi^\star) \triangleq$*

[3] Regularization in both the evaluation and improvement step has been proven successful in the infinite horizon linear mixture MDP setting [28]. Their regularization in the evaluation step is helpful to improve the horizon dependence. In our case, we use unregularized evaluation because in the linear MDP setting our analysis presents additional leading terms that can not be bounded as $\mathcal{O}(\sqrt{K})$ with regularization in the policy evaluation routine.

$\sum_{k=1}^{K} \left\langle d^{\pi^k} - d^{\pi^\star}, \mathbf{c}^k \right\rangle$ *is upper bounded with probability* $1-2\delta$ *by*

$$\frac{\tau \log|\mathcal{A}|}{\eta} + \frac{\tau+1}{1-\gamma} + \frac{\eta K}{(1-\gamma)^2} + 12\beta K\sqrt{\frac{d}{\tau}}\log\left(\frac{2Kd}{\tau\delta}\right) + \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau}.$$

Proof. Sketch The proof is based on the following decomposition that holds in virtue of Lemma 2. Denoting $\delta^k(s,a) \triangleq \mathbf{c}^k(s,a) + \gamma P V^k(s,a) - Q^{k+1}(s,a)$ and $g^k(s,a) \triangleq Q^{k+1}(s,a) - Q^k(s,a)$

$$(1-\gamma)\text{Regret}(K;\pi^\star) = \sum_{k=1}^{K} \mathbb{E}_{s\sim d^{\pi^\star}}\left[\left\langle Q^k(s,\cdot), \pi^k(s) - \pi^\star(s)\right\rangle\right] \tag{OMD}$$

$$+ \sum_{k=1}^{K}\sum_{s,a}\left[d^{\pi^k}(s,a) - d^{\pi^\star}(s,a)\right]\cdot\left[\delta^k(s,a)\right] \tag{Optimism}$$

$$+ \sum_{k=1}^{K}\mathbb{E}_{s,a\sim d^{\pi^k}}\left[g^k(s,a)\right] - \sum_{k=1}^{K}\mathbb{E}_{s,a\sim d^{\pi^\star}}\left[g^k(s,a)\right] \tag{Shift}$$

Then, we have that Optimism can be bounded similarly to the finite horizon case using Lemma 4 while for Shift we rely on the regularization of the policy improvement step and on the fact that the policy is updated only every τ steps. All in all, we have that the first term in (Shift) can be bounded as $\frac{1}{(1-\gamma)^2}\sum_{j=2}^{\lfloor K/\tau\rfloor}\sqrt{2\eta} = \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau}$. The second term just telescopes therefore (Shift) $\leq \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau} + \frac{1}{1-\gamma}$. Finally, (OMD) can be bounded as in the finite horizon case. □

5 Imitation Learning in Infinite Horizon MDPs

In this section, we apply Theorem 4 to imitation learning. Indeed, we design Algorithm 3 using the insights from the decomposition in Equation (2): we use a no regret algorithm to update the cost at each round and we update the learner's policy using a no regret algorithm for infinite horizon full information adversarial Linear MDP, of which Algorithm 2 is the first example in the literature. The

Algorithm 3 Imitation Learning via Adversarial Reinforcement Learning (**ILARL**) for Infinite Horizon Linear MDPs.

1: **Input:** Access to Algorithm 2 (with inputs τ,β, η, π_0) , Step size for Cost Update α, Expert dataset $\mathcal{D}_{\pi_{\text{E}}} = \{\boldsymbol{\tau}_i\}_{i=1}^{\tau_E}$.
2: Estimate for expert feature visitation $\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\text{E}}}} \triangleq \frac{(1-\gamma)}{\tau_E}\sum_{\boldsymbol{\tau}\in\mathcal{D}_{\pi_{\text{E}}}}\sum_{s_h,a_h\in\boldsymbol{\tau}}\gamma^h\phi(s_h,a_h)$.
3: **for** $k = 1,\dots,K$ **do**
4: // Cost Update (to control $\text{Regret}_w(K;\mathbf{w}_{\text{true}})$)
5: Estimate $\widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}} \triangleq \tau^{-1}\sum_{s,a\in\mathcal{D}^{(j)}}\phi(s,a)$ where $\mathcal{D}^{(j)}$ is defined in Step 6 in Algorithm 2.
6: $\mathbf{w}^{k+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}^k - \alpha(\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\text{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}})\right]$ with $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq 1\}$.
7: // Policy Update (to control $\text{Regret}_\pi(K;\pi_{\text{E}})$)
8: The cost $\Phi\mathbf{w}^k$ is revealed to the learner.
9: The learner updates their policy π^k performing one iteration of Algorithm 2.
10: **end for**

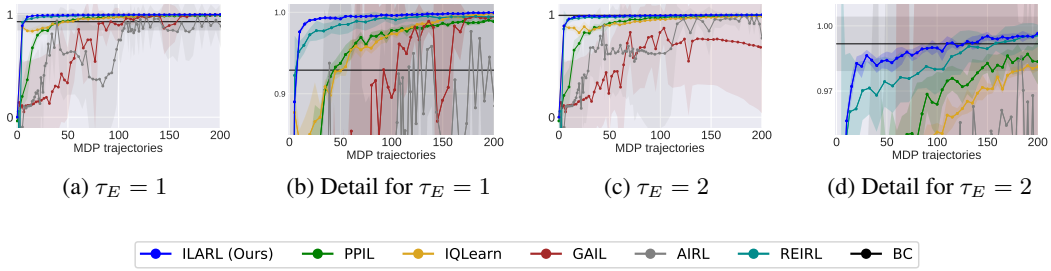
guarantees for Algorithm 3 are given in the following theorem.

Theorem 5. *Under Assumptions 2,3, let us consider K iterations of Algorithm 3 with $K \geq \widetilde{\mathcal{O}}\left(\frac{\log|\mathcal{A}|d\beta^2\log^2(1/\delta)}{(1-\gamma)^6\epsilon^4}\right)$ where β is chosen as in Lemma 6 (i.e. $\beta = \widetilde{\mathcal{O}}\left(d(1-\gamma)\right)^{-1}$). Moreover, let us consider the following choices $\alpha = \frac{1}{\sqrt{2K}}$, $\tau = \mathcal{O}\left(\frac{\beta(1-\gamma)\sqrt{dK}\log(2dK/\delta)}{\sqrt{\log|\mathcal{A}|}}\right)$, expert trajectories $\tau_E = \frac{8d\log(d/\delta)}{(1-\gamma)^2\epsilon_E^2}$ and $\eta = \sqrt{\frac{\tau\log|\mathcal{A}|(1-\gamma)^2}{K}}$.Then, the above conditions ensure $\frac{1}{1-\gamma}\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k}\right\rangle \leq \epsilon + \epsilon_E$ with probability $1-4\delta$.*

The proof included in Appendix G starts with the decomposition in Equation (2). Then, we control the term $\mathrm{Regret}_w(K; \mathbf{w}_{\mathrm{true}})$ with the standard online gradient descent analyses and the term $\mathrm{Regret}_\pi(K; \pi_{\mathrm{E}})$ with Theorem 4. Finally, we control the statistical estimation error for the losses seen by the $\mathbf{w}$-player with an application of Lemma 8.

Remark 2. *The resulting algorithm improves over [48] in two ways: (i) We bypass all kind of exploration assumptions, such as the persistent excitation assumption. We remark that this is a qualitative improvement. Indeed, the persistent excitation assumption is easily violated by deterministic policies with tabular features. (ii) Moreover, the sample complexity improves from $\mathcal{O}(\epsilon^{-5})$ to $\mathcal{O}(\epsilon^{-4})$.*

6 Empirical evaluation



(a) $\tau_E = 1$ (b) Detail for $\tau_E = 1$ (c) $\tau_E = 2$ (d) Detail for $\tau_E = 2$

Figure 1: Experiments on a continuous gridworld with a stochastic expert.The y-axis reports the normalized return. 1 correpsonds to the expert performance and 0 to the uniform policy one.

We numerically verify the main theoretical insights derived in the previous sections (i) We aim to verify that for a general stochastic expert, the efficiency in terms of expert trajectories improves upon behavioural cloning. (ii) ILARL is more efficient in terms of MDP trajectories compared to PPIL [48] which has worst theoretical guarantees and with popular algorithms that are widely used in practice but do not enjoy theoretical guarantees: GAIL [20], AIRL [15], REIRL [7] and IQLearn [16] The experiments are run in a continuous state MDP explained in Appendix I.

Expert trajectory efficiency with stochastic expert For the first claim, we use a stochastic expert obtained following with equal probability either the action taken by a deterministic experts previously trained with LSVI-UCB or an action sampled uniformly at random. We collect with such policy τ_E trajectories. From Figure 1, we observe that all imitation learning we tried have a final performance improving over behavioural cloning for the case $\tau_E = 1$ while only REIRL and ILARL do so for $\tau_E = 2$. In both cases, ILARL achieves the highest return that even matches the expert performance.

MDP trajectories efficiency For the second claim, we can see in Figure 1 that ILARL is the most efficient algorithm in terms of MDP trajectories for both values of τ_E.

7 Conclusions

In this paper, we proposes ILARL which greatly reduces the number of MDP trajectories in imitation learning in Linear MDP and BRIG that provides a further improvement for the finite horizon case. Both results build on the connection between imitation learning and MDPs with adversarial losses.

There is a number of exciting future directions. In particular, the estimation of $\widehat{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}}$ could be carried out with **fewer expert trajectories** using trajectory access to the MDP. This observation has been proven successful having access to the exact transitions of the MDP in the tabular case [36] or under linear function approximation with further assumption on the expert policy and the feature distribution [44, 35]. Whether the same is possible for general stochastic experts in Linear MDP is an interesting open question.

Finally, a **better sample complexity** can be achieved designing better no regret algorithm for infinite horizon adversarial discounted linear MDP with full-information feedback and apply them in Step 9 of Algorithm 3 building for example on the recent result for the finite horizon case in [42].

References

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning (ICML)*, 2019.

[2] Yasin Abbasi-Yadkori, Peter L. Bartlett, and Csaba Szepesvari. Online learning in markov decision processes with adversarially chosen transition probability distributions, 2013.

[3] Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced politex. *arXiv:1908.10479*, 2019.

[4] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

[5] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems (NeurIPS)*, 2020.

[6] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

[7] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 182–189. JMLR Workshop and Conference Proceedings, 2011.

[8] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning (ICML)*, 2020.

[9] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[10] Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *arXiv:20031004*, 2020.

[11] Robert Dadashi, Leonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[12] Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.

[13] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[14] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning (ICML)*, 2020.

[15] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2021.

[17] Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[18] Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

[19] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[20] J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning (ICML)*, 2016.

[21] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[22] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation, 2019.

[23] W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving, 2021.

[24] Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, and Dale Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward MDPs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] Yichen Li and Chicheng Zhang. On efficient online imitation learning via classification. *Advances in Neural Information Processing Systems*, 35:32383–32397, 2022.

[26] 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 (ICML)*, 2022.

[27] Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

[28] Antoine Moulin and Gergely Neu. Optimistic planning by regularized dynamic programming. *arXiv preprint arXiv:2302.14004*, 2023.

[29] Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567, 2003.

[30] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pages 529–551. PMLR, 2021.

[31] Francesco Orabona. A modern introduction to online learning, 2023.

[32] T Osa, J Pajarinen, G Neumann, JA Bagnell, P Abbeel, and J Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

[33] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

[34] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[35] Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. *Advances in Neural Information Processing Systems*, 34:1325–1336, 2021.

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

[37] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. SQIL: imitation learning via regularized behavioral cloning. *arXiv:1905.11108*, 2019.

[38] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[39] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[40] Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

[41] Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. *arXiv:2102.06924*, 2021.

[42] Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

[43] Uri Sherman, Tomer Koren, and Yishay Mansour. Improved regret for efficient online reinforcement learning with linear function approximation, 2023.

[44] 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. *Advances in Neural Information Processing Systems*, 35:7077–7088, 2022.

[45] Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pages 33299–33318. PMLR, 2023.

[46] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[47] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34:25917–25931, 2021.

[48] Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. *Advances in Neural Information Processing Systems*, 35:24309–24326, 2022.

[49] Joe Watson, Sandy H. Huang, and Nicolas Heess. Coherent soft imitation learning, 2023.

[50] Feiyang Wu, Jingyang Ke, and Anqi Wu. Inverse reinforcement learning with the average reward criterion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[51] Tian Xu, Ziniu Li, Yang Yu, and Zhi-Quan Luo. Provably efficient adversarial imitation learning with unknown transitions. *arXiv preprint arXiv:2306.06563*, 2023.

[52] Siliang Zeng, Mingyi Hong, and Alfredo Garcia. Structural estimation of markov decision processes in high-dimensional state space with finite-time guarantees. *arXiv preprint arXiv:2210.01282*, 2022.

[53] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

[54] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Understanding expertise through demonstrations: A maximum likelihood framework for offline inverse reinforcement learning. *arXiv preprint arXiv:2302.07457*, 2023.

[55] Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841*, 2023.

[56] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[57] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Tom Fawcett and Nina Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 928–936. AAAI Press, 2003.

A An improvement for the finite horizon case

We notice that in Algorithm 3 we missed an opportunity. In fact, we could use prior knowledge of the cost function $\mathbf{c}_h^k$ to update the policy π_h^k. In other words, in Step 8 of Algorithm 3 we could reveal $\mathbf{w}^{k+1}$ to the algorithm that controls $\mathrm{Regret}_\pi(K; d^{\pi_{\mathrm{E}}})$. From an online learning perspective we can play best response to control more effectively the regret term $\mathrm{Regret}_\pi(K; d^{\pi_{\mathrm{E}}})$ using LSVI-UCB [22]. This idea leads to the Algorithm 4. Theorem 6 proves that Algorithm 4 improves

Algorithm 4 Best Response Imitation learninG (**BRIG**).

Input: Exploration parameter β, Step size η , Reward step size α, expert dataset $\mathcal{D}_{\pi_{\mathrm{E}}} = \cup_{h\in[H]}\mathcal{D}_{\pi_{\mathrm{E}},h}$.
Initialize π_0 as uniform distribution over $\mathcal{A}$.
Estimate expert features expectation vectors $\widehat{\mathbf{\Phi}^\intercal d_h^{\pi_{\mathrm{E}}}} = \frac{1}{|\mathcal{D}_{\pi_{\mathrm{E}},h}|}\sum_{s,a\in\mathcal{D}_{\pi_{\mathrm{E}},h}}\phi(s,a)$ for all $h \in [H]$.
for $k = 1, \dots K$ **do**
  Collect one episodes with policy π^k denoted as $\boldsymbol{\tau}^k = \left\{(s_h^k, a_h^k, s_{h+1}^k)\right\}_{h=1}^{H-1}$ and for every $h \in [H]$ append data $(s_h^k, a_h^k, s_{h+1}^k)$ to $\mathcal{D}_h$.
  // Cost Update (to control $\mathrm{Regret}_w(K; \mathbf{w}_{\mathrm{true}})$)
  $\mathbf{w}_h^{k+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}_h^k - \alpha(\widehat{\mathbf{\Phi}^\intercal d_h^{\pi_{\mathrm{E}}}} - \phi\left(s_h^k, a_h^k\right))\right]$ with $\mathcal{W} = \{\mathbf{w} : 0 \le \mathbf{w} \le 1\}$ for all $h \in [H]$.
  // Full information LSVI-UCB (to control $\mathrm{Regret}_\pi(K; \pi_{\mathrm{E}})$)
  Initialize $V_{H+1}^k = 0$
  for $h = H, \dots, 1$ **do**
    $\Lambda_h^k = \sum_{(s,a)\in\mathcal{D}_h}\phi(s,a)\phi(s,a)^\intercal + I$
    $\mathbf{v}_h^k = (\Lambda_h^k)^{-1}\sum_{(s,a,s')\in\mathcal{D}_h}\phi(s,a)V_{h+1}^k(s')$
    $b_h^k(s,a) = \beta\left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}}$
    $Q_h^k = \left[\mathbf{\Phi}\mathbf{w}_h^{k+1} + \mathbf{\Phi}\mathbf{v}_h^k - b_h^k\right]_{-H+h-1}^{H-h+1}$ // We use the future loss $\mathbf{w}_h^{k+1}$.
    $\pi_h^{k+1}(s) = \arg\min\left(Q_h^k(s,\cdot)\right)$ // Greedy policy update, Best Response.
    $V_h^k(s) = \left\langle \pi_h^{k+1}(s), Q_h^k(s,\cdot)\right\rangle$.
  end for
end for

the required number of interaction to $K = \widetilde{\mathcal{O}}(H^4 d^3 \epsilon^{-2}\log^2(1/\delta))$ which greatly improves over $K = \widetilde{\mathcal{O}}(H^6 \log|\mathcal{A}|\, d^3\epsilon^{-4}\log^2(1/\delta))$ achieved by Algorithm 3 applied to finite horizon problems which does not use the best response observation. A core step in the proof is to show that the regret of LSVI-UCB is still $\mathcal{O}(\sqrt{K})$ if the cost function is not fixed but it is observed in advanced by the agent.

Theorem 6. *Let us consider $K = \mathcal{O}\left(\frac{H^4 d^3 \log(dH/(\epsilon\delta))}{\epsilon^2}\right)$ iterations of Algorithm 4 run with $\alpha = \sqrt{\frac{1}{2K}}$ and expert demonstrations $\tau_E = \frac{2H^2 d\log(2d/\delta)}{\epsilon_E^2}$. Moreover, let $\hat{k}$ be an iteration index sampled uniformly at random from $\{1, 2, \dots, K\}$, then it holds that with probability $1 - 3\delta$, it holds that $\mathbb{E}_{\hat{k}}\left[V_1^{\pi^{\hat{k}}}(s_1; \mathbf{c}_{\text{true}}) - V_1^{\pi_{\mathrm{E}}}(s_1; \mathbf{c}_{\text{true}})\right] \le \mathcal{O}(\epsilon_E + \epsilon)$.*

Remark 3. *Unfortunately, the best response idea does not help improving the infinite horizon result because the use of greedy policies makes the term* (Shift) *impossible to control.*

B On the β of the persistent excitation assumption

Lemma 1. *Let the persistent excitation assumption holds, then it holds that $\beta \le d^{-1}$.*

Proof.

$$
\begin{aligned}
\beta &\leq \lambda_{\min}(\mathbb{E}_{s,a\sim d^{\pi^k}}\phi(s,a)\phi(s,a)^T) \\
&\leq \frac{1}{d}\mathrm{Trace}(\mathbb{E}_{s,a\sim d^{\pi^k}}\phi(s,a)\phi(s,a)^T) \\
&= \frac{1}{d}\mathbb{E}_{s,a\sim d^{\pi^k}}\mathrm{Trace}(\phi(s,a)\phi(s,a)^T) \\
&= \frac{1}{d}\mathbb{E}_{s,a\sim d^{\pi^k}}\mathrm{Trace}(\phi(s,a)^T\phi(s,a)) \\
&= \frac{1}{d}\mathbb{E}_{s,a\sim d^{\pi^k}}\|\phi(s,a)\|_2^2 \\
&\leq \frac{1}{d}\mathbb{E}_{s,a\sim d^{\pi^k}}\|\phi(s,a)\|_1^2 \\
&\leq \frac{1}{d}\max_{s,a}\|\phi(s,a)\|_1^2 \\
&= \frac{1}{d}\|\Phi\|_{1,\infty}^2 \leq \frac{1}{d}.
\end{aligned}
$$

where the last step follows from $\|\Phi\|_{1,\infty} \leq 1$ as assumed in Assumptions 1 and 2. □

Using this result, we obtain that the dimension dependence in the bound in Viano et al. 2022 is in the best case d^8. Therefore, our new algorithm improves the dimension dependence as well.

C Interaction Protocol

Protocol 1 Interaction in Adversarial MDPs

for Episode index $k \in [1, K]$ **do**
  Sample initial state $s_1^k \sim \boldsymbol{\nu}_0$
  if Finite Horizon **then**
    for stage $h \in [1, H]$ **do**
      The learner plays an action sampled from the policy $a_h^k \sim \pi_h^k(\cdot|s_h^k)$.
      The environment sample next state $s_h^k \sim P_h(\cdot|s_h^k, a_h^k)$.
      The agent observes the vector c_h^k.
    end for
  end if
  if Infinite Horizon **then**
    Initialize $Z = 0$, $i = 1$, $s^1 \sim \boldsymbol{\nu}_0$.
    while $Z == 0$ **do**
      The learner plays an action sampled from the policy $a^i \sim \pi^k(\cdot|s^i)$.
      The environment sample next state $s'^{,i} \sim P(\cdot|s^i, a^i)$, $s^{i+1} = s'^{,i}$.
      // Restart with probability $1 - \gamma$.
      Sample $Z \sim \mathrm{Bernoulli}(1 - \gamma)$.
    end while
    The agent observes the vector c^k.
  end if
  The learner chooses her next policy, i.e. π^{k+1}.
end for

D Additional Related Works in the infinite horizon discounted setting

The practical advantage is that in the infinite horizon setting the optimal policy can be sought in the class of stationary policies which are much easier to store in memory than the nonstationary ones. All the works we mentioned so far focused on the finite horizon, however the infinite horizon setting is the most common in practice [20, 19, 15, 37, 11, 49, 16]. The practical advantage is that in the infinite horizon setting the optimal policy can be sought in the class of stationary policies

which are much easier to store in memory than the nonstationary ones. Despite this fact, there are only few previous result studying IL in infinite horizon MDP and all of them operate under limiting assumptions. As mentioned, [48] requires the persistent excitation assumption, [50] requires a uniformly good evaluation policy evaluation error which is possible only if the policies generated by the algorithm visits every state with high probability. [53] provided the first guarantees for IL with non linear reward functions but it assumes ergodic dynamics and that the the soft action value function of every policy can be perfectly evaluated at every state action pair. The latter assumption has been relaxed later in [52] and in [54] that allows for a uniformly bounded policy evaluation error. In the latter case, the policies are evaluated under the transitions learned from expert data. All in all, their bound on the number of expert trajectories scale as $(1-\gamma)^{-4}$ and with the number of states visited by the expert while our bound leverages online access to the MDP to obtain a better horizon and to avoid the dependence on the number of states. Moreover, the bounds in [54, 52, 50] depend on the number of states which can be prohibitively large in the function approximation setting.

E Future directions

Reducing the number of expert trajectories [35] showed that under known transitions for a particular choice of features and when the linear expert occupancy measure is uniform over the state space the necessary expert trajectories are $\mathcal{O}\left(\frac{H^{3/2}d^2}{\epsilon}\right)$.

In the linear MDP case, under the same assumption on the features we can show that the same amount of expert trajectories is sufficient even for the unknown transition case. Moreover, the algorithm we propose is computationally efficient while it is unclear how the output policy in linear Mimic-MD [35] can be computed with complexity independent on the number of states and actions. We detail this in Appendix J.

In addition, it is an interesting open question to see if the same improvement in terms of expert trajectory can be achieved under weaker conditions. A first step has been already made in [44]. They still requires a linear expert model but avoids the uniform expert occupancy measure assumption replacing it with the bounded density assumption (see [44, Assumption 9]). A natural follow up would be to investigate if the same expert trajectory bound can be obtained bringing the Linear MDP assumption but dropping the linear expert and the other assumptions used in [35, 44]. An intermediate step could benefit from using the persistent excitation assumption which is the natural counterpart of the bounded density assumption [44, Assumption 9] in Linear MDPs.

Results for Linear Mixture MDPs Analogous ideas can be used for the case of Linear Mixture MDPs. In particular, one can use the same structure as in Algorithm 3 but replacing an algorithm that deals with adversarial losses in Linear Mixture MDPs. For the infinite horizon case one can use [28] while for the finite horizon case one can choose [56]. In the latter case one can improve the Sample Complexity of OGAIL [26] to $\mathcal{O}\left(H^3d^2\epsilon^{-2}\right)$. Moreover, we do not think that that the Linear Expert assumption [35] is meaningful in Linear Mixture MDPs because this would require the learner to know in advance the features $\int_{\mathcal{S}}\phi(s,a,s')V^\star(s')ds'$ where $V^\star$ is the optimal state value function.

Extension to Bilinear Classes The current results can be extended to Bilinear Classes [13] at least in the finite horizon case. To this end we would need the observation that we can update the cost first and then using an algorithm which is allowed to see the next cost vector one round in advance. This is the same fact that allowed us to obtain $\mathcal{O}(\epsilon^{-2})$ sample complexity bound in the finite horizon case using LSVI-UCB.

In the following we present an informal discussion of the proof technique that would prove polynomial sample complexity for Imitation Learning in bilinear classes.

If the cost at round k is known before the learner needs to take an action the agent can form the discrepancy function

$$\ell^k(s_h,a_h,s_{h+1},g) = Q_{h,g}(s_h,a_h) + c_h^k - V_{h+1,g}(s_{h+1})$$

where the upper script k highlights the fact that the discrepancy function depends on the adversarial cost c^k. At each round, we can then compute

$$\text{argmax}_{g\in\mathcal{H}} V_{0,g}(s_1) \quad \text{s.t.} \quad \frac{1}{m}\sum_{\tau=1}^{k}\sum_{i=1}^{m} \ell^k(s_h^i, a_h^i, s_{h+1}^i, g) \leq k\epsilon_{\text{gen},k}^2(m,\delta) \quad \forall h \in [H]$$

Where the generalization error at round k denoted $\epsilon_{\text{gen},k}$ satisfies that

$$\sup_{g\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^{m} \ell^k(s_h^i, a_h^i, s_{h+1}^i, g) - \mathbb{E}_{s,a\sim d^{\pi^k}, s'\sim P(\cdot|s,a)} \left[\ell^k(s_h, a_h, s_{h+1}, g)\right] \right| \leq \epsilon_{\text{gen},k}(m,\delta)$$

with probability at least $1-\delta$.

At this point, we modify the Bilinear Classes assumption to keep into account the adversarial costs setting as follows. We assume that all the adversarial costs belongs to a convex set $\mathcal{C}$. Then we consider an MDP for which there exists a function $f^\star \in \mathcal{H}$ such that for all $c \in \mathcal{C}$, it holds that

$$\left|\mathbb{E}_{s,a\sim d_h^{\pi_f}} \left[Q_{h,f}(s_h,a_h) + c_h(s_h,a_h) - V_{h+1,f}\right]\right| \leq |\langle W_h^c(f) - W_h^c(f^\star), X_h^c(f)\rangle|$$

and

$$\left|\mathbb{E}_{s,a\sim d_h^{\pi_f}} \left[Q_{h,g}(s_h,a_h) + c_h(s_h,a_h) - V_{h+1,g}\right]\right| \leq |\langle W_h^c(g) - W_h^c(f^\star), X_h^c(f)\rangle|$$

This modified assumption for Bilinear classes with time changing rewards implies that the comparator hypothesis $f^\star$ is realizable for all k, that is, for all k it holds that

$$\frac{1}{m}\sum_{\tau=1}^{k}\sum_{i=1}^{m} \ell^k(s_h^i, a_h^i, s_{h+1}^i, f^\star) \leq k\epsilon_{\text{gen},k}^2(m,\delta)$$

so optimism holds and with the same steps in [13] it can be proven that:

$$\begin{aligned}
\text{Regret}(K,\pi^\star) &\leq \sum_{k=1}^{K} \left|\mathbb{E}_{s,a\sim d_h^{\pi_{f^k}}} \left[Q_{h,f^k}(s_h,a_h) + c_h(s_h,a_h) - V_{h+1,f^k}\right]\right| \\
&\leq \sum_{k=1}^{K} \left|\left\langle W_h^{r^k}(f^k) - W_h^{r^k}(f^\star), X_h^{r^k}(f^k)\right\rangle\right|
\end{aligned}$$

Then, using [13, Equation 8] and the elliptical potential lemma we obtain

$$\frac{1}{K}\text{Regret}(K,\pi^\star) = \mathcal{O}\left(\sqrt{K \max_{k\in[K]} \epsilon_{\text{gen},k}^2(m,\delta)\left(\exp\frac{\log K}{K} - 1\right)}\right)$$

Then, denoting $\epsilon_{\text{gen}}(m,\delta) = \max_{k\in[K]} \epsilon_{\text{gen},k}(m,\delta)$ and choosing $K = \mathcal{O}(\log \epsilon_{\text{gen}}^{-2}(m,\delta))$ gives $\frac{1}{K}\text{Regret}(K,\pi^\star) \leq \mathcal{O}\left(\epsilon_{\text{gen}}(m,\delta)\right)$.

This concludes the regret proof for the π-player. The player updating the sequences of cost can still use OGD projecting on the set $\mathcal{C}$.

Our algorithm for the finite horizon setting, BRIG, uses greedy policies. In this case, [13] showed that the generalization error can be controlled effectively in many instances such as Linear $Q^\star/V^\star$, Low Occupancy measure models [13], Bellman complete models [29] and finite Bellman Rank [21].

However in the infinite horizon case we need to use regularization for which it is currently not known if $\text{gen}(m,\delta)$ can be controlled effectively. This is again related to the issue with the covering number of softmax policies in linear MDPs (see [42, 55]). This is an interesting open question.

A final comment is that the algorithm proposed in [13] for bilinear classes is not computationally efficient is general. Therefore also its adversarial extension presented above will have this drawback. In this paper we focused on the smaller class of Linear MDP for which we provide a computationally efficient algorithm.

Improving dependence on ϵ, d and H For what concerns the finite horizon case we presented BRIG which has a dependence on ϵ which can not be improved further if not bypassing the reduction to online learning in MDPs. However the dependence in d and H can be improved. To circumvent this problem, we could think that one could apply LSVI-UCB++ [18]. However, it turns out that LSVI-UCB++ fails if the cost changes adversarially even if the learner knows the cost at the next round. Therefore, a new algorithm design is needed to improve by a factor dH upon the MDP trajectory complexity of Algorithm 4.

F Omitted Proofs

F.1 Proof of Lemma 2

Lemma 2. *Consider the MDP $M = (\mathcal{S}, \mathcal{A}, \gamma, P, \mathbf{c})$ and two policies $\pi, \pi' : \mathcal{S} \to \Delta_{\mathcal{A}}$. Then consider for any $\widehat{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\widehat{V}^{\pi}(s) = \left\langle \pi(\cdot|s), \widehat{Q}(s,\cdot) \right\rangle$ and $Q^{\pi'}, V^{\pi'}$ be respectively the state action and state value function of the policy π in MDP M and let $E \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ be the matrix such that for an arbitrary vector $f \in \mathbb{R}^{|\mathcal{S}|}$ it holds $(Ef)(s,a) = f(s)$. Then, it holds that*

$$\left\langle \boldsymbol{\nu}_0, \widehat{V}^{\pi} - V^{\pi'} \right\rangle = \frac{1}{1-\gamma} \left(\left\langle d^{\pi'}, \widehat{Q} - \mathbf{c} - \gamma P \widehat{V}^{\pi} \right\rangle + \left\langle d^{\pi'}, E\widehat{V}^{\pi} - \widehat{Q} \right\rangle \right)$$

Proof.

$$\begin{aligned}
\left\langle d^{\pi'}, \widehat{Q} \right\rangle &= \left\langle d^{\pi'}, \widehat{Q} - \mathbf{c} - \gamma P \widehat{V}^{\pi} \right\rangle + \left\langle d^{\pi'}, \mathbf{c} + \gamma P \widehat{V}^{\pi} \right\rangle \\
&= \left\langle d^{\pi'}, \widehat{Q} - \mathbf{c} - \gamma P \widehat{V}^{\pi} \right\rangle + (1-\gamma) \left\langle \boldsymbol{\nu}_0, V^{\pi'} \right\rangle + \left\langle \gamma P^{\intercal} d^{\pi'}, \widehat{V}^{\pi} \right\rangle \\
&= \left\langle d^{\pi'}, \widehat{Q} - \mathbf{c} - \gamma P \widehat{V}^{\pi} \right\rangle + (1-\gamma) \left\langle \boldsymbol{\nu}_0, V^{\pi'} - \widehat{V}^{\pi} \right\rangle + \left\langle E^{\intercal} d^{\pi'}, \widehat{V}^{\pi} \right\rangle
\end{aligned}$$

Rearranging the terms leads to the conclusion. □

F.2 Proof of Theorem 4

Proof. Let $V^{\pi^k,k} \in [-(1-\gamma)^{-1}, (1-\gamma)^{-1}]^{|\mathcal{S}|}$ denote $V^{\pi^k}(\cdot; \mathbf{c}^k)$, then

$$\begin{aligned}
\mathrm{Regret}(K; \pi^\star) &\triangleq \sum_{k=1}^{K} \left\langle d^{\pi^k} - d^{\pi^\star}, \mathbf{c}^k \right\rangle \\
&= (1-\gamma) \sum_{k=1}^{K} \left\langle \boldsymbol{\nu}_0, V^{\pi^k,k} - V^{\pi^\star,k} \right\rangle \\
&= \sum_{k=1}^{K} \mathbb{E}_{s \sim d^{\pi^\star}} \left[\left\langle Q^k(s,\cdot), \pi^k(s) - \pi^\star(s) \right\rangle \right] && \text{(OMD)} \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{s,a \sim d^{\pi^\star}} \left[Q^{k+1}(s,a) - \mathbf{c}^k(s,a) - \gamma P V^k(s,a) \right] && \text{(Optimism 1)} \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{s,a \sim d^{\pi^k}} \left[\mathbf{c}^k(s,a) + \gamma P V^k(s,a) - Q^{k+1}(s,a) \right] && \text{(Optimism 2)} \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{s,a \sim d^{\pi^k}} \left[Q^{k+1}(s,a) - Q^k(s,a) \right] && \text{(Shift 1)} \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{s,a \sim d^{\pi^\star}} \left[Q^k(s,a) - Q^{k+1}(s,a) \right] && \text{(Shift 2)}
\end{aligned}$$

Then, we have that Optimism 1,Optimism 2 can be bounded using Lemma 4 while for Shift 1 we crucially rely on the regularization of the policy improvement step and on the fact that the policy is

updated only every τ steps. Both these observations allow to derive

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}_{s,a\sim d^{\pi^k}} \left[Q^k(s,a) - Q^{k+1}(s,a)\right] &\leq \sum_{k=2}^{K} \sum_{s,a} Q^k(s,a) \left(d^{\pi^k}(s,a) - d^{\pi^{k-1}}(s,a)\right) + \sum_{s,a} Q^1(s,a) d^{\pi^1}(s,a) \\
&\leq \sum_{k=2}^{K} \left\|Q^k(s,a)\right\|_\infty \left\|d^{\pi^k} - d^{\pi^{k-1}}\right\|_1 \\
&\leq \frac{1}{1-\gamma} \sum_{k=2}^{K} \left\|d^{\pi^k} - d^{\pi^{k-1}}\right\|_1 \\
&= \frac{1}{1-\gamma} \sum_{j=2}^{\lfloor K/\tau \rfloor} \left\|d^{\pi^{(j)}} - d^{\pi^{(j-1)}}\right\|_1
\end{aligned}$$

At this point applying Pinkser's inequality and [28, Lemma A.1] we obtain

$$\begin{aligned}
\left\|d^{\pi^{(j)}} - d^{\pi^{(j-1)}}\right\|_1 \leq \sqrt{2D_{KL}(d^{\pi^{(j)}}, d^{\pi^{(j-1)}})} &\leq \sqrt{\frac{2}{1-\gamma} \mathbb{E}_{s\sim d^{\pi^{(j)}}} D_{KL}(\pi^{(j)}(\cdot|s), \pi^{(j-1)}(\cdot|s))} \\
&\leq \sqrt{\frac{2\eta}{1-\gamma} \mathbb{E}_{s,a\sim d^{\pi^{(j)}}} \bar{Q}^{(j-1)}} \\
&\leq \frac{\sqrt{2\eta}}{1-\gamma}
\end{aligned}$$

All in all, we have that

$$\text{(Shift 1)} \leq \frac{1}{(1-\gamma)^2} \sum_{j=2}^{\lfloor K/\tau \rfloor} \sqrt{2\eta} = \frac{\sqrt{2\eta} K}{(1-\gamma)^2 \tau}$$

For the second shift term, we can use a trivial telescoping argument to obtain that (Shift 2) $\leq (1-\gamma)^{-1}$.

The terms (Optimism 1) and (Optimism 2) can be bounded exactly as in the finite horizon case thanks to Lemma 4 we have that with probability $1-\delta$

$$\text{(Optimism 1)} + \text{(Optimism 2)} \leq 2 \sum_{k=1}^{K} \mathbb{E}_{s,a\sim d^{\pi^k}} \left[b^k(s,a)\right]. \tag{5}$$

Therefore, we just need to adapt the argument for bounding the exploration term $2\sum_{k=1}^{K} \mathbb{E}_{s,a\sim d^{\pi^k}} \left[b^k(s,a)\right]$ in the infinite horizon case. We start by exploiting the fact that both b^k and π^k change in fact only every τ updates. So we have

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}_{s,a\sim d^{\pi^k}} \left[b^k(s,a)\right] &= \tau \sum_{j=1}^{K/\tau} \mathbb{E}_{s,a\sim d^{\pi^{(j)}}} \left[b^{(j)}(s,a)\right] \\
&= \sum_{j=1}^{K/\tau} \mathbb{E}_{\mathcal{D}^{(j)}\sim d^{\pi^{(j)}}} \left[\sum_{s,a\in\mathcal{D}^{(j)}} b^{(j)}(s,a)\right] \\
&= \sum_{j=1}^{K/\tau} \mathbb{E}_{\mathcal{D}^{(j)}\sim d^{\pi^{(j)}}} \left[\sum_{s,a\in\mathcal{D}^{(j)}} \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}}\right]
\end{aligned}$$

At this point, fixing an arbitrary order for the state action pairs in $\mathcal{D}^{(j)}$ we can define for any $i \in [1,\tau]$ the matrix

$$\Lambda_i^{(j)} = \sum_{\ell=1}^{i} \phi(s^\ell, a^\ell)\phi(s^\ell, a^\ell)^T + \lambda I$$

Then, it holds that

$$\sum_{j=1}^{K/\tau} \mathbb{E}_{\mathcal{D}^{(j)}\sim d^{\pi^{(j)}}} \left[\sum_{s,a\in\mathcal{D}^{(j)}} \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} \right] = \sum_{j=1}^{K/\tau} \mathbb{E}_{\mathcal{D}^{(j)}\sim d^{\pi^{(j)}}} \left[\sum_{\ell=1}^{\tau} \beta \left\|\phi(s^\ell,a^\ell)\right\|_{(\Lambda^{(j)})^{-1}} \right]$$
$$\leq \sum_{j=1}^{K/\tau} \mathbb{E}_{\mathcal{D}^{(j)}\sim d^{\pi^{(j)}}} \left[\sum_{\ell=1}^{\tau} \beta \left\|\phi(s^\ell,a^\ell)\right\|_{(\Lambda_\ell^{(j)})^{-1}} \right]$$

At this point, we can notice that for any index pair ℓ, j it holds that

$$\left\|\phi(s^\ell,a^\ell)\right\|_{(\Lambda_\ell^{(j)})^{-1}} \leq \sqrt{\lambda_{\max}((\Lambda_\ell^{(j)})^{-1})} \left\|\phi(s^\ell,a^\ell)\right\| \leq \sqrt{\frac{1}{\lambda_{\min}(\Lambda_\ell^{(j)})}} \leq 1$$

where the norm $\phi(s^\ell,a^\ell)$ is upper bounded by 1 thanks to the Linear MDP assumption. Then, via [43, Lemma F.1] (where the random variable X_i is in this context $\sum_{\ell=1}^{\tau} \beta \left\|\phi(s^\ell,a^\ell)\right\|_{(\Lambda_\ell^{(j)})^{-1}}$ which is supported in $[0,\beta\tau]$) we can continue upper bounding with probability $1-\delta$ the bonus sum as

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}_{s,a\sim d^{\pi^k}} \left[b^k(s,a)\right] &\leq 2\sum_{j=1}^{K/\tau}\sum_{\ell=1}^{\tau} \beta \left\|\phi(s^\ell,a^\ell)\right\|_{(\Lambda_\ell^{(j)})^{-1}} + 4\beta\tau\log(2K/(\tau\delta)) \\
&= 2\sum_{j=1}^{K/\tau}\sum_{\ell=1}^{\tau} \beta\sqrt{\phi(s^\ell,a^\ell)^T(\Lambda_\ell^{(j)})^{-1}\phi(s^\ell,a^\ell)} + 4\beta\tau\log(2K/(\tau\delta)) \\
&\leq 2\sum_{j=1}^{K/\tau} \beta\sqrt{\tau\sum_{\ell=1}^{\tau}\phi(s^\ell,a^\ell)^T(\Lambda_\ell^{(j)})^{-1}\phi(s^\ell,a^\ell)} + 4\beta\tau\log(2K/(\tau\delta)) \\
&\leq 2\sum_{j=1}^{K/\tau} \beta\sqrt{\tau\sum_{i=1}^{d}\log\left(1+\lambda_i\right)} + 4\beta\tau\log(2K/(\tau\delta))
\end{aligned}$$

where the last inequality uses [9, Lemma 11.11] and the notation $\{\lambda_i\}_{i=1}^d$ stands for the eigenvalues of the matrix $\Lambda^{(j)} - I$. At this point we can recognize the determinant inside the log and use the

determinant trace inequality.

$$
\begin{aligned}
\sum_{j=1}^{K/\tau} \beta \sqrt{\tau \sum_{i=1}^{d} \log\left(1+\lambda_i\right)} &= \sum_{j=1}^{K/\tau} \beta \sqrt{\tau \log\left(\prod_{i=1}^{d} 1+\lambda_i\right)} \\
&= \sum_{j=1}^{K/\tau} \beta \sqrt{\tau \log\left(\det(\Lambda^{(j)})\right)} \\
&\leq \sum_{j=1}^{K/\tau} \beta \sqrt{\tau d \log\left(\frac{\mathrm{Trace}(\Lambda^{(j)})}{d}\right)} \\
&\leq \sum_{j=1}^{K/\tau} \beta \sqrt{\tau d \log\left(\frac{d+\mathrm{Trace}(\sum_{\ell=1}^{\tau} \phi(s^\ell,a^\ell)\phi(s^\ell,a^\ell)^T)}{d}\right)} \\
&\leq \sum_{j=1}^{K/\tau} \beta \sqrt{\tau d \log\left(\frac{d+\tau \max_\ell \mathrm{Trace}(\phi(s^\ell,a^\ell)\phi(s^\ell,a^\ell)^T)}{d}\right)} \\
&= \sum_{j=1}^{K/\tau} \beta \sqrt{\tau d \log\left(\frac{d+\tau \max_\ell \phi(s^\ell,a^\ell)^T\phi(s^\ell,a^\ell)}{d}\right)} \\
&\leq \sum_{j=1}^{K/\tau} \beta \sqrt{\tau d \log\left(\frac{d+\tau}{d}\right)} \\
&\leq \beta K \sqrt{\frac{d \log 2\tau d}{\tau}}.
\end{aligned}
$$

Hence, with probability $1-2\delta$ (union bound between the event under which Equation (5) holds and the the application of the concentration result [43, Lemma F.1] in bounding the exploration bonuses sum), it holds that

$$
\begin{aligned}
\mathrm{Regret}(K;\pi^\star) \leq \sum_{k=1}^{K} \mathbb{E}_{s\sim d^{\pi^\star}}\left[\left\langle Q^k(s,\cdot), \pi^k(s)-\pi^\star(s)\right\rangle\right] + 4\beta K\sqrt{\frac{d\log(2\tau d)}{\tau}} + 8\beta\tau\log(2K/(\tau\delta)) \\
+ \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau} + \frac{1}{(1-\gamma)}
\end{aligned}
$$

The last step is to bound (OMD) invoking [43, Lemma F.5] noticing that the gradient norm is upper bounded by $\frac{1}{1-\gamma}$. This gives

$$
\sum_{k=1}^{K} \mathbb{E}_{s\sim d^{\pi^\star}}\left[\left\langle Q^k(s,\cdot), \pi^k(s)-\pi^\star(s)\right\rangle\right] \leq \frac{\tau\log|\mathcal{A}|}{\eta} + \frac{\tau}{1-\gamma} + \frac{\eta K}{(1-\gamma)^2}
$$

Putting all together , we have that for $\tau \leq \frac{K}{\sqrt{\tau}}$

$$
\mathrm{Regret}(K;\pi^\star) \leq \frac{\tau\log|\mathcal{A}|}{\eta} + \frac{\tau+1}{1-\gamma} + \frac{\eta K}{(1-\gamma)^2} + 12\beta K\sqrt{\frac{d}{\tau}}\log\left(\frac{2Kd}{\tau\delta}\right) + \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau}
$$

□

F.3 Proof of Theorem 5

To improve the readability of the proof we restate Algorithm 3 hereafter.

Algorithm 5 ILARL (detailed version).

Input: Dataset size τ, Exploration parameter β, Step size η, Expert dataset $\mathcal{D}_{\pi_{\mathrm{E}}} = \{\boldsymbol{\tau}_i\}_{i=1}^{\tau_E}$.
Estimate for expert feature visitation $\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} \triangleq \frac{(1-\gamma)}{\tau_E} \sum_{\boldsymbol{\tau}\in\mathcal{D}_{\pi_{\mathrm{E}}}} \sum_{s_h,a_h\in\boldsymbol{\tau}} \gamma^h \phi(s_h,a_h)$.
Initialize π_0 as uniform distribution over $\mathcal{A}$
Initialize $V^1 = 0$
for $j = 1, \dots \lfloor K/\tau \rfloor$ **do**
  // Collect on-policy data
  Denote the indices interval $T_j \triangleq [(j-1)\lfloor K/\tau \rfloor, j\lfloor K/\tau \rfloor)$.
  Sample $\mathcal{D}^{(j)} = \left\{s^i, a^i, s'^{,i}\right\}_{i\in T_j} \sim d^{\pi^{(j)}}$
  Compute $\Lambda^{(j)} = \sum_{(s,a)\in\mathcal{D}^{(j)}} \phi(s,a)\phi(s,a)^\intercal + I$.
  Compute $b^{(j)}(s,a) = \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}}$.
  for $k \in T_j$ **do**
    // Cost update
    Estimate features expectation vector $\widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}}$ as $\tau^{-1}\sum_{s,a\in\mathcal{D}^{(j)}} \phi(s,a)$.
    $\mathbf{w}^{k+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}^k - \alpha(\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}})\right]$ with $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 \le 1\}$.
    // Optimistic Policy Evaluation
    $\mathbf{v}^k = (\Lambda^{(j)})^{-1}\sum_{(s,a,s')\in\mathcal{D}^{(j)}} \phi(s,a)V^k(s')$
    $Q^{k+1} = \left[\boldsymbol{\Phi}\mathbf{w}^k + \gamma\boldsymbol{\Phi}\mathbf{v}^k - b^{(j)}\right]_{-(1-\gamma)^{-1}}^{(1-\gamma)^{-1}}$
    $V^{k+1}(s) = \left\langle \pi^{(j)}(a|s), Q^{k+1}(s,a)\right\rangle$ (notice that $\pi^{(j)} = \pi^{k+1}$).
  end for
  // Policy Improvement Step
  Compute average Q value $\bar{Q}^{(j)}(s,a) = \frac{1}{\tau}\sum_{k\in T_j} Q^k(s,a)$.
  Update policy

$$\pi^{(j+1)}(a|s) \propto \exp\left(-\eta\sum_{i=1}^{j}\bar{Q}^{(i)}(s,a)\right)$$

end for

Proof. Consider the following decomposition

$$\sum_{k=1}^{K}\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\mathrm{E}}} - d^{\pi^k}\right\rangle = \sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \boldsymbol{\Phi}^\intercal d^{\pi^k}\right\rangle + \sum_{k=1}^{K}\left\langle c^k, d^{\pi_{\mathrm{E}}} - d^{\pi^k}\right\rangle \tag{6}$$

$$+ \sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}\right\rangle \tag{7}$$

For the first term we can use the following steps.

$$\sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \boldsymbol{\Phi}^\intercal d^{\pi^k}\right\rangle \le \sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}}\right\rangle + \sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}} - \boldsymbol{\Phi}^\intercal d^{\pi^k}\right\rangle$$

Now, using the regret bound for OMD [31, Theorem 6.10] we can bound the first term in the decomposition above as

$$\begin{aligned}
\sum_{k=1}^{K}\left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}}\right\rangle &\le \frac{\max_{\mathbf{w}\in\mathcal{W}}\left\|\mathbf{w} - \mathbf{w}^1\right\|_2^2}{2\alpha} + \frac{\alpha}{2}\sum_{k=1}^{K}\left\|\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}}\right\|_2^2 \\
&\le \frac{\max_{c\in\mathcal{C}}\left\|c - c^1\right\|_2^2}{2\alpha} + \frac{\alpha}{2}\sum_{k=1}^{K}\left\|\widehat{\boldsymbol{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} - \widehat{\boldsymbol{\Phi}^\intercal d^{\pi^k}}\right\|_1^2 \\
&\le \frac{1}{2\alpha} + 2\alpha K
\end{aligned}$$

Then, for $\alpha = \frac{1}{2\sqrt{K}}$, then $\sum_{k=1}^{K} \left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}} - \widehat{\mathbf{\Phi}^\intercal d^{\pi^k}} \right\rangle \leq 2\sqrt{K}$. For the estimation term,

$$\begin{aligned}
\sum_{k=1}^{K} \left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \mathbf{\Phi}^\intercal d^{\pi^k} - \widehat{\mathbf{\Phi}^\intercal d^{\pi^k}} \right\rangle &\leq \sum_{k=1}^{K} \left\| \mathbf{w} - \mathbf{w}^k \right\|_1 \left\| \mathbf{\Phi}^\intercal d^{\pi^k} - \widehat{\mathbf{\Phi}^\intercal d^{\pi^k}} \right\|_\infty \\
&\leq \sqrt{d} \sum_{k=1}^{K} \left\| \mathbf{w} - \mathbf{w}^k \right\|_2 \left\| \mathbf{\Phi}^\intercal d^{\pi^k} - \widehat{\mathbf{\Phi}^\intercal d^{\pi^k}} \right\|_\infty \\
&\leq 2\sqrt{d} \sum_{k=1}^{K} \left\| \mathbf{\Phi}^\intercal d^{\pi^k} - \widehat{\mathbf{\Phi}^\intercal d^{\pi^k}} \right\|_\infty \\
&\leq 2\sqrt{d} \sum_{k=1}^{K} \sqrt{\frac{2\log(2dK/\delta)}{\tau}} = 2K\sqrt{\frac{2d\log(2dK/\delta)}{\tau}},
\end{aligned}$$

where the last inequality holds with probability $1-\delta$ thanks to Azuma-Hoeffding inequality. Therefore, using Theorem 4 to control the second term in Equation (7) and a union bound we obtain that with probability $1 - 3\delta$.

$$\begin{aligned}
\sum_{k=1}^{K} \left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - d^{\pi^k} \right\rangle &\leq 2\sqrt{K} + 2K\sqrt{\frac{2d\log(2dK/\delta)}{\tau}} + \frac{\tau \log|\mathcal{A}|}{\eta} + \frac{\tau+1}{1-\gamma} + \frac{\eta K}{(1-\gamma)^2} \\
&\quad + 12\beta K\sqrt{\frac{d}{\tau}} \log\left(\frac{2K}{\tau\delta}\right) + \frac{\sqrt{2\eta}K}{(1-\gamma)^2\tau} + + \sum_{k=1}^{K} \left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}} - \mathbf{\Phi}^\intercal d^{\pi_{\text{E}}} \right\rangle
\end{aligned}$$

and using that for the empirical expert an application of Lemma 8 gives that with probability $1 - \delta$.

$$\sum_{k=1}^{K} \left\langle \mathbf{w}_{\text{true}} - \mathbf{w}^k, \widehat{\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}} - \mathbf{\Phi}^\intercal d^{\pi_{\text{E}}} \right\rangle \leq 2K\sqrt{d} \left\| \mathbf{\Phi}^\intercal d^{\pi_{\text{E}}} - \widehat{\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}} \right\|_\infty \leq 2K\sqrt{\frac{2d\log(d/\delta)}{\tau_E}}$$

Therefore, selecting $\tau_E \geq \frac{8d\log(d/\delta)}{\epsilon_E^2}$ and using a last union bound gives that with probability $1 - 4\delta$

$$\begin{aligned}
\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k} \right\rangle &\leq \epsilon_E + \frac{2}{\sqrt{K}} + \sqrt{\frac{8d\log(2dK/\delta)}{\tau}} + \frac{\tau\log|\mathcal{A}|}{\eta K} + \frac{\tau+1}{(1-\gamma)K} + \frac{\eta}{(1-\gamma)^2} \\
&\quad + 12\beta\sqrt{\frac{d}{\tau}} \log\left(\frac{2Kd}{\tau\delta}\right) + \frac{\sqrt{2\eta}}{(1-\gamma)^2\tau}
\end{aligned}$$

Using $\eta = \sqrt{\frac{\tau\log|\mathcal{A}|(1-\gamma)^2}{K}}$,

$$\begin{aligned}
\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k} \right\rangle &\leq \epsilon_E + \frac{2}{\sqrt{K}} + \sqrt{\frac{8d\log(2dK/\delta)}{\tau}} + \frac{2}{(1-\gamma)}\sqrt{\frac{\tau\log|\mathcal{A}|}{K}} + \frac{\tau+1}{(1-\gamma)K} \\
&\quad + 12\beta\sqrt{\frac{d}{\tau}} \log\left(\frac{2K}{\tau\delta}\right) + \frac{\sqrt{2}}{(1-\gamma)^2\tau}\sqrt[4]{\frac{\tau\log|\mathcal{A}|\,(1-\gamma)^2}{K}}
\end{aligned}$$

Neglecting lower order terms we obtain

$$\begin{aligned}
\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k} \right\rangle &\leq \epsilon_E + \mathcal{O}\left(\sqrt{\frac{8d\log(2dK/\delta)}{\tau}} + \frac{2}{(1-\gamma)}\sqrt{\frac{\tau\log|\mathcal{A}|}{K}} + 12\beta\sqrt{\frac{d}{\tau}}\log\left(\frac{2K}{\tau\delta}\right)\right) \\
&\leq \epsilon_E + \mathcal{O}\left(\frac{2}{(1-\gamma)}\sqrt{\frac{\tau\log|\mathcal{A}|}{K}} + 15\beta\sqrt{\frac{d}{\tau}}\log\left(\frac{2dK}{\tau\delta}\right)\right)
\end{aligned}$$

Therefore, choosing $\tau = \mathcal{O}\left(\frac{\beta(1-\gamma)\sqrt{dK}\log(2dK/\delta)}{\sqrt{\log|\mathcal{A}|}}\right)$, gives

$$\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k} \right\rangle \leq \epsilon_E + \widetilde{\mathcal{O}}\left(\frac{\log^{1/4}|\mathcal{A}|\, d^{1/4}\sqrt{\beta}}{\sqrt{1-\gamma}K^{1/4}}\right)$$

Therefore, choosing $K \geq \widetilde{\mathcal{O}}\left(\frac{\log|\mathcal{A}|d\beta^2 \log^2(2dK/\delta)}{(1-\gamma)^2\epsilon^4}\right)$ which is attained by $K = \mathcal{O}\left(\frac{\log|\mathcal{A}|d\beta^2}{(1-\gamma)^2\epsilon^4}\log^2(\frac{d\beta\log|\mathcal{A}|}{\delta(1-\gamma)\epsilon})\right)$ ensures $\left\langle \mathbf{c}_{\text{true}}, d^{\pi_{\text{E}}} - \frac{1}{K}\sum_{k=1}^{K} d^{\pi^k}\right\rangle \leq \epsilon + \epsilon_E$ with probability $1-4\delta$. □

G Technical Lemmas

Lemma 3. *Assume that in Algorithm 1, we set $\beta = \mathcal{O}\left(dH\log(RdH\delta^{-1})\right)$ with $R = \tau^2\sqrt{d}H$. Then, with probability $1-\delta$ it holds that*

$$-2b_h^k(s,a) \leq Q_h^k(s,a) - \mathbf{c}_h^k(s,a) - PV_{h+1}^k(s,a) \leq 0 \quad \forall s,a \in \mathcal{S}\times\mathcal{A}, k\in[K], h\in[H]$$

Proof. We first see that Lemma 5 holds for every j. Then, thanks to union bound we have that with probability $1-\delta$ it holds that for any state action pairs we have that

$$\left|\phi(s,a)^T\mathbf{v}_h^k - PV_{h+1}^k(s,a)\right| \leq \beta\left\|\phi(s,a)\right\|_{(\Lambda_h^{(j)})^{-1}} = b_h^k(s,a)$$

From this fact the conclusion follows immediately if no truncation happens. That is , if $Q_h^k = \mathbf{c}_h^k + \mathbf{\Phi}\mathbf{v}_h^k - b_h^k$. Now, we consider the case where a lower truncation takes place, in this case, we have

$$Q_h^k = -H + h - 1 \leq \mathbf{c}_h^k + PV_{h+1}^k$$

If there a truncation from above it holds that

$$Q_h^k \leq \mathbf{c}_h^k + \mathbf{\Phi}\mathbf{v}_h^k - b_h^k \leq \mathbf{c}_h^k + PV_{h+1}^k + b_h^k - b_h^k = \mathbf{c}_h^k + PV_{h+1}^k$$

To show the lower bound in the lemma in case of a lower truncation, we have that

$$Q_h^k \geq \mathbf{c}_h^k + \mathbf{\Phi}\mathbf{v}_h^k - b_h^k \geq \mathbf{c}_h^k + PV_{h+1}^k - b_h^k - b_h^k = \mathbf{c}_h^k + PV_{h+1}^k - 2b_h^k$$

finally, if the truncation from above is triggered, we have that

$$Q_h^k = H - h + 1 \geq \mathbf{c}_h^k + PV_{h+1}^k \geq \mathbf{c}_h^k + PV_{h+1}^k - 2b_h^k$$

□

Lemma 4. *For any $k \in [K]$, let the bonus b^k be defined as in Algorithm 2 with $\beta = \mathcal{O}\left(d(1-\gamma)^{-1}\log(Rd(1-\gamma)^{-1}\delta^{-1})\right)$ with $R = \tau^2\sqrt{d}(1-\gamma)^{-1}$. Then, with probability $1-\delta$ it holds that*

$$-2b^k(s,a) \leq Q^{k+1}(s,a) - \mathbf{c}^k(s,a) - \gamma PV^k(s,a) \leq 0 \quad \forall s,a \in \mathcal{S}\times\mathcal{A}, k\in[K]$$

Proof. We first see that Lemma 6 holds for every j. Then, thanks to union bound we have that with probability $1-\delta$ it holds that for any state action pairs we have that

$$\left|\phi(s,a)^T\mathbf{v}^k - PV^k(s,a)\right| \leq \beta\left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} = b^k(s,a)$$

From this fact the conclusion follows immediately if no truncation happens. That is , if $Q^{k+1} = \mathbf{c}^k + \mathbf{\Phi}\mathbf{v}^k - b^k$. Now, we consider the case where a upper truncation takes place, in this case, we have

$$Q^{k+1} = \frac{1}{1-\gamma} = 1 + \frac{\gamma}{1-\gamma} \geq \mathbf{c}^k + \gamma PV^k - 2b^k$$

While for the upper bound, we have that

$$Q^{k+1} \leq \mathbf{c}^k + \gamma\mathbf{\Phi}\mathbf{v}^k - b^k \leq \mathbf{c}^k + \gamma PV^k + b^k - b^k = \mathbf{c}^k + \gamma PV^k$$

Now, we handle the case of a lower truncation in this case

$$Q^{k+1} \geq \mathbf{c}^k + \gamma\mathbf{\Phi}\mathbf{v}^k - b^k \geq \mathbf{c}^k + \gamma PV^k - b^k - b^k = \mathbf{c}^k + \gamma PV^k - 2b^k$$

for the upper bound we have that

$$Q^{k+1} = -\frac{1}{1-\gamma} = -1 - \frac{\gamma}{1-\gamma} \leq \mathbf{c}^k + \gamma PV^k$$

□

Lemma 5. *Let V_h^k be the sequence of value functions generated by Algorithm 1, fix a batch index j and let T_j denote the set of indices in the jth batch. Then, it holds that for $\beta = \widetilde{\mathcal{O}}(dH)$, the estimator*

$$\mathbf{v}_h^k = (\Lambda^{(j)})^{-1} \sum_{(s,a,s') \in \mathcal{D}_h^{(j)}} \phi(s,a) V_{h+1}^k(s')$$

satisfies

$$\left|\phi(s,a)^T \mathbf{v}_h^k - PV_{h+1}^k(s,a)\right| \leq \beta \left\|\phi(s,a)\right\|_{(\Lambda_h^{(j)})^{-1}} = b^k(s,a) \quad \forall k \in T_j, h \in [H], \forall s,a \in \mathcal{S} \times \mathcal{A}$$

with probability $1 - \delta\tau/K$.

Proof. The proof is analogous to the proof of Lemma 6 but invoking Theorem 7 with $B = H$ and applying a further union bound over the set $[H]$. Thus, the proof is skipped for brevity. □

Lemma 6. *Let V^k be the sequence of value functions generated by Algorithm 2, fix a batch index j and let T_j denote the set of indices in the jth batch. Then, it holds that for $\beta = \widetilde{\mathcal{O}}\left(\frac{d}{1-\gamma}\right)$, the estimator*

$$\mathbf{v}^k = (\Lambda^{(j)})^{-1} \sum_{(s,a,s') \in \mathcal{D}^{(j)}} \phi(s,a) V^k(s')$$

satisfies

$$\left|\phi(s,a)^T \mathbf{v}^k - PV^k(s,a)\right| \leq \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} = b^k(s,a) \quad \forall k \in T_j, \forall s,a \in \mathcal{S} \times \mathcal{A}$$

with probability $1 - \delta\tau/K$.

Proof. With standard manipulation one can prove that

$$PV^k = \mathbf{\Phi} M V^k = \mathbf{\Phi}(\Lambda^{(j)})^{-1} M V^k + \mathbf{\Phi}(\Lambda^{(j)})^{-1} \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a) PV^k(s,a)$$

and by definition

$$\mathbf{\Phi}\mathbf{v}^k = \mathbf{\Phi}(\Lambda^{(j)})^{-1} \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a) V^k(s')$$

Therefore,

$$PV^k - \mathbf{\Phi}\mathbf{v}^k = \mathbf{\Phi}(\Lambda^{(j)})^{-1} M V^k + \mathbf{\Phi}(\Lambda^{(j)})^{-1} \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a)(PV^k(s,a) - V^k(s'))$$

Then, for any state action pair (s,a), we have

$$\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| = \phi(s,a)^T (\Lambda^{(j)})^{-1} M V^k + \phi(s,a)^T (\Lambda^{(j)})^{-1} \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a)(PV^k(s,a) - V^k(s'))$$

by applying Holder's inequality,

$$\begin{aligned}
\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| &\leq \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} \left\|(\Lambda^{(j)})^{-1} M V^k\right\|_{(\Lambda^{(j)})} \\
&\quad + \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} \left\|(\Lambda^{(j)})^{-1} \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a)(PV^k(s,a) - V^k(s'))\right\|_{(\Lambda^{(j)})} \\
&= \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} \left\|M V^k\right\|_{(\Lambda^{(j)})^{-1}} \\
&\quad + \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}} \left\|\sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a)(PV^k(s,a) - V^k(s'))\right\|_{(\Lambda^{(j)})^{-1}}
\end{aligned}$$

where in the equality we used that for a symmetric matrix A we have that $\|Ax\|_{A^{-1}} = \|x\|_A$. Then, we can use that $\left\|MV^k\right\|_{(\Lambda^{(j)})^{-1}} \leq \frac{1}{1-\gamma}$ to obtain

$$\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| \leq \|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}} \left(\frac{1}{1-\gamma} + \left\| \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a)(PV^k(s,a) - V^k(s')) \right\|_{(\Lambda^{(j)})^{-1}} \right)$$

To handle the second term in brackets we use that $V^k \in \mathcal{V}^{\pi^{(j)}}$ defined as

$$\mathcal{V}^{\pi^{(j)}} = \left\{ \left\langle \pi^{(j)}(\cdot|s), Q(s,\cdot) \right\rangle | Q(s,a) \in \mathcal{Q}(\beta, \Lambda, \mathbf{w}, \mathbf{v}) \right\}$$

where $\mathcal{Q}(\beta, \Lambda, \mathbf{w}, \mathbf{v})$ is defined as in Theorem 8. Denote as $\mathcal{N}_\epsilon^j$ the $\|\cdot\|_\infty$-covering number of the class $\mathcal{V}^{\pi^{(j)}}$ and notice that $\mathcal{V}^{\pi^{(j)}}$ and $\mathcal{D}^{(j)}$ are conditionally independent given $\pi^{(j)}$.

Under this setting we can use Theorem 7 with $B = (1-\gamma)^{-1}$ to obtain that with probability $1-\delta\tau/K$

$$\begin{aligned}\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| &\leq \|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}} \left(\frac{1}{1-\gamma} + \sqrt{\frac{2d}{(1-\gamma)^2} \log\left(\frac{K(\tau+1)}{\delta\tau}\right) + \frac{4}{(1-\gamma)^2} \log \mathcal{N}_\epsilon^j + 8\tau^2\epsilon^2} \right) \\ &\leq \frac{\|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}}}{1-\gamma} \left(1 + \sqrt{2d \log\left(\frac{K(\tau+1)}{\delta\tau}\right)} + 2\sqrt{\log \mathcal{N}_\epsilon^j} + 2\sqrt{2}\tau\epsilon \right).\end{aligned}$$

Then, we can conclude that the covering number is upper bounded by Theorem 8 with $L = \frac{\tau}{1-\gamma}$ since

$$\left\|\mathbf{v}^k\right\| \leq \frac{1}{(1-\gamma)} \left\|(\Lambda^{(j)})^{(-1)}\right\| \left\| \sum_{s,a,s' \in \mathcal{D}^{(j)}} \phi(s,a) \right\| \leq \frac{\tau}{(1-\gamma)}$$

we obtain that

$$\log \mathcal{N}_\epsilon^j \leq d \log\left(1 + \frac{4}{\epsilon}\sqrt{1 + \frac{\gamma^2\tau^2}{(1-\gamma)^2}}\right) + d^2 \log(1 + 8\sqrt{d}\beta^2\epsilon^{-2})$$

Therefore,

$$\begin{aligned}\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| \leq \frac{\|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}}}{1-\gamma} \Bigg[&1 + \sqrt{2d \log\left(\frac{K(\tau+1)}{\delta\tau}\right)} + \sqrt{2d \log\left(1 + \frac{4}{\epsilon}\sqrt{1 + \frac{\gamma^2\tau^2}{(1-\gamma)^2}}\right)} \\ &+ \sqrt{2}d\sqrt{\log(1 + 8\sqrt{d}\beta^2\epsilon^{-2})} + 2\sqrt{2}\tau\epsilon \Bigg]\end{aligned}$$

At this point, using $\epsilon = \tau^{-1}$, we obtain

$$\begin{aligned}\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| \leq \frac{\|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}}}{1-\gamma} \Bigg[&1 + \sqrt{2d \log\left(\frac{K(\tau+1)}{\delta\tau}\right)} + \sqrt{2d \log\left(1 + 4\tau\sqrt{1 + \frac{\gamma^2\tau^2}{(1-\gamma)^2}}\right)} \\ &+ \sqrt{2}d\sqrt{\log(1 + 8\sqrt{d}\beta^2\tau^2)} + 2\sqrt{2} \Bigg]\end{aligned}$$

To simplify the above expression, we notice that there exists a constant c such that

$$\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| \leq c\frac{\|\phi(s,a)\|_{(\Lambda^{(j)})^{-1}}}{1-\gamma} d\sqrt{\log\left(\frac{\tau K\beta\sqrt{d}}{\delta(1-\gamma)}\right)}$$

Now, using [43, Lemma D.2], we have that $\beta = \mathcal{O}(d(1-\gamma)^{-1}\log\left[Rd(1-\gamma)^{-1}\right])$ with $R = \tau K\sqrt{d}\delta^{-1}(1-\gamma)^{-1}$ ensures

$$\beta \geq c\frac{d}{(1-\gamma)} \log\left(\frac{\tau^2\beta\sqrt{d}}{\delta(1-\gamma)}\right),$$

and therefore

$$\left|PV^k(s,a) - \mathbf{\Phi}\mathbf{v}^k(s,a)\right| \leq \beta \left\|\phi(s,a)\right\|_{(\Lambda^{(j)})^{-1}}.$$

□

Theorem 7. *For a fixed policy π consider a function class $\mathcal{V}$ and a state action pair dataset $\mathcal{D}$ collected with a fixed policy π such that $\mathcal{D}$ and $\mathcal{V}$ are conditionally independent given π . Then, for any $f \in \mathcal{V}$ such that $\|f\|_\infty \leq B$, it holds with probability $1 - \delta\tau/K$*

$$\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(Pf(s,a) - f(s'))\right\|^2_{(\Lambda^{(j)})^{-1}} \leq 2dB^2 \log\left(\frac{K(\tau+1)}{\delta\tau}\right) + 4B^2 \log \mathcal{N}_\epsilon + 8\tau^2\epsilon^2$$

where N_ϵ is the the $(\epsilon, \|\cdot\|_\infty)$- covering number of the class $\mathcal{V}^\pi$.

Proof. Consider the decomposition in [22, Lemma D.4]. In particular, let $\mathcal{C}_\epsilon(\mathcal{V})$ denote the $(\epsilon, \|\cdot\|_\infty)$-covering set of $\mathcal{V}$ and pick $\tilde{f} \in \mathcal{C}_\epsilon(\mathcal{V})$ such that $\left\|f - \tilde{f}\right\|_\infty \leq \epsilon$. The existence of $\tilde{f}$ is guaranteed by the properties of covering sets. Then, we have

$$\begin{aligned}\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(Pf(s,a) - f(s'))\right\|^2_{(\Lambda^{(j)})^{-1}} &\leq 2\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(P\tilde{f}(s,a) - \tilde{f}(s'))\right\|^2_{(\Lambda^{(j)})^{-1}} \\ &\quad + 2\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(P(f-\tilde{f})(s,a) - (f-\tilde{f})(s'))\right\|^2_{(\Lambda^{(j)})^{-1}}\end{aligned}$$

The second term can be bounded by $8\tau^2\epsilon^2$ as in [22] so now we focus on the first term via a uniform bound over the set $\mathcal{C}_\epsilon(\mathcal{V})$. We need to index the dataset $\mathcal{D}$, i.e. $\mathcal{D} = \left\{(s^\ell, a^\ell)\right\}_{\ell=1}^{|\mathcal{D}|}$ and consider the filtration $\mathcal{F}_j = \left\{(s^\ell, a^\ell)\right\}_{\ell=1}^{j}$. Since the features mapping is deterministic, $\phi(s^\ell, a^\ell)$ is $\mathcal{F}_\ell$-measurable. Then, notice that by assumption $\mathcal{D}$ and $\mathcal{V}^\pi$ are conditionally independent given π. Therefore, we also have that $\mathcal{D}$ and $\mathcal{C}_\epsilon(\mathcal{V}^\pi)$ are conditionally independent given π. So for any function $\bar{f} \in \mathcal{C}_\epsilon(\mathcal{V}^\pi)$ we have that $\mathbb{E}[\bar{f}(s^{\ell+1})|\mathcal{F}_\ell] = P\bar{f}(s^\ell, a^\ell)$. Finally, from the assumption $\left\|\bar{f}\right\|_\infty \leq B$ we have that $\bar{f}$ is B^2-subgaussian. Therefore, all the conditions of [22, Theorem D.3] are met and via a union bound over the covering set allows to conclude that with probability $1 - \delta\tau/K$

$$2\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(P\bar{f}(s,a) - \bar{f}(s'))\right\|^2_{(\Lambda^{(j)})^{-1}} \leq 2dB^2 \log\left(\frac{K(\tau+1)}{\delta\tau}\right) + 4B^2 \log \mathcal{N}_\epsilon \quad \forall \bar{f} \in \mathcal{C}_\epsilon(\mathcal{V}),$$

and since $\tilde{f} \in \mathcal{C}_\epsilon(\mathcal{V})$,

$$2\left\|\sum_{s,a,s'\in\mathcal{D}} \phi(s,a)(P\tilde{f}(s,a) - \tilde{f}(s'))\right\|^2_{(\Lambda^{(j)})^{-1}} \leq 2dB^2 \log\left(\frac{K(\tau+1)}{\delta\tau}\right) + 4B^2 \log \mathcal{N}_\epsilon.$$

□

Theorem 8. *Let us consider the function class $\mathcal{Q}$ defined as follows*

$$\mathcal{Q}(\beta, \Lambda, \mathbf{w}, \mathbf{v}) = \{Q(s,a;\beta,\Lambda,\mathbf{w},\mathbf{v}) | \beta \in \mathbb{R}, \lambda_{\min}(\Lambda) \geq 1, \|\mathbf{w}\| \leq 1, \|\mathbf{v}\| \leq L\}$$

$$\textit{where} \quad Q(s,a;\beta,\Lambda,\mathbf{w},\mathbf{v}) = [\phi(s,a)^\intercal(\mathbf{w} + \gamma\mathbf{v}) + \beta\left\|\phi(s,a)\right\|_{\Lambda^{-1}}]_0^{(1-\gamma)^{-1}}$$

and the classes

$$\mathcal{V}^\pi = \{\langle \pi(\cdot|s), Q(s,\cdot)\rangle \,|Q(s,a) \in \mathcal{Q}(\beta,\Lambda,\mathbf{w},\mathbf{v})\}$$

for any $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$.

Then, it holds that for any $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$

$$\mathcal{N}_\epsilon(\mathcal{V}^\pi) \leq \mathcal{N}_\epsilon(\mathcal{Q}) = (1 + 4\sqrt{1+\gamma^2L^2}/\epsilon)^d(1 + 8\sqrt{d}\beta^2\epsilon^{-2})^{d^2}$$

Proof. Let us remove clipping that can only decreasing the covering number of the function class and let us consider the matrix $A = \beta^2\Lambda^{-1}$, then we can rewrite the function class of interest as parameterized only by A rather then β and Λ separately. In addition, let us consider a vector $\mathbf{z} = \mathbf{w} + \gamma\mathbf{v}$

$$\mathcal{Q}(A, \mathbf{z}) = \left\{ Q(s, a; A, \mathbf{w}, \mathbf{v}) | \lambda_{\min}(\Lambda) \geq \beta^2, \|\mathbf{z}\|^2 \leq 2 + 2\gamma^2 L^2 \right\}$$

with

$$Q(s, a; A, \mathbf{z}) = \phi(s, a)^\intercal \mathbf{z} + \|\phi(s, a)\|_A$$

Then, we have that

$$\begin{aligned}
|Q(s, a; A_1, \mathbf{z}_1) - Q(s, a; A_2, \mathbf{z}_2)| &\leq \|\phi(s, a)\| \|\mathbf{z}_1 - \mathbf{z}_2\| + \left| \sqrt{\phi(s, a)^\intercal A_1 \phi(s, a)} - \sqrt{\phi(s, a)^\intercal A_2 \phi(s, a)} \right| \\
&\leq \|\phi(s, a)\| \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{|\phi(s, a)^\intercal (A_1 - A_2)\phi(s, a)|} \\
&\leq \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{\sup_{\phi : \|\phi\| \leq 1} |\phi(s, a)^\intercal (A_1 - A_2)\phi(s, a)|} \\
&\leq \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{\|A_1 - A_2\|} \\
&\leq \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{\|A_1 - A_2\|_F}
\end{aligned}$$

where $\|A_1 - A_2\|$ is the spectral norm of the matrix $A_1 - A_2$ and $\|A_1 - A_2\|_F$ is the Frobenius norm. We also used the inequality $\left|\sqrt{x} - \sqrt{y}\right| \leq \sqrt{|x - y|}$ that holds for any $x, y \geq 0$. At this point we can constructing an ϵ-covering set for $\mathcal{Q}(A, \mathbf{z})$ as product of the $\epsilon^2/4$ covering set for the set $\mathcal{Y} = \left\{ A \in \mathbb{R}^{d \times d} | \|A\|_F \leq \sqrt{d}\beta^{-2} \right\}$ which has cardinality $\mathcal{N}_\epsilon(\mathcal{Y}) = (1 + 8\sqrt{d}\beta^2\epsilon^{-2})^{d^2}$ while the covering for the set $\mathcal{Z} = \left\{ \mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|^2 \leq 1 + \gamma^2 L^2 \right\}$ satisfies $\mathcal{N}_\epsilon(\mathcal{Z}) = (1 + 4\sqrt{1 + \gamma^2 L^2}/\epsilon)^d$. Hence, taking the product, we have that

$$\mathcal{N}_\epsilon(\mathcal{Q}) = (1 + 4\sqrt{1 + \gamma^2 L^2}/\epsilon)^d (1 + 8\sqrt{d}\beta^2\epsilon^{-2})^{d^2}$$

.

At this point, let us consider the set

$$\mathcal{V}^\pi = \{\langle \pi(\cdot|s), Q(s, \cdot) \rangle \, | Q(s, a) \in \mathcal{Q}(A, \mathbf{z})\}$$

Since the policy π is fixed and averaging is a non expansive operation, we have that $\mathcal{N}_\epsilon(\mathcal{V}^\pi) \leq \mathcal{N}_\epsilon(\mathcal{Q})$.

However, for the set

$$\mathcal{V} = \{\langle \pi(\cdot|s), Q(s, \cdot) \rangle \, | \pi \in \Pi, Q(s, a) \in \mathcal{Q}(\beta, \Lambda, \mathbf{w}, \mathbf{v})\}$$

the averaging is not wrt to a fixed distribution therefore we would need to proceed as follow

$$\begin{aligned}
|\langle \pi_1(\cdot|s), Q_1(s, \cdot) \rangle - \langle \pi_2(\cdot|s), Q_2(s, \cdot) \rangle| &\leq \frac{1}{(1 - \gamma)} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + 2 \|Q_1(s, \cdot) - Q_2(s, \cdot)\|_\infty \\
&\leq \frac{1}{(1 - \gamma)} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + 2 \|Q_1(s, \cdot) - Q_2(s, \cdot)\|_\infty \\
&\leq \frac{1}{(1 - \gamma)} \max_{s \in \mathcal{S}} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + 2 \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{\|A_1 - A_2\|_F} \\
&\leq \frac{1}{(1 - \gamma)} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_{\infty,1} + 2 \|\mathbf{z}_1 - \mathbf{z}_2\| + \sqrt{\|A_1 - A_2\|_F}
\end{aligned}$$

Therefore, we can conclude that $\mathcal{N}_\epsilon(\mathcal{V}) = \mathcal{N}_\epsilon(\Pi, \|\cdot\|_{\infty,1}) \mathcal{N}_\epsilon(\mathcal{Q})$. □

Next, we prove the Lemma that we use to state Theorems 1 and 2 using $\delta = \epsilon$.

Lemma 7. ***High probability to expectation conversion for a bounded random variable*** *Let us consider a random variable X such that $-X_{\max} \leq X \leq X_{\max}$ almost surely and that $\mathbb{P}[X \geq \mu] \leq \delta$, then it holds that*

$$\mathbb{E}[X] \leq \mu + \delta(X_{\max} - \mu)$$

Proof.

$$\mathbb{E}[X] = (1-\delta)\mathbb{E}[X|X \le \mu] + \delta\mathbb{E}[X|X \ge \mu] \le (1-\delta)\mu + \delta X_{\max}$$

□

Lemma 8. ***Expert concentration*** *[46, Theorem 1] Let $\mathcal{D}_{\pi_{\mathrm{E}}} \triangleq \{(s_0^\ell, a_0^\ell, s_1^\ell, a_1^\ell, \ldots, s_H^\ell, a_H^\ell)\}_{\ell=1}^{n_{\mathrm{E}}}$ be a finite set of i.i.d. truncated sample trajectories collected with an expert policy π_{E}. We consider the empirical expert feature expectation vector $\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}$ by taking sample averages, i.e.,*

$$\widehat{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} \triangleq (1-\gamma)\frac{1}{n_{\mathrm{E}}}\sum_{t=0}^{H}\sum_{\ell=1}^{N}\gamma^t\phi_i(s_t^\ell, a_t^\ell),\ \forall\, i \in [d].$$

Suppose the trajectory length is $H \ge \frac{1}{1-\gamma}\log(\frac{1}{\varepsilon})$, and the number of of expert trajectories is $n_{\mathrm{E}} \ge \frac{2\log(\frac{2d}{\delta})}{\varepsilon^2}$. Then, with probability at least $1-\delta$, it holds that $\left\|\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}} - \widehat{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}}\right\|_\infty \le \varepsilon$.

H Omitted proofs for Best Response Imitation Learning

H.1 Proof of Theorem 6

Proof.

$$\begin{aligned}
\mathrm{Regret}(K, \pi^\star) &= \sum_{k=1}^{K} V_1^{\pi^k,k}(s_1) - V_1^{\pi^\star,k} \\
&= \sum_{k=1}^{K} V_1^{\pi^k,k}(s_1) - V_1^{k-1}(s_1) + V_1^{k-1}(s_1) - V_1^{\pi^\star,k} \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[\mathbf{c}_h^k(s,a) + P_h V_{h+1}^{k-1}(s,a) - Q_h^{k-1}(s,a)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^\star}}\left[Q_h^{k-1}(s,a) - \mathbf{c}_h^k(s,a) - P_h V_{h+1}^{k-1}(s,a)\right] \\
&\quad + \sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{k=1}^{K}\left\langle \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s), Q_h^{k-1}(s,a)\right\rangle\right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[\mathbf{c}_h^k(s,a) + P_h V_{h+1}^{k-1}(s,a) - Q_h^{k-1}(s,a)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^\star}}\left[Q_h^{k-1}(s,a) - \mathbf{c}_h^k(s,a) - P_h V_{h+1}^{k-1}(s,a)\right]
\end{aligned}$$

where the last inequality is due to the use of the best response (greedy policy) in Step 17 of Algorithm 4. At this point we can prove the optimistic properties of the estimator (that follows combining Lemmas 4 and 9),i.e. for any $h = H, \ldots, 1$, it holds that

$$\mathbf{c}_h^k(s,a) + P_h V_{h+1}^{k-1}(s,a) - 2b_h^{k-1}(s,a) \le Q_h^{k-1}(s,a) \le \mathbf{c}_h^k(s,a) + P_h V_{h+1}^{k-1}(s,a) \quad \forall s,a \in \mathcal{S}\times\mathcal{A} \quad w.p. \quad 1-\delta.$$

Thus, it holds with probability $1-\delta$ that

$$\mathrm{Regret}(K, \pi^\star) \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[b_h^{k-1}(s,a)\right]$$

and then, using Cauchy-Schwartz and the elliptical potential lemma (see Lemma 10), we obtain that with probability $1-2\delta$,

$$\mathrm{Regret}(K, \pi^\star) \le \mathcal{O}\left(H^2 d^{3/2}\sqrt{K\log(K\delta^{-1})}\right)$$

Then, we apply this result in the imitation learning setting. We start with our usual decomposition

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_{\text{true},h}, d_h^{\pi^k} - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_h^k, d_h^{\pi^k} - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle + \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_{\text{true},h} - \mathbf{c}_h^k, d_h^{\pi^k} - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_h^k, d_h^{\pi^k} - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle + \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_{\text{true},h} - \mathbf{c}_h^k, \mathbb{1}\left[s_k^h, a_k^h\right] - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle \\ &\quad - \sum_{h=1}^{H}\sum_{k=1}^{K}\left\langle \mathbf{c}_{\text{true},h} - \mathbf{c}_h^k, \mathbb{1}\left[s_k^h, a_k^h\right] - d_h^{\pi^k}\right\rangle\end{aligned}$$

Then, notice that $z_h^k = -\left\langle \mathbf{c}_{\text{true},h} - \mathbf{c}_h^k, \mathbb{1}\left[s_k^h, a_k^h\right] - d_h^{\pi^k}\right\rangle$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_h^k = \left\{\left(\tau^k, \mathbf{c}_h^k\right)\right\}$ almost surely bounded by 4. Therefore, by Azuma Hoeffding inequality, we obtain $\sum_{h=1}^{H}\sum_{k=1}^{K} z_h^k \leq H\sqrt{8K\log\delta^{-1}}$. Therefore, via a union bound, we have that with probability $1-3\delta$, it holds

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \mathbf{c}_{\text{true},h}, d_h^{\pi^k} - d_h^{\widehat{\pi_{\text{E}}}}\right\rangle &\leq \widetilde{\mathcal{O}}\left(H^2 d^{3/2}\sqrt{K\log(K\delta^{-1})}\right) + \frac{H}{\alpha} + 2\alpha KH + H\sqrt{8K\log\delta^{-1}} \\ &= \widetilde{\mathcal{O}}\left(H^2 d^{3/2}\sqrt{K\log(K\delta^{-1})}\right) + 4H\sqrt{K} + H\sqrt{8K\log\delta^{-1}}\end{aligned}$$

where last step follows from choosing $\alpha = \frac{1}{\sqrt{2K}}$. At this point, the conclusion holds plugging in the value for K in the statement of the main theorem which is $K = \mathcal{O}\left(\frac{H^4 d^3 \log(dH/(\epsilon\delta))}{\epsilon^2}\right)$. Finally, we need to control the error in the estimation of the expert occupancy measure that can be done as in the proof for Algorithm 1.

$$\begin{aligned}V_1^{\pi_{\text{E}}}(s_1;\mathbf{c}_{\text{true}}) - V_1^{\widehat{\pi_{\text{E}}}}(s_1;\mathbf{c}_{\text{true}}) &= \sum_{h=1}^{H}\left\langle \mathbf{\Phi}^\intercal d_h^{\pi_{\text{E}}} - \mathbf{\Phi}^\intercal d_h^{\widehat{\pi_{\text{E}}}}, w_{\text{true},h}\right\rangle \\ &\leq H\sqrt{d}\max_{h\in[H]}\left\|\mathbf{\Phi}^\intercal d_h^{\pi_{\text{E}}} - \mathbf{\Phi}^\intercal d_h^{\widehat{\pi_{\text{E}}}}\right\|_\infty \\ &\leq H\sqrt{\frac{2d\log(2d/\delta)}{\tau_E}}\end{aligned}$$

where the last inequality holds with probability $1-\delta$. Therefore, the choice of τ_E in the theorem statement ensures that $V_1^{\pi_{\text{E}}}(s_1;\mathbf{c}_{\text{true}}) - V_1^{\widehat{\pi_{\text{E}}}}(s_1;\mathbf{c}_{\text{true}}) \leq \epsilon_E$. □

Lemma 9. *For $\beta = \mathcal{O}\left(dH\log(\frac{dT}{\delta})\right)$, the estimator used in Algorithm 4*

$$\mathbf{v}_h^k = \left(\Lambda_h^k\right)^{-1}\sum_{l=1}^{k}\phi(s_h^l, a_h^l)V_h^k(s_{h+1}^l)$$

satisfies for any $h, k \in [H]\times[K]$ and for any state action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$.

$$\left|\phi(s,a)^\intercal\mathbf{v}_h^k - P_h V_{h+1}^k(s,a)\right| \leq \beta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^k\right)^{-1}} \tag{8}$$

with probability $1-\delta$.

Proof. With analogous steps to the proof of Lemma 6 that

$$\begin{aligned}\left|\phi(s,a)^\intercal\mathbf{v}_h^k - P_h V_{h+1}^k(s,a)\right| &\leq \left\|MV_{h+1}^k\right\|_{\left(\Lambda_h^k\right)^{-1}}\left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}} \\ &\quad + \left\|\sum_{l=1}^{k}\phi(s_h^l, a_h^l)\left(V_{h+1}^k(s_{h+1}^l) - P_h V_{h+1}^k(s_h^l, a_h^l)\right)\right\|_{\left(\Lambda_h^k\right)^{-1}}\left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}}\end{aligned}$$

Then, using the fact that by assumption on M and by the clipping of the value function, we have that $\left\|MV_{h+1}^k\right\|_{\left(\Lambda_h^k\right)^{-1}} \leq H$. Then, using [22, Lemma B.3] it holds that with probability $1-\delta$

$$\left\|\sum_{l=1}^{k} \phi(s_h^l, a_h^l)\left(V_{h+1}^k(s_{h+1}^l) - P_h V_{h+1}^k(s_h^l, a_h^l)\right)\right\|_{\left(\Lambda_h^k\right)^{-1}} \leq \mathcal{O}\left(dH \log\left(\frac{dK}{\delta}\right)\right)$$

□

Then, noticing that this is the main term we conclude that

$$\left|\phi(s,a)^\intercal \mathbf{v}_h^k - P_h V_{h+1}^k(s,a)\right| \leq \mathcal{O}\left(dH \log\left(\frac{dK}{\delta}\right)\right) \|\phi(s,a)\|_{\left(\Lambda_h^k\right)^{-1}}.$$

Lemma 10. *It holds that with probability* $1-\delta$

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[b_h^{k-1}(s,a)\right] \leq \mathcal{O}\left(d^{3/2}H^2\sqrt{K\log\left(\frac{2K}{\delta}\right)}\right)$$

Proof. We have that

$$\begin{aligned}
2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[b_h^{k-1}(s,a)\right] &\leq 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^{k-1}(s_h^k, a_h^k) + \beta H\sqrt{K\log\delta^{-1}} \quad \text{(Azuma-Hoeffding)} \\
&\leq 2\sum_{h=1}^{H}\sqrt{K\sum_{k=1}^{K}(b_h^{k-1}(s_h^k,a_h^k))^2} + \beta H\sqrt{K\log\delta^{-1}} \quad \text{(Cauchy-Schwartz)} \\
&= 2\beta\sum_{h=1}^{H}\sqrt{K\sum_{k=1}^{K}\phi(s_h^k,a_h^k)^\intercal(\Lambda_h^{k-1})\phi(s_h^k,a_h^k)} + \beta H\sqrt{K\log\delta^{-1}} \\
&\leq 2\beta\sum_{h=1}^{H}\sqrt{dK\log(2K)} + \beta H\sqrt{K\log\delta^{-1}} \\
&= \mathcal{O}\left(d^{3/2}H^2\sqrt{K\log\left(\frac{2K}{\delta}\right)}\right)
\end{aligned}$$

□

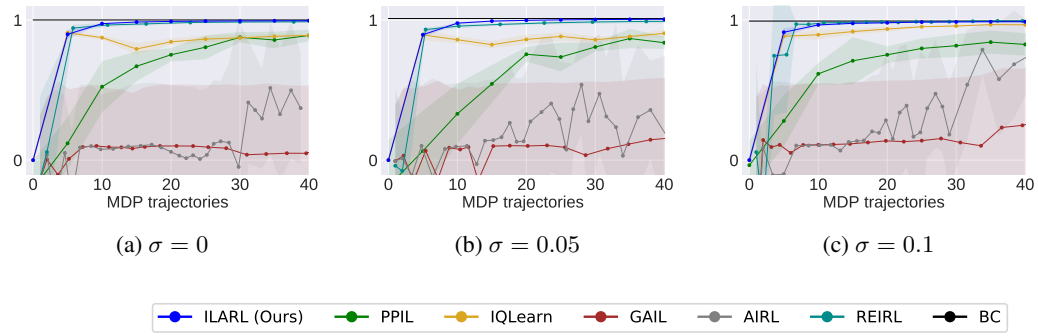
I Experiments

I.1 Experiments with deterministic expert

We also run an experiment where the expert is deterministic and see if ILARL can compete with BC in this setting. The results are provided in Figure 2. The parameter σ is the probability at which the system does not evolve according to the agent's action but in an adversarial way. We experiment with $\sigma = \{0, 0.05, 0.1\}$. The details about the transition dynamics are given in Appendix I.2. Form Figure 2, we can see that ILARL and REIRL are again the most efficient algorithms in terms of MDP trajectories and they are able to match the performance of behavioural cloning despite the fact it has better guarantees for the case of deterministic experts. For Figure 1, we used $\sigma = 0.1$.

I.2 Environment description

We run the experiment in the following MDP with continuous states space. We consider a 2D environment, where we denote the horizontal coordinate as $x \in [-1, 1]$ and vertical one as $y \in [-1, 1]$. The agent starts in the upper left corner, i.e., the coordinate $[-1, 1]^\intercal$ and should learn to reach the opposite corner (i.e. $[1, -1]^\intercal$) while avoiding the central high cost area depicted in Figure 3. The reward function is given by: $\mathbf{c}_{\text{true}}(s,a) = \mathbf{c}_{\text{true}}([x,y]^\intercal, a) = (x-1)^2 + (y+1)^2 + 80 \cdot$



(a) $\sigma = 0$ (b) $\sigma = 0.05$ (c) $\sigma = 0.1$

Figure 2: Experiments on the continuous gridworld with one trajectory from a deterministic expert.

$e^{-8(x^2+y^2)} - 100 \cdot \mathbb{1}\{x \in [0.95, 1], y \in [-1, -0.95]\}$. The action space for the agent is given by $\mathcal{A} = \left\{ \underbrace{[0.01, 0]^\intercal}_{\triangleq A_1}, \underbrace{[0, 0.01]^\intercal}_{\triangleq A_2}, \underbrace{[-0.01, 0]^\intercal}_{\triangleq A_3}, \underbrace{[0, -0.01]^\intercal}_{\triangleq A_4} \right\}$, and the transition dynamics are given by:

$$s_{t+1} = \begin{cases} \Pi_{[-1,1]^2}\left[s_t + \frac{a_t}{10}\right] & \text{w.p.} \quad 1-\sigma \\ \Pi_{[-1,1]^2}\left[s_t - \frac{s_t}{10\|s_t\|_2}\right] & \text{w.p.} \quad \sigma \end{cases}$$

Thus, with probability σ, the environment does not respond to the action taken by the agent, but

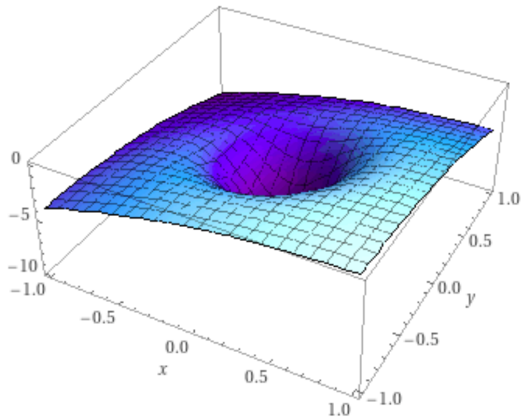


Figure 3: Graphical representation of $-\mathbf{c}_{\text{true}}$ of the linear MDP used in Figures 1 and 2.

it takes a step towards the low reward area centered at the origin, i.e., $-\frac{s_t}{10\|s_t\|_2}$. The agent should therefore pass far enough from the origin. Consider

$$\phi(s, a) = \phi([x, y], a) = \left[x^2, y^2, x, y, e^{-8(x^2+y^2)}, \mathbb{1}\left\{x \in [0.95, 1], y \in [-1, -0.95]\right\}, \mathbf{e}_a^\intercal\right]$$

with

$$\mathbf{e}_a = \left[\mathbb{1}\left\{a = A_1\right\}, \mathbb{1}\left\{a = A_2\right\}, \mathbb{1}\left\{a = A_3\right\}, \mathbb{1}\left\{a = A_4\right\}\right]^\intercal.$$

Notice that Assumption 2 holds only for the cost $\mathbf{c}_{\text{true}} = \mathbf{\Phi}[1, 1, -2, -2, 80, -100, 2, 2, 2, 2]^\intercal$ while for the dynamics the linearity assumption does not hold.

I.3 Numerical verification of the finite horizon improvement.

We test BRIG (Algorithm 4) in a toy finite horizon problem. In particular, we consider a linear bandits problem ($H = 1$) with true cost function $\mathbf{c}_{\text{true}} = \mathbf{\Phi}\mathbf{w}_{\text{true}}$ where $\mathbf{\Phi}$ entries are sampled from a normal

distribution. For $\mathbf{w}_{\text{true}}$ we choose $\mathbf{w}_{\text{true}}(i) = 0$ if i is odd and $\mathbf{w}_{\text{true}}(i) = 1$ otherwise. We generate the expert dataset sampling 10 actions from a softmax expert. The results are shown in Figure 4. They confirm the theoretical findings that BRIG outperforms ILARL for finite horizon problems in terms of MDP trajectories.

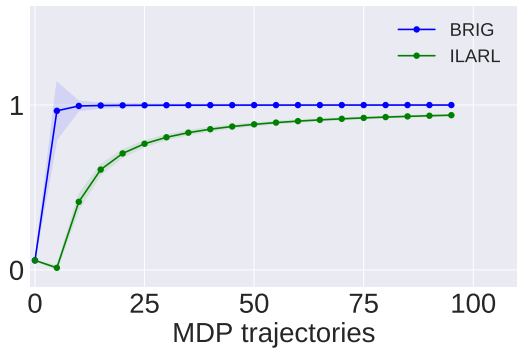


Figure 4: Experiment in finite horizon setting to assess the better efficiency of BRIG.

I.4 Hyperparameters

For the experiments in Figures 1 and 2 we used $\eta = 1$, $\tau = 5$ and $\beta = 8$. For IQlearn, we also collect 5 trajectory to perform each update on the Q-function, and we use again $\eta = 1$ and 0.005 as stepsize for the Q-function weights. For PPIL, we use batches of 5 trajectories, 20 gradient updates between each batch collection, $\eta = 1$ and and 0.005 as stepsize for the Q-function weights. For GAIL and AIRL, we use the default hyperparameters in `https://github.com/Khrylx/PyTorch-RL` but we obtained a better prerformance with a larger batch size of 6144 states and we use linear models rather than neural networks. For REIRL, we used the implementation in [47] but again we increased the batch size equal to 6144 states for achieving a better performance.

J Reducing the number of expert trajectories.

In this section, we show that the number of required expert trajectories can be further reduced at the price of additional assumption on the expert policy, features and expert occupancy measure. The estimator we use is build on the ideas underling Mimic-MD in the linear case [35].

Remark 4. *Using such an estimator in ILARL or BRIG allows to improve upon Mimic-MD in two ways. Indeed ILARL and BRIG are provably computationally efficient algorithms and do not require knowledge of the dynamics. On the other hand, Mimic-MD requires perfect knowledge of the transition dynamics and it is unclear if the output policy can be computed efficiently in Linear MDPs.*

To this goal, we need to consider the following estimator for $\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}$, where we denote via (s_h^τ, a_h^τ) the state action pair encountered at step h in the trajectory τ

$$\widetilde{\mathbf{\Phi}^\intercal d^{\pi_{\text{E}}}} = (1-\gamma)\mathbb{E}_{\tau\sim\pi_{\text{E}}}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau, a_h^\tau)\mathbb{1}\left[s_t \in \mathcal{K} \quad \forall s_t \in \tau\right]\right] + (1-\gamma)\mathbb{E}_{\tau\sim\text{Unif}(D_1)}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau, a_h^\tau)\mathbb{1}\left[\exists s_h \in \tau \; s.t. \; s_h \notin \mathcal{K}\right]\right] \tag{9}$$

where we split the expert dataset $\mathcal{D}_{\pi_{\text{E}}}$ in two disjoint halves D_0, D_1. The first D_0 is used to compute the set $\mathcal{K}$ which is according to [35, Definition 7] the set where the policy π_{BC} learned via Behavioural Cloning on the input dataset D_0 equals the expert policy. That is, $\mathcal{K} = \{s \in \mathcal{S} \;\; s.t. \;\; \pi_{BC}(s) = \pi_{\text{E}}(s)\}$[4]. The other half denoted via D_1 is used for the second

[4]Notice that we consider a deterministic expert in this section as done in [35]. Therefore, we consider policies as mapping from states to actions, i.e. $\pi : \mathcal{S} \to \mathcal{A}$

term in 9. In the analysis of [35] the first term can be computed thanks to the perfect knowledge of the dynamics. In our case, we have only trajectory access so we use the estimator

$$\underline{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}} = (1-\gamma)\mathbb{E}_{\tau\sim\mathrm{Unif}(\mathcal{D}_{\pi_{BC}})}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]\right]$$
$$+(1-\gamma)\mathbb{E}_{\tau\sim\mathrm{Unif}(D_1)}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[\exists s_h\in\tau\quad s.t.\quad s_h\notin\mathcal{K}\right]\right] \tag{10}$$

where the dataset $\mathcal{D}_{\pi_{BC}}$ contains trajectories sampled according to π_{BC}.

Lemma 11. *Let us consider the estimator $\underline{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}}$ with the set $\mathcal{K}$ be the confidence set for a binary linear classifier as defined in [35, Section 4.1], let the expert policy be deterministic ans satisfy the Linear Expert Assumption [35, Definition 4].Moreover consider features that satisfy $-\phi(s,1)=\phi(s,0)=s/2$ for all $s\in\mathcal{S}$ where the state space is chosen to be $\mathbb{R}^d$. Finally, let consider that $\sum_{a\in\mathcal{A}}d^{\pi_{\mathrm{E}}}(\cdot,a)$ is the uniform distribution $\mathrm{Unif}(\mathcal{S})$, then it holds that for any $\delta>0$*

$$\mathbb{E}\left\|\underline{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}}-\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}\right\|_\infty \le \frac{1}{1-\gamma}\sqrt{\frac{\log(d/\delta)}{2\left|\mathcal{D}_{\pi_{BC}}\right|}}+\frac{\delta}{1-\gamma}+\mathcal{O}\left(\frac{d^{5/4}\log d}{(1-\gamma)^{3/2}\left|\mathcal{D}_{\pi_{\mathrm{E}}}\right|}\right) \tag{11}$$

Remark 5. *The Lemma above follows the construction in [35] to show that there exists one example under which ILARL used with estimator $\underline{\mathbf{\Phi}^\intercal d^{\pi_{\mathrm{E}}}}$ requires only $\widetilde{\mathcal{O}}(d^{5/4}\epsilon^{-1}(1-\gamma)^{-3/2})$ expert trajectories. However, it remains open to prove that the same holds true for general expert in Linear MDPs without further assumptions on the features and expert occupancy measure.*

Proof. The error can be controlled as follow

$$E\triangleq\mathbb{E}_{\tau\sim\pi_{\mathrm{E}}}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]\right]-\mathbb{E}_{\tau\sim\mathrm{Unif}(\mathcal{D}_{\pi_{BC}})}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]\right]$$

so denoting $X(\tau)\triangleq\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]$ and noticing that by definition of $\mathcal{K}$ we have that

$$\mathbb{E}_{\tau\sim\mathrm{Unif}(\mathcal{D}_{\pi_{BC}})}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]\right]=\mathbb{E}_{\tau\sim\mathrm{Unif}(\mathcal{D}_{\pi_{\mathrm{E}}})}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[s_t\in\mathcal{K}\quad\forall s_t\in\tau\right]\right],$$

we can rewrite E as a martingale difference sequence

$$\begin{aligned}\|E\|_\infty &= \left\|\mathbb{E}_{\tau\sim\pi_{\mathrm{E}}}\left[X(\tau)\right]-\frac{1}{\left|\mathcal{D}_{\pi_{BC}}\right|}\sum_{\tau\in\mathcal{D}_{\pi_{BC}}}X(\tau)\right\|_\infty\\ &= \left\|\frac{1}{\left|\mathcal{D}_{\pi_{BC}}\right|}\sum_{\tau\in\mathcal{D}_{\pi_{BC}}}\left(X(\tau)-\mathbb{E}_{\tau\sim\pi_{\mathrm{E}}}\left[X(\tau)\right]\right)\right\|_\infty\\ &\le \frac{1}{1-\gamma}\sqrt{\frac{\log(d/\delta)}{2\left|\mathcal{D}_{\pi_{BC}}\right|}}\quad\text{w.p.}\quad 1-\delta\end{aligned}$$

Therefore choosing $|\mathcal{D}_{\pi_{BC}}|=\frac{\log(d/\delta)}{2\epsilon^2(1-\gamma)^2}$ ensures $E\le\epsilon$ with probability at least $1-\delta$. Therefore by Lemma 7,

$$\mathbb{E}\left\|E\right\|_\infty\le\frac{1}{1-\gamma}\sqrt{\frac{\log(d/\delta)}{2\left|\mathcal{D}_{\pi_{BC}}\right|}}+\frac{\delta}{1-\gamma}$$

These trajectories can be simulated in the MDP therefore the latter it is not a requirement on the expert dataset size. The number of expert trajectories is crucial to control the error due to the trajectories

containing trajectories not in $\mathcal{K}$, i.e. Equation (10). Denoting this error as E_2 we have

$$
\begin{aligned}
&\mathbb{E}_{D_0,D_1} \|E_2\|_\infty \\
&= \mathbb{E}_{D_0,D_1}\Bigg[\Bigg|\Bigg|\mathbb{E}_{\tau\sim\pi_{\mathrm{E}}}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[\exists s_h\in\tau\ s.t.\ s_h\notin\mathcal{K}\right]\right] \\
&\quad - \mathbb{E}_{\tau\sim\mathrm{Unif}(D_1)}\left[\sum_{h=1}^{\infty}\gamma^h\phi(s_h^\tau,a_h^\tau)\mathbb{1}\left[\exists s_h\in\tau\ s.t.\ s_h\notin\mathcal{K}\right]\right]\Bigg|\Bigg|_\infty\Bigg] \\
&\leq \frac{1}{(1-\gamma)}\sqrt{\frac{d}{|D_1|}\mathbb{E}_{\tau\sim D_0}\left[\mathbb{1}\left[\exists s_h\in\tau\ s.t.\ s_h\notin\mathcal{K}\right]\right]} \\
&= \frac{1}{(1-\gamma)}\sqrt{\frac{d}{|D_1|}\mathbb{E}_{\mathrm{lenght}(\tau)}\left[\mathbb{E}_{\tau\sim D_0|\mathrm{lenght}(\tau)}\left[\mathbb{1}\left[\exists s_h\in\tau\ s.t.\ s_h\notin\mathcal{K}\right]\right]\right]} \quad \text{Tower Property of Expectation} \\
&\leq \frac{1}{(1-\gamma)}\sqrt{\frac{d}{|D_1|}\mathbb{E}_{\mathrm{lenght}(\tau)}\left[\sum_{h=1}^{\mathrm{lenght}(\tau)}\mathbb{E}_{\tau\sim D_0|\mathrm{lenght}(\tau)}\left[\mathbb{1}\left[s_h\notin\mathcal{K}\right]\right]\right]} \quad \text{Union Bound} \\
&\leq \frac{1}{(1-\gamma)}\sqrt{\frac{d}{|D_1|}\mathbb{E}_{\mathrm{lenght}(\tau)}\left[\sum_{h=1}^{\mathrm{lenght}(\tau)}\mathcal{O}\left(\frac{d^{3/2}\log d}{|D_0|}\right)\right]} \quad \text{Thanks to [35, Theorem 7]} \\
&\leq \mathcal{O}\left(\frac{1}{(1-\gamma)}\sqrt{\frac{d^{5/2}\log d}{|D_1|^2}\mathbb{E}_{\mathrm{lenght}(\tau)}\left[\mathrm{lenght}(\tau)\right]}\right) \quad \text{Using that } |D_0|=|D_1| \text{ by construction} \\
&\leq \mathcal{O}\left(\frac{1}{(1-\gamma)}\sqrt{\frac{d^{5/2}\log d}{|D_1|^2}\frac{1}{1-\gamma}}\right) \\
&= \mathcal{O}\left(\frac{d^{5/4}\log d}{(1-\gamma)^{3/2}|D_1|}\right)
\end{aligned}
$$

Where we used [35, Theorem 7] to bound

$$\mathbb{E}_{\tau\sim D_0|\mathrm{lenght}(\tau)}\left[\mathbb{1}\left[s_h\notin\mathcal{K}\right]\right] \leq \mathcal{O}\left(\frac{d^{3/2}\log d}{|D_0|}\right)$$

so overall

$$\mathbb{E}_{D_0,D_1}\|E_2\|_\infty \leq \mathcal{O}\left(\frac{d^{5/4}\log d}{(1-\gamma)^{3/2}|D_1|}\right)$$

□