ILeNA: ϵ^{-2} Sample Complexity for Value-Based Imitation Learning

Anonymous authors
Paper under double-blind review

Abstract

Value-based imitation learning is a new approach to eliminate assumptions about the policy class. We study offline imitation learning where expert realizability is replaced by Q^π-realizability. The objective is to find a policy with a return at least as good as the expert's. Our proposed algorithm, ILeNA, obtains a sample complexity in $O(\frac{\log \mathcal{N}(\mathcal{Q})}{(1-\gamma)^4\epsilon^2})$ under general Q^π-realizability, and $\tilde{O}(\frac{d}{(1-\gamma)^4\epsilon^2})$ under linear Q^π-realizability. This solves the open problem of obtaining sample complexity in the order of ϵ^{-2} under non-convex action-value function sets. The best previous sample complexity in the general non-convex case was in $O(\epsilon^{-4})$, which is improved by a factor of $\frac{1}{(1-\gamma)^4\epsilon^2}$ in this work. We reduce the problem to a two-player repeated game between an actor (π-player) and a critic (Q-player), and translate any no-regret dynamic between the players into a guarantee for the value of the output policy. The generality of this result enables us to use an appropriate pair of no-regret algorithms for the two players, which gives us the aforementioned sample complexity. We also support our theoretical result with numerical experiments.

Imitation learning (also known as *apprenticeship learning* and *learning from demonstration*) is a sub-domain of reinforcement learning where the objective is to find a policy that mimics an expert policy. The offline version of imitation learning (IL), which is relatively close to the supervised learning setting, addresses the case where a batch of data from the expert policy's behaviour is provided for the learner. There is a wide variety of applications for IL, but it can especially be a reasonable approach when the reward is defined based on a desired way of acting in the environment, such as in robotics (Celemin et al., 2022; Schaal, 1999; Singh et al., 2020), LLMs fine-tuning, and reasoning (Wulfmeier et al., 2024; Cundy & Ermon, 2023; Wu et al., 2021; Yang et al., 2022; Sun, 2024).

Typically, there are two main types of objectives in imitation learning. First, the output policy can imitate the unknown expert policy in the distribution over actions it takes. Second, the aim is to propose a policy whose value is at least as high as the expert's value. The former objective is stronger than the latter; if the output policy is close to the expert at the level of actions, then it also obtains a value close to that of the expert. However, the latter, and the more reward-oriented objective, is beneficially less restrictive. For example, if the algorithm happens to output the optimal policy, the second objective will be satisfied, even if the distributions over the actions in the optimal policy are not close to those of the expert, violating the first objective. In this work, we consider the second value-based objective.

In earlier work, this problem was studied under structural assumptions such as known transition kernels (or simulator access) and linear reward functions (Abbeel & Ng, 2004; Syed & Schapire, 2007). A more recent line of work replaces these assumptions with expert realizability, i.e., knowing a set of policies that contains the expert policy (Ross & Bagnell, 2010; Rohatgi et al., 2025; Foster et al., 2024; Rajaraman et al., 2021).

In many tasks, the variety of possible returns is much less than the number of different ways of taking actions. For example, in a large environment with sparse rewards, although there are many states and different ways to act, the final possible cumulative returns are limited (refer to Figure 7). This is the case in reasoning tasks, where we are interested in achieving a sound proof (sparse reward of 1 at the end). This observation is in favor of putting assumptions on different possible rewards that can be achieved (Q^π-realizability) rather than different possible ways to act in an environment (expert realizability). This underlying logic is formalized by Joshi et al. (2025), for the case of contextual

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

bandit. Also you can refer to Appendix D.2 for an example supporting this argument, in which there is an exponential difference between the size of the policy class, $\log \mathcal{N}(\Pi, \epsilon)$, and the size of the possible action value function, $\log \mathcal{N}(\mathcal{Q}, \epsilon)$. In the spirit of this logic, we study the problem of imitation learning under the Q^π-realizability assumption.

Assumption 0.1. *(Q^π Realizability) An MDP is said to be Q^π -realizable if there exists a function class $\mathcal{Q} \subset \mathbb{R}^{S \times A}$ such that for any policy π, it holds that $Q^\pi \in \mathcal{Q}$, and for any $Q \in \mathcal{Q}$, $\|Q\|_\infty \leq \frac{1}{1-\gamma}$.*

In some environments, although there might be many states or actions (possibly infinitely many), there exists an underlying structure that allows us to represent the involved elements in a feature space. In this work, in addition to the general case, we consider linear-function approximation for the set of action-value functions.

Assumption 0.2. *(Linear-Q^π Realizability) An MDP is said to be linearly Q^π -realizable if there exists a known mapping $\phi : \mathcal{S} \times A \to \mathbb{R}^d$, such that for any policy π, there is a vector $\theta^\pi \in \mathbb{R}^d$, such that $Q^\pi(s,a) = \langle \phi(s,a), \theta^\pi \rangle$. In addition, we assume, $\|\theta^\pi\| \leq B_\theta$ for all π, and $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\| \leq B_\phi$.*

Main Question

How many samples from the behaviour of an expert policy (in the form of state-action) is needed to learn an imitating policy, only under the Q^π realizability assumption?

By *only*, we mean that no further assumptions are made: neither knowledge of the transition kernel or the reward, nor expert realizability, nor technical assumptions such as convexity of $\mathcal{Q}$.

One might imagine that the dataset gives us an empirical estimate for the expert policy distribution over the actions for each state, i.e. $\pi_E(.|s)$. So why not just proposing that empirical distribution? In the tabular case, this is what *behavioral cloning* (Pomerleau, 1988; Sammut et al., 1992) does; more generally, BC fits a policy from a class Π^E to the expert's actions by supervised learning, and it is a reasonable approach when no further assumptions are available. However, having more assumptions, e.g. Q^π-realizability, enables better approaches. Using supervised learning terminology, the assumption of expert realizability is analogous to having a realizable hypothesis class, and a data set of state-action pairs resembles a data set of feature-labels. In this case, a factor of $\log |\Pi^E|$ contributes to the sample complexity of the algorithms (Table 1). In comparison to the expert realizability assumption, Q^π realizability is not a direct assumption on the policies, which in turn makes it technically harder to exploit

Recently, Moulin et al. (2025) established sample complexity guarantees under Q^π-realizability assumptions. We will discuss their primal-dual based algorithm, SPOIL , in Section A along with other related work. Moulin et al. (2025) achieve the sample complexity in the order of $\tilde{O}(\epsilon^{-2})$ under *linear* Q^π-realizability and $\tilde{O}(\epsilon^{-4})$ for (general) Q^π-realizability assumption. Achieving a tighter sample complexity, such as in the order of $\tilde{O}(\epsilon^{-2})$ under the general Q^π-realizability assumption was left as an open problem.

Our Contribution: We propose the meta algorithm, ILENA , for offline imitation learning in the discounted infinite horizon setting, achieving the sample complexity of $O(\frac{\log \mathcal{N}(\mathcal{Q})}{(1-\gamma)^4\epsilon^2})$ for general Q^π realizability. This resolves the open problem of finding $O(\epsilon^{-2})$ sample complexity for general non-convex sets of action value functions, and improves the best previously known sample complexity by a factor of $\frac{1}{(1-\gamma)^4\epsilon^2}$ (Moulin et al., 2025). ILENA also achieves $\tilde{O}(\frac{d}{(1-\gamma)^4\epsilon^2})$ sample complexity for linear-function approximation. As a technical contribution, the meta-algorithm and its analysis are compatible with any underlying pair of no-regret algorithms for the Q-player (critic) and the π-player (actor). The incorporated ideas, particularly the pair of no-regret algorithms we employ, may be of independent interest for addressing non-convex action sets. The results are only followed by the Q^π-realizability assumption, which substitutes for the conventionally used assumption of expert policy realizability. There are no further assumptions on knowing the transition kernel or reward function. We also empirically evaluate ILENA to support the theoretical results.

We approach the problem through constructing a minimax optimization between the π-player (actor) and the Q-player (critic). The key distinguishing point of our work, compared to the previous results using primal-dual approach, including (Moulin et al., 2025; Viano et al., 2026), is that ILENA and

its analysis are compatible with any underlying pair of no-regret algorithms to solve the min-max optimization problem. We believe the underlying idea that leads us to have this freedom in choosing underlying no-regret algorithms can be transferred to other domains to improve minimax approaches. Additionally, our novel choice of no-regret algorithms can be applied to other problems with non-convex action sets. Our approach avoids working with convex hulls in the analysis—which often have significantly larger covering numbers than the original sets—as discussed further in Section 3.1.

1 Notation and Problem Setting

We consider a Markov Decision Process (MDP) M to be a tuple, $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0)$ where $\mathcal{S}$ and $\mathcal{A}$ are respectively state and action spaces, with the cardinality $|\mathcal{A}| = \boldsymbol{A}$ and $|\mathcal{S}| = \boldsymbol{S}$. The reward $R_t \in [0, 1]$ is a random reward obtained at time step t with the mean of $r(s_t, a_t)$, where $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. γ is the discount factor and ρ_0 is the initial state distribution. Given a pair of state-action, the transition kernel P shows the probability distribution which the next state belongs to, i.e. $P(.|s_t, a_t) \in \Delta_{\mathcal{S}}$, and $S_{t+1} \sim P(.|s_t, a_t)$. A memory-less policy is a mapping $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ from the state space to distributions over the action space. We gather all such policies in Π. We also use $\pi(a|s)$ to denote the probability mass that π puts on a when it acts on state s. The interaction between an agent using a policy π and an MDP M is as follows: The policy π operates on the initial state $S_0 \sim \rho_0$ and takes action $A_0 \sim \pi(S_0)$. The agent gets the reward of $R_0 = R(S_0, A_0)$. Then the next state S_1 is drawn from $P(.|S_0, A_0)$, and the interaction continues. The policy π together with the initial distribution ρ_0, transition kernel P, and random reward R, induces a probability distribution $\mathbb{P}^\pi$ over the trajectories of interactions. $\mathbb{E}^\pi$ is also its corresponding expectation operator. For any state $s \in \mathcal{S}$, the (discounted) state value function of the policy π is defined as $V^\pi(s) := \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t R(S_t, A_t)|S_0 = s]$. The state-action value function Q^π is similarly defined with an additional condition on the first action, i.e. $Q^\pi(s, a) := \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t R(S_t, A_t)|S_0 = s, A_0 = a]$. These two value functions are related to each other through Bellman equations:

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)Q^\pi(s, a), \quad \text{and} \quad Q^\pi(s, a) = r(s, a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s, a)V^\pi(s').$$

Any policy π (not necessarily memory-less), together with the initial distribution ρ_0 and transition kernel P, generates a state occupancy measure $\rho_\pi \in \Delta_{\mathcal{S}}$ over states.

$\rho_\pi(s) := (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}^\pi(S_t = s)$, denotes the expected discounted total times that the state s is occupied in the interaction between π and the MDP. The state-action occupancy measure $\mu_\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}$ is defined similarly, i.e. $\mu_\pi(s, a) := (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}^\pi(S_t = s, A_t = a)$. These two quantities are related to each other through flow conditions,

$$\rho_\pi(s) = (1 - \gamma)\rho_0(s) + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} P(s|s', a')\mu_\pi(s', a').$$

One can also verify that the value of a (potentially non-memory-less) policy is related to the occupancy measure of the policy and the reward function as follows,

$$V^\pi(\rho_0) = \frac{1}{(1 - \gamma)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_\pi(s, a)r(s, a),$$

where $V^\pi(\rho_0) = \sum_{s\in\mathcal{S}} \rho_0(s)V^\pi(s)$. Also $g^\pi := (1-\gamma)\mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t R(S_t, A_t)]$ is the expected *normalized* return of policy π. One can verify that $g^\pi = (1-\gamma)V^\pi(\rho_0) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_\pi(s, a)r(s, a)$.

Imitation Learning. We assume a data set of state-action pairs from the unknown expert policy π^E i.e. $\mathcal{D} = \{(S_i, A_i)\}_{i=1}^\tau$ where $(S_i, A_i) \sim \mu_{\pi^E}$. We also denote the set of states in the data set by $\mathcal{D}_{\mathcal{S}} := \{S_i\}_{i=1}^\tau$. Given this data set, we aim to output a policy $\bar{\pi}$ that imitates the behavior of π^E. Note that the transition kernel P and the reward function r are unknown to the learner. The goal is to achieve at least as much return as the expert policy π^E achieves. More specifically, the goal is to output a policy $\bar{\pi}$ such that with high probability of $1 - \delta$:

$$V^{\pi_E}(\rho_0) - V^{\bar{\pi}}(\rho_0) \leq \epsilon. \tag{1}$$

It is important to note that the objective is defined over the value of the policies (V^π), and not on the normalized return (g^π) which is smaller by a factor of $(1 - \gamma)$.

Notations. By $\mathcal{C}_p(\mathcal{Q}, \epsilon)$ and $\mathcal{N}_p(\mathcal{Q}, \epsilon)$, respectively, we denote the covering set and the covering number of a set $\mathcal{Q}$ under the norm $\|.\|_p$ with precision ϵ. $\boldsymbol{\delta}(a)$ is the Dirac distribution in $\Delta_{\mathcal{A}}$ which has the value of zero everywhere except at the entry a (with the value of 1).

1.1 The Loss Functions

From the *performance-difference lemma* (Kakade & Langford, 2002; Howard, 1960), we can write $g_{\pi^E} - g_{\bar{\pi}} = \mathbb{E}_{\mu_{\pi^E}}[Q^{\bar{\pi}}(S, A) - Q^{\bar{\pi}}(S, \bar{\pi}(S))]$, where (S, A) are sampled from μ_{π^E}, and $Q(S, \pi(S)) = \sum_{a \in \mathcal{A}} \pi(a|S) Q(S, a)$. We can now define a loss function $\mathcal{L}$,over the set of $\Pi \times \mathcal{Q}$ as,

$$\mathcal{L}(\pi, Q) = \mathbb{E}_{\mu_{\pi^E}}[Q(S, A) - Q(S, \pi(S))]. \tag{2}$$

The objective is now translated into finding a policy $\bar{\pi}$ such that $\mathcal{L}(\bar{\pi}, Q^{\bar{\pi}}) \leq (1 - \gamma)\epsilon$. Since μ_{π^E} is unknown we can not directly compute $\mathcal{L}$. However, having the set of samples $\mathcal{D} = \{(S_i, A_i)\}_{i=1}^{\tau}$ suggests considering the following empirical loss,

$$\hat{\mathcal{L}}(\pi, Q) = \frac{1}{\tau} \sum_{i=1}^{\tau} \big[Q(S_i, A_i) - \sum_{a \in \mathcal{A}} \pi(a|S_i) Q(S_i, a) \big]. \tag{3}$$

From this point, there are different ways to find a near-optimal pair of $(\bar{\pi}, Q^{\bar{\pi}})$. Moulin et al. (2025) directly work on $\hat{\mathcal{L}}$, making a min-max problem with the two players of π-player and Q-player (min and max player respectively). The analysis of their algorithm is highly reliant on the Q-player being a best responder. In contrast, we define an online learning algorithm that allows both players to select any no-regret algorithm.

2 Towards the ϵ^{-2} Sample Complexity

In this section, we explain the main ideas we deploy to have a sample complexity of order $O(\epsilon^{-2})$, which also helps to obtain the logic of algorithm ILeNA in Section 3, and its analysis in Section 4. To give a summary, based on the aforementioned empirical loss function, we construct τ many two-player repeated games, such that running no-regret algorithms in the games leads us to a desired output $\bar{\pi}$. We use variants of best responding for the π player. The simplicity of best responding enables us to control the set of policies that can be played, and in turn having a tighter concentration inequalities. We will discuss this in details in Section 3.1. To prevent potential confusion, it is worth noting that this is an online approach, inside the learner, to solve an offline imitation learning problem.

2.1 The Game Between Actor (π-player) and Critic (Q-player)

We propose an online learning problem, which has a special form of cost function. Let

$$\hat{c}_n(\pi, Q) := \hat{\mathcal{L}}(\pi, Q_n) - \hat{\mathcal{L}}(\pi_n, Q), \tag{4}$$

where (π_n, Q_n) is the tuple played by the learner at time step n. $\hat{c}$ is designed in a way that having a no-regret algorithm controls the duality gap defined on $\hat{\mathcal{L}}$ (Cheng et al., 2020). The regret of the online learning problem over N iterations is defined as,

$$\mathsf{Reg} := \frac{1}{N} \big[\sum_{n=1}^{N} \hat{c}_n(\pi_n, Q_n) - \min_{\pi \in \Pi, Q \in \mathcal{Q}} \sum_{n=1}^{N} \hat{c}_n(\pi, Q) \big]. \tag{5}$$

Note that $\hat{c}_n(\pi_n, Q_n) = 0$ which makes the first term zero. The second term in the regret is related to the duality gap. Let us inspect the cost function $\hat{c}_n(\pi, Q)$ by expanding Definition 4,

$$\frac{1}{\tau} \sum_{i=1}^{\tau} \big[Q_n(S_i, A_i) - \sum_{a \in \mathcal{A}} \pi(a|S_i) Q_n(S_i, a) - Q(S_i, A_i) + \sum_{a \in \mathcal{A}} \pi_n(a|S_i) Q(S_i, a) \big]$$

The first term $Q_n(S_i, A_i)$ only makes an offset on the cost function $\hat{c}_n$, so we can eliminate it from both terms, $\hat{c}_n(\pi_n, Q_n)$ and $\hat{c}_n(\pi, Q)$, and define $\hat{h}_n(.,.) := \hat{c}_n(.,.) - \frac{1}{\tau} \sum_{i=1}^{\tau} Q_n(S_i, A_i)$. The regrets induced by $\hat{h}_n$ and $\hat{c}_n$ are equal. Now we can write

$$\hat{h}_n(\pi, Q) = \frac{1}{\tau} \sum_{i=1}^{\tau} \big[\langle \pi_n(.|S_i) - \boldsymbol{\delta}(A_i), Q(S_i, .) \rangle + \langle \pi(.|S_i), -Q_n(S_i, .) \rangle \big],$$

where the inner product is in $\mathbb{R}^{\boldsymbol{A}}$. Note that the new cost function $\hat{h}_n$ has been decomposed into two parts, i.e. $\hat{h}_n(\pi, Q) = h_n^{\mathcal{Q}}(Q) + \frac{1}{\tau}\sum_{i=1}^{\tau} h_{n,i}^{\Pi}(\pi)$ where,

$$h_n^{\mathcal{Q}}(Q) := \frac{1}{\tau}\sum_{i=1}^{\tau}\langle \pi_n(.|S_i) - \boldsymbol{\delta}(A_i), Q(S_i, .)\rangle, \quad \text{and} \quad h_{n,i}^{\Pi}(\pi) := \langle \pi(.|S_i), -Q_n(S_i, .)\rangle. \tag{6}$$

This leads us to the following regret decomposition, which will be notably useful.

Proposition 2.1. *The regret of the online learning problem is upper-bounded by the sum of Q-player's and π-player's regret,*

$$\mathsf{Reg} \leq \mathsf{Reg}^{\mathcal{Q}} + \frac{1}{\tau}\sum_{i=1}^{\tau}\mathsf{Reg}_i^{\Pi} \tag{7}$$

where

$$\mathsf{Reg}^{\mathcal{Q}} := \frac{1}{N}\max_{Q\in\mathcal{Q}}\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)], \quad \textit{and} \quad \mathsf{Reg}_i^{\Pi} := \frac{1}{N}\max_{\pi\in\Pi}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)].$$

The proof can be found in Appendix B.7. Similar regret decomposition holds for the linear Q^π realizability (Refer to Appendix). From this regret decomposition we can see that a low Reg value, which in turn implies a small duality gap and achieving the desired policy, can be obtained by deploying *any* no-regret learning algorithms for the Q-player and π-player that control Reg^Q and Reg_i^{Π}. This freedom, which is not available in many other works, including Moulin et al. (2025); Swamy et al. (2021), helps notably in the analysis.

2.2 From Regret to The Imitating Policy

In the previous section, following Proposition 2.1, it became clear that having low regret for both players implies a small value of Reg. In this and the following section, we discuss: How does a small value of Reg give us a desirable policy? Here is the core proposition that leads us to the main theorems. This is a general result that apply to any no-regret dynamic between the actor and critique.

Proposition 2.2. *(Core Proposition) Given a data set $\mathcal{D}$ of size τ, any dynamic in the game between Q and π player with $\mathsf{Reg} \leq \epsilon$, and a covering over $\mathcal{Q}$ with precision ϵ', with probability of $1-\delta$ we have,*

$$V^{\pi_E} - V^{\bar{\pi}} \leq \underbrace{\frac{2\epsilon}{(1-\gamma)}}_{\textit{Opt. Gap}} + \underbrace{\frac{8\epsilon'}{(1-\gamma)^2} + \sqrt{\frac{2\log(2\mathcal{N}_\infty(\mathcal{Q},\epsilon')/\delta)}{\tau(1-\gamma)^4}} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q},\epsilon')/\delta)}{\tau(1-\gamma)^4}}}_{\textit{Statistical Gap}},$$

Where $\tilde{\Pi} \subseteq \Pi$ is a subset of policies which covers any policy that can be played by the π player union with $\{\pi^E\}$, and $\bar{\pi}(.|s) = \frac{1}{N}\sum_{n=1}^{N}\pi_n(.|s)$ for all $s \in \mathcal{S}$, is the state-wise mixture policy.

Proof Sketch: Here we present the main steps in the proof:

$$\begin{aligned} g_{\pi^E} - g_{\bar{\pi}} = \mathcal{L}(\bar{\pi}, Q^{\bar{\pi}}) &\leq \hat{\mathcal{L}}(\bar{\pi}, Q^{\bar{\pi}}) + \Delta(\bar{\pi}, Q^{\bar{\pi}}) \leq \sup_{Q\in\mathcal{Q}}\hat{\mathcal{L}}(\bar{\pi}, Q) + \Delta(\bar{\pi}, Q^{\bar{\pi}}) \\ &\leq \hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) + \epsilon + \Delta(\bar{\pi}, Q^{\bar{\pi}}) \leq \frac{1}{N}\sum_{n=1}^{N}\Delta(\pi^E, Q_n) + \epsilon + \epsilon + \Delta(\bar{\pi}, Q^{\bar{\pi}}) \\ &\leq \frac{1}{N}\sum_{n=1}^{N}\Delta(\pi^E, Q_n) + \epsilon + \epsilon + \frac{1}{N}\sum_{n=1}^{N}\Delta(\pi_n, Q^{\bar{\pi}}) \end{aligned}$$

The first inequality holds due to the definition of $\Delta(\pi, Q) := |\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)|$, the third is from the result of Lemma B.1. The fourth inequality holds due to Lemma B.2, and the last one

is from the triangle inequality and linearity of $\mathcal{L}$ and $\hat{\mathcal{L}}$ in π. Finally, we need to replace the Δ terms with the results of Lemma B.3. The third inequality and Lemma B.1, which is the key lemma, demonstrate how finding the saddle point solution of $\hat{\mathcal{L}}$ through a pair of no-regret algorithms helps us in controlling our objective.

There is a similar proposition to this in (Moulin et al., 2025) (Proposition 1). The proof in their work is reliant on the Q-player being best responding, which imposes some limitation on the applicability of that proposition. However, Proposition 2.2 is more general since it is compatible with any pair of no-regret algorithms for both Q and π players. Also, it is important to note that no technical assumptions, such as convexity of $\mathcal{Q}$, are used in the proof.

Proposition 2.2 shows how the regret upper bound (ϵ), and the statistical error (Δ) contribute to control our objective, $V^{\pi_E} - V^{\bar{\pi}}$. Δ decreases by having more data and finer covering over $\mathcal{Q}$, and ϵ decreases by spending more iterations in the online learning problem (equivalently, the two player repeated games). Here we discuss how to control the two gaps mentioned in Proposition 2.2:

Statistical Gap. The statistical gap is the result of differences between $\mathcal{L}$ and empirical loss $\hat{\mathcal{L}}$ in the policies and the Q functions played by the two players. This gap can be controlled by having enough data (large enough τ). It is important to note the two metric entropies in the numerator of the fractions. $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ is the covering number of $\tilde{\Pi} \times \mathcal{Q}$ under the norm of $\|(\pi, Q)\|_{\infty,1,\infty} := \max_{s \in \mathcal{S}} \|\pi(.|s)\|_1 + \|Q\|_\infty$. Since $\tilde{\Pi}$ is the set of policies that can be played by the π-player's algorithm (refer to 13 for formal definition of $\tilde{\Pi}$), this motivates us to use simple algorithms for the π-player to have a controlled and restricted set of $\tilde{\Pi}$ with a small covering number (refer to Remark 4.3).

Optimization Gap. This gap is controlled by the underlying two no-regret algorithms of the Q and π player, and is associated to their convergence rate. Using typical no-regret algorithms for N iterations, ϵ will be in the order of $\frac{1}{\sqrt{N}}$. We are going to have other convergence rates of $\frac{1}{N^{1/3}}$ and $\frac{1}{N}$ for different situations.

3 ALGORITHM: ILENA

In this section we propose our algorithm and discuss how the choice of no-regret algorithms for the two players contributes to the performance of the algorithm. We also specify the pair of no-regret algorithms and the ordering of players in different settings. Our (meta-)algorithm ILENA (**I**mitation **Le**arning with **N**o-regret **A**lgorithms) admits two arbitrary no-regret algorithms for the π-player and Q-player, and uses them in the underlying two-player games with the aforementioned designed utilities (refer to 6). The pseudocode of ILENA is on Algorithm 1.

Algorithm 1: ILeNA

Input: $\mathcal{D} = \{(S_i, A_i)\}_{i=1}^{\tau}$, $\mathcal{Q}$, N, Alg^Q, Alg^π

for $n \in \{1, 2, \ldots, N\}$ **do**
- Q-player takes Q_n suggested by Alg^Q ▷ Hedge (Mirror Descent for linear case)
- **for** $i \in \{1, 2, \ldots, \tau\}$ **do**
  - π-player takes $\pi_n(.|S_i)$ suggested by Alg_i^π ▷ Regularized Delayed Best Response
- $h_n^{\mathcal{Q}}(.)$ is revealed to the Q-player, and $h_{n,i}^{\Pi}(.)$ to the π-player ▷ Full info. feedback

Result $\bar{\pi}(.|s) = \frac{1}{N} \sum_{n=1}^{N} \pi_n(.|s)$ *for all* $s \in \mathcal{S}$

The Q-player first chooses an action Q_n (or θ_n for linear Q^π realizability case) from its action set. Then the π-player, corresponding to every data point of (S_i, A_i), responds with π_n. The Q-player receives the utilities of $h_n^{\mathcal{Q}}(.)$, and similarly the π player receives $h_{n,i}^{\Pi}(.)$ for all $i \in \tau$. They interact for N iterations. Since this is a procedure inside the learner, we can assume both players know the utilities of the games, and thus the interaction is with full information feedback.

Here we specify the underlying no-regret algorithms under the two assumptions of (i) general Q^π-realizability, and (ii) linear Q^π-realizability.

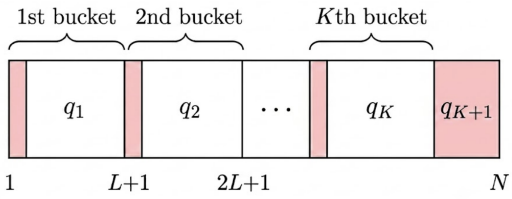


Figure 1: The lazy hedge algorithm runs over K buckets of length L. By optimizing these two values we obtain $K \in \Theta(N^{\frac{2}{3}})$ and a $L \in \Theta(N^{\frac{1}{3}})$. Since the algorithm does not change its action within a bucket it benefits from a switching cost of K.

3.1 No-regret algorithms under Q^π-realizability

The main factor that should be controlled through a proper choice of no-regret algorithms is the term $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ in the statistical gap. It will be insightful if we first show why the analysis of SPOIL (Moulin et al., 2025) does not achieve the sample complexity of $O(\epsilon^{-2})$. Moulin et al. (2025) use typical adversarial no regret algorithms (e.g. FTRL) for the π player. Since the optimization in these algorithms depends on the previous loss vectors (i.e. previous actions of Q-player), Moulin et al. (2025) work with the upper bound of $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta A})^{N+1}$, which is exponential in the number of iterations between the two players, N. We also know the convergence rate of these adversarial no-regret algorithms are in $O(\frac{1}{\sqrt{N}})$. Leaving aside whether their upper bound on $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ is tight or not, Moulin et al. (2025) control the gap using a similar result to Proposition 2.2, $g_{\pi^E} - g_{\bar{\pi}} \leq \tilde{O}(\frac{1}{\sqrt{N}}) + \tilde{O}(\sqrt{\frac{N \log(\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta A}))}{\tau}})$. This means that to have a gap in the order of ϵ, they need to choose $N \in O(\epsilon^{-2})$, which in turn results in the sample complexity of $\tau \in O(\epsilon^{-4})$. Their sub-optimal sample complexity originates from the factor of N before the $\log$. In our work, thanks to Proposition 2.2, and having freedom in choosing any no-regret algorithm we eliminate that additional factor of N, and achieve the sample complexity of $\tau \in \tilde{O}(\epsilon^{-2})$.

To have an upper bound on $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ not being exponential in N, we need to have simple algorithm for the π-player, such as best responding. In contrast to FTRL, best responding only depends on the single action of Q-player, which prevents us having the exponent of N in upper bounding $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$. To connect the covering number $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ to $\mathcal{N}_\infty(\mathcal{Q}, \epsilon)$ we also need to have a Lipschitz behavior for the π-player with respect to the input action of the Q-player. Thus we also add a regularization. The π-player's algorithm is *Regularized Delayed Best Responding (RDBR)*, with the following optimization at iteration $n \in [N]$,

$$\pi_n(.|s) = \arg\min_{\pi \in \Delta_A} \eta \langle \pi, -Q_{n-1}(s, .) \rangle + \frac{1}{2}\|\pi\|_2^2. \tag{8}$$

Note that it is a delayed best responding to the previous action of Q_{n-1}. The reason for this is that the π-player needs to best respond to the actual *realized* action of Q-player, and not its mixed strategy. Otherwise, the set $\tilde{\Pi}$ would be connected to the convex hull $Conv(\mathcal{Q})$ instead of $\mathcal{Q}$, which would blow up the covering number and is not desirable. On the other hand, we know if the π-player best responds to the realized action Q_n (and not the mixed strategy), the Q-player would not be able to have a sublinear guarantee. The novel choice of best responding to the previous realized action Q_{n-1} solves this issue. First of all, it prevents us to control $\tilde{\Pi}$ with the convex hull of $\mathcal{Q}$. Secondly, if the Q-players algorithm enjoys a sublinear switching cost, replacing Q_n with Q_{n-1} in the π-player's best responding, does not break its no-regret guarantee.

This is where the choice of Q-player's algorithm will be clear. Over the covering set of $\mathcal{C}_\infty(\mathcal{Q}, \epsilon)$ which is treated as the set of experts, the Q-player deploys a lazy version of Hedge (a.k.a. multiplicative weight update) algorithm. In Lazy Hedge, the number of times that the Q-player changes his actions is in $O(N^{\frac{2}{3}})$. As mentioned above, this sublinear switching cost is crucial for the π-player to have a sublinear regret. The pseudo code for the lazy Hedge algorithm is in Algorithm 3 and is described in Figure 1.

Overall, with this pair of algorithms, i.e. Regularized Delayed Best Responding (RDBR) for the π-player, and Lazy Hedge for the Q-player, we can have convergence rate of $O(N^{-\frac{1}{3}})$, and more importantly we obtain the upper bound of $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta A})^2$. This means that

$$g_{\pi^E} - g_{\bar{\pi}} \leq \underbrace{\tilde{O}(\frac{1}{N^{\frac{1}{3}}})}_{\text{Optimization Gap}} + \underbrace{\tilde{O}(\sqrt{\frac{\log(\mathcal{N}_\infty(\mathcal{Q}, \epsilon))}{\tau}})}_{\text{Statistical Gap}}. \tag{9}$$

which results in the sample complexity of $\tau \in O(\epsilon^{-2})$. This improvement in the sample complexity comes with a slower convergence rate, i.e. N needs to be in $O(\epsilon^{-3})$.

3.2 No-regret algorithms under Linear Q^π-realizability

Under the linear Q^π realizability assumption, the additional structure allows us to have $\mathcal{N}_\infty(\mathcal{Q}_{lin}, \epsilon) \leq (1 + \frac{2B_\theta B_\phi}{\epsilon})^d$ and $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}_{lin}, \epsilon) \leq (1 + \frac{4B_\theta B_\phi \eta A}{\epsilon})^{2d}$ (refer to Appendix B.18). In this case, the Q-player runs Mirror Descent on the convex set of θs, ball $\mathcal{B}_\theta$, and obtain a regret less than $\frac{2B_\theta B_\phi}{\sqrt{N}}$ (refer to Lemma B.15). The π player also deploys a regularized best responding to θ_n. The convexity of $\mathcal{B}_\theta$, the set of Q-players actions, is the key that allows this pair of algorithms having regrets in $O(1/\sqrt{N})$.

4 Analysis

The explanation in the previous sections brings the high-level view of the analysis. A notable portion of the proof is on the shoulders of Proposition 2.2. On top of that, we need to control the optimization gap and the statistical gap resulting from the aforementioned choice of base algorithms.

4.1 General Q^π Realizability

We first start with controlling the optimization gap, which gives us the number of rounds the underlying repeated game needs to be played. The proof of the following proposition is based on the regret decomposition (Proposition 2.1), and the two lemmas, Lemma B.8 and Lemma B.10, which are regret guarantees for each of the players (refer to Appendix B.11).

Proposition 4.1. *For any $\delta \in (0, 1)$, the dynamic with the horizon of $N \in \Omega(\frac{\ln(\mathcal{N}_\infty(\mathcal{Q}, \psi/8)\delta^{-1})}{(1-\gamma)^3\psi^3})$ induced by the two players, where the π-player deploys the regularized delayed best response (RDBR) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \frac{4}{\psi}$, and the Q-player deploys the lazy Hedge over the expert set of $\mathcal{C}_\infty(\mathcal{Q}, \psi/8)$ with bucket length of $L = \lceil (\frac{5}{2})^{\frac{2}{3}} N^{\frac{1}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \psi/8)\delta^{-1})^{\frac{-1}{3}} \rceil$ and parameter $\lambda = (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q}, \psi/8))/K}$, where $K = \lfloor (\frac{2}{5})^{\frac{2}{3}} N^{\frac{2}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \psi/8)\delta^{-1})^{\frac{1}{3}} \rfloor$, with probability of at least $1 - \delta$ has the regret of $\mathsf{Reg} \leq \psi$.*

Now we can state the lemma that shows how the covering number of the set of actions playable in the no-regret dynamic can be controlled. Proof of the following lemma is in Appendix B.4.

Lemma 4.2. *In the dynamic induced by the two players for the length of any $N \in \mathbb{N}$, where the π-player deploys the regularized best response (either delayed or not) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and the hyperparameter η, and the Q-player using any algorithm over $\mathcal{Q}$,*

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}}) + 1 \quad \textit{and} \quad \mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta \boldsymbol{A}})^2. \tag{10}$$

Remark 4.3. *The use of regularized best response by the π-player allows us to have an upper bound over $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$ which is not exponential in N. This is the central fact that leads us to have a tighter statistical gap, and obtain the ϵ^{-2} sample complexity bound (refer to Section 3.1).*

The main theorem for the general Q^π-realizability is as follows. The proof can be found in Appendix B.6. It is based on the result of Proposition 2.2 where the covering numbers are controlled through Lemma 4.2.

Theorem 4.4. *Given a data set of size $\tau \in \Omega(\frac{\log(\mathcal{N}_\infty(\mathcal{Q},\epsilon_\tau)/\delta)}{\epsilon^2(1-\gamma)^4})$, and running the two-player repeated game for $N \in \Omega(\frac{\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon_N)/\delta)}{(1-\gamma)^9\epsilon^3})$ iterations with the π-player deploying the regularized delayed best response (RDBR) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \frac{24}{\epsilon(1-\gamma)^2}$, and the Q-player deploying Lazy Hedge over the expert set of $\mathcal{C}_\infty(\mathcal{Q},\epsilon_N)$, with the bucket length of $L \in \Theta(N^{\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon_N)/\delta)^{\frac{-1}{3}})$ and parameter$\lambda = (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q},\frac{\epsilon(1-\gamma)^2}{48}))/K}$, where $K \in \Theta(N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon_N)/\delta)^{\frac{1}{3}})$, with probability of $1-2\delta$,*

$$V_{\pi^E} - V_{\bar{\pi}} \leq \epsilon,$$

where $\bar{\pi} = \frac{1}{N}\sum_{n=1}^N \pi_n$ is the state-wise mixture policy over the ones taken by the π-player. Also $\epsilon_\tau = \frac{\epsilon^2(1-\gamma)^4}{2304\boldsymbol{A}}$ and $\epsilon_N = \frac{\epsilon(1-\gamma)^2}{48}$.

Remark 4.5. *It is worth mentioning that the covering over $\mathcal{Q}$ with the precision ϵ_N, i.e. $\mathcal{C}_\infty(\mathcal{Q},\epsilon_N)$, is needed in the algorithm, since this is the set of experts for the Lazy Hedge algorithm. However, the other covering with precision ϵ_τ only appears in the analysis.*

4.2 Linear Q^π-Realizability

As the procedure is notably shared between general Q^π realizability and linear Q^π realizability, we leave the analysis of the linear case for Appendix B.6, and only state the main results here.

Theorem 4.6. *Given a data set of size $\tau \in \tilde{\Omega}(\frac{d}{(1-\gamma)^4\epsilon^2})$, and running two player repeated game for $N \in \Omega(\frac{B_\phi^2 B_\theta^2}{(1-\gamma)^4\epsilon^2})$ iterations with the π-player deploying the regularized best response with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \sqrt{N}$, and the Q-player deploying Mirror Descent on the ball $\mathcal{B}_\theta$ with potential function $F(\theta) = \frac{1}{2}\|\theta\|_2^2$ and parameter$\lambda = \frac{B_\theta}{2B_\phi\sqrt{N}}$, with probability of $1-\delta$,*

$$V_{\pi^E} - V_{\bar{\pi}} \leq \epsilon,$$

where $\bar{\pi} = \frac{1}{N}\sum_{n=1}^N \pi_n$ is the state-wise mixture policy over the policies taken by the π-player.

5 Discussion

Experiment. The main thesis of this work is theoretical. However, we have included some experiments. In a nutshell, ILENA and SPOIL , which exploit the non-convex class, are far more robust and sample-efficient than BC and the linear-feature methods (IQLearn and ValueDICE), especially when data is scarce relative to the size of the class. With enough data, the linear-feature methods catch up. In our environment, SPOIL performs better than its ϵ^{-4} guarantee, suggesting that Moulin et al. (2025) worst-case bound might be loose. See Appendix C for extensive explanations.

Trading Samples for Computation. What if we had a lot of data!? In that case, the statistical gap is easy to control and the bottleneck becomes the number of iterations N. Since Proposition 2.2 holds for any pair of no-regret algorithms, under a convex set of $\mathcal{Q}$, both players can instead use fast-rate dynamics such as Optimistic FTRL (Soleymani et al., 2025; Daskalakis et al., 2021; Farina et al., 2022), which give $\mathsf{Reg}^{\mathcal{Q}}, \mathsf{Reg}_i^{\Pi} \in \tilde{O}(1/N)$ for all $i \in [\tau]$, and hence $\mathsf{Reg} \in \tilde{O}(1/N)$. The price is that the π-player's policy now depends on all past Q-functions, which reintroduces the exponential blow-up of the covering number discussed in Section 3.1. The resulting bound, $\tilde{O}\big(\frac{1}{N}\big) + \tilde{O}\Big(\sqrt{N\log\mathcal{N}_\infty\big(\mathcal{Q},\epsilon\big)/\tau}\Big)$, gives $N \in \tilde{O}(\epsilon^{-1})$ and $\tau \in \tilde{O}(\epsilon^{-3})$. In other words, we can trade off a factor of ϵ between these two quantities. This is again a direct result of the generality of Proposition 2.2, and being free in the choice of base algorithms. See Table 2 and Appendix D.3 for the computational complexity of ILENA .

Future Work and Further Discussion There are many future directions for this work, which are mentioned in Appendix D alongside further discussions.

5.1 AI USE STATEMENT

We used large language models (LLMs), namely Claude (Anthropic), Gemini (Google) and ChatGPT (OpenAI), to assist with writing and editing the paper, identifying and checking related work. Additionally, we used generative AI tools for experiments. We also used them as a discussion aid during the formalization of the problem and the development of the theoretical results. All content, including the problem formulation, proofs, experimental design and results, was verified by the authors, who take full responsibility for the paper.

REFERENCES

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

Carlos Celemin, Rodrigo Pérez-Dattari, Eugenio Chisari, Giovanni Franzese, Leandro de Souza Rosa, Ravi Prakash, Zlatan Ajanović, Marta Ferraz, Abhinav Valada, Jens Kober, et al. Interactive imitation learning in robotics: A survey. *Foundations and Trends® in Robotics*, 10(1-2):1–197, 2022.

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

Ching-An Cheng, Remi Tachet Combes, Byron Boots, and Geoff Gordon. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3514–3524. PMLR, 2020.

Chris Cundy and Stefano Ermon. Sequencematch: Imitation learning for autoregressive sequence modelling with backtracking. *arXiv preprint arXiv:2306.05426*, 2023.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, 35:39076–39089, 2022.

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

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34: 4028–4039, 2021.

Ronald A Howard. Dynamic programming and markov processes. 1960.

Nirmit Joshi, Gene Li, Siddharth Bhandari, Shiva Prasad Kasiviswanathan, Cong Ma, and Nathan Srebro. Learning to answer from correct demonstrations. *arXiv preprint arXiv:2510.15464*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

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

Antoine Moulin, Gergely Neu, and Luca Viano. Inverse q-learning done right: Offline imitation learning in q^π-realizable mdps. *arXiv preprint arXiv:2505.19946*, 2025.

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

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.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv:2502.12465*, 2025.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Claude Sammut, Scott Hurst, Dana Kedzier, and Donald Michie. Learning to fly. In *Machine Learning Proceedings 1992*, pp. 385–393. Elsevier, 1992.

Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6): 233–242, 1999.

Avi Singh, Eric Jang, Alexander Irpan, Daniel Kappler, Murtaza Dalal, Sergey Levinev, Mohi Khansari, and Chelsea Finn. Scalable multi-task imitation learning with autonomous improvement. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2167–2173. IEEE, 2020.

Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 518–529, 2025.

Hao Sun. Supervised fine-tuning as inverse reinforcement learning. *arXiv preprint arXiv:2403.12017*, 2024.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

Qingyang Wu, Lei Li, and Zhou Yu. Textgail: Generative adversarial imitation learning for text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14067–14075, 2021.

Markus Wulfmeier, Michael Bloesch, Nino Vieillard, Arun Ahuja, Jorg Bornschein, Sandy Huang, Artem Sokolov, Matt Barnes, Guillaume Desjardins, Alex Bewley, et al. Imitating language via scalable inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 90714–90735, 2024.

Mengjiao Sherry Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. Chain of thought imitation with procedure cloning. *Advances in Neural Information Processing Systems*, 35:36366–36381, 2022.

A Related Work

Ross & Bagnell (2010) show that a per-state error ϵ of behavioral cloning can lead to a sub-optimality of $O(\epsilon H^2)$, and Ross et al. (2011) show that interacting with the expert (DAgger) reduces it to $O(\epsilon H)$. There is plenty of work that quantifies the expert sample complexity under the expert realizability assumption, i.e. the learner knows a policy class Π^E containing π^E. Rajaraman et al. (2021) assume the expert policy is in the class of deterministic linear policies, $\Pi_{det,lin}$. Using techniques from multiclass classification in supervised learning, they establish an expert sample complexity in $\tilde{O}(H^2 d/\epsilon)$, where d is the feature dimension of policies. Foster et al. (2024) showed that finding an ϵ-optimal policy is achievable with $\tilde{O}(H \log |\Pi^E|/\epsilon)$ samples for deterministic expert policies, and with $\tilde{O}(H^2 \log |\Pi^E|/\epsilon^2)$ samples for stochastic ones. Rohatgi et al. (2025); Foster et al. (2024) also step further to study misspecified expert classes. Rohatgi et al. (2025) capture the misspecification error with an additive term of $\tilde{O}(H \log(W)\, \varepsilon_{\text{miss}})$, where $W = \max_{\pi \in \Pi^E, (s,a) \in \mathcal{S},\mathcal{A}, h \in [H]} \frac{\pi_{E,h}(a|s)}{\pi_h(a|s)}$ and $\epsilon_{miss} = \min_{\pi \in \Pi^E} D_H^2(\mathbb{P}^\pi, \mathbb{P}^{\pi_E})$.

There are related early classic works of (Abbeel & Ng, 2004; Syed & Schapire, 2007) in IL which do not assume expert realizability. However, their approach needs knowledge of the transition kernel (or access to a simulator) and assumes the unknown reward function is linear in a set of known features. ADVIL (Swamy et al., 2021) and IQ-LEARN(Garg et al., 2021) are also based on primal-dual approaches. Garg et al. (2021) approach the IL problem from an inverse RL perspective. IQ-LEARNlearns a single soft Q-function that implicitly represents both a reward and a policy, but it provides no sample complexity guarantees. The ADVIL algorithm of Swamy et al. (2021) comes with an error-propagation guarantee rather than a sample complexity. They state that if the min-max problem is solved to within ϵ, the suboptimality of the learned policy is $O(\epsilon H^2)$. As in (Ross & Bagnell, 2010), such results do not quantify how many expert samples are needed to reach that ϵ, whereas our bounds do.

As mentioned earlier, another assumption that might be more aligned with the underlying tasks in some cases (if not many (Joshi et al., 2025)), is Q^π-realizability. Our paper is mostly related to the recent work of Moulin et al. (2025) (SPOIL), which, to the best of our knowledge, establishes the first sample complexity theoretical results when expert realizability is replaced with Q^π-realizability. (Moulin et al., 2025) achieve the sample complexity in the order of $\tilde{O}(\epsilon^{-2})$ under *linear* Q^π-realizability and $\tilde{O}(\epsilon^{-4})$ for (general) Q^π-realizability assumption. Thus, obtaining better sample complexities for the general Q^π-realizability assumption was left as an open problem. We achieve a sample complexity of $\tilde{O}(\epsilon^{-2})$. Also, ILENA has a better dependence on $1/(1-\gamma)$ (effective horizon) than SPOIL for general Q^π-realizability (refer to Table 1).

Extending this line of inquiry to the online setting, the very recent work of Viano et al. (2026), builds upon the SPOIL framework by introducing OVI, an interactive on-policy imitation learning algorithm. They demonstrate that querying the expert interactively relaxes representational demands, requiring only Q^{π^E}-realizability instead of realizing the expert's policy.

B Omitted Proofs and Lemmas

B.1 Saddle Point Lemmas

The following is the key lemma that helps us not to be limited to best responding for the Q player and is inspired by Cheng et al. (2020).

Lemma B.1. *Suppose $\bar{Q} = \frac{1}{N}\sum_{n=1}^N Q_n$ and $\bar{\pi} = \frac{1}{N}\sum_{n=1}^N \pi_n$, where $\{\pi_n\}_{n=1}^N$ and $\{Q_n\}_{n=1}^N$ are the sequences of actions played by the two base no-regret algorithms. If* $\mathsf{Reg} \leq \epsilon$*, then,*

$$\sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) - \epsilon \leq \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q}) \leq \hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) \leq \sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) \tag{11}$$

Algorithm	**Assumptions**	**Avoids expert realizability**	**Infinite horizon**	**Quadratic Sample Comp.**	**Sample complexity ($\tilde{\Omega}$)**
BC with log loss (Foster et al., 2024)	Π^{E}	$\times$	$\times$	✓	$\frac{H^2 \log \lvert\Pi^{\mathrm{E}}\rvert}{\varepsilon^2}$
BC with 0-1 loss (Rajaraman et al., 2021)	Trajectory Data, $\Pi^{\mathrm{E}}_{\mathrm{det,lin}}$	$\times$	$\times$	✓*	$\frac{H^2 d}{\varepsilon}$
BoostedLogLossBC (Rohatgi et al., 2025)	Π^{E}	✓**	$\times$	✓	$\frac{H^2 \log \lvert\Pi^{\mathrm{E}}\rvert}{\varepsilon^2}$
PROJECTION (Abbeel & Ng, 2004)	Linear reward Known transitions	✓	✓	✓	$\frac{d}{(1-\gamma)^2\varepsilon^2}$
MWAL (Syed & Schapire, 2007)	Linear reward Known transitions	✓	✓	✓	$\frac{\log(d)}{(1-\gamma)^2\varepsilon^2}$
SPOIL (Moulin et al., 2025)	Linear Q^π-realizability	✓	✓	✓	$\frac{d}{(1-\gamma)^4\varepsilon^2}$
SPOIL (Moulin et al., 2025)	Q^π-realizability	✓	✓	$\times$	$\frac{\log \mathcal{N}_{\varepsilon''}(\mathcal{Q})}{(1-\gamma)^8\varepsilon^4}$
ILENA THIS WORK (Thrm 4.6)	Linear Q^π-realizability	✓	✓	✓	$\frac{d}{(1-\gamma)^4\varepsilon^2}$
ILENA THIS WORK (THRM. 4.4)	Q^π-realizability	✓	✓	✓	$\frac{\log(\mathcal{N}_{\epsilon'}(\mathcal{Q}))}{(1-\gamma)^4\varepsilon^2}$

Table 1: Sample complexity of imitation learning for some of the algorithms in the literature. Note that the objective of $V^{\pi_E}(\rho_0) - V^{\bar{\pi}}(\rho_0) \leq \epsilon$ (not the sub-optimality of *normalized* expected return). $\Pi^{\mathrm{E}}_{\mathrm{det,lin}} = \{\pi \mid \exists \theta \in \mathcal{B}(B_\theta), \pi(.) = \arg\max_{a \in \mathcal{A}} \langle \theta, \phi(., a) \rangle\}$. $\epsilon' \in O(\frac{\epsilon^2(1-\gamma)^4}{\boldsymbol{A}})$ and $\varepsilon'' \in \tilde{O}(\epsilon^3)$. ✓*: Knowing that expert is a deterministic policy enables us to have ϵ^{-1} sample complexity. ✓**: With a misspecification error linear in H.

Proof.

$$\begin{aligned}
\mathsf{Reg} \leq \epsilon &\Rightarrow \frac{1}{N} \min_{\pi,Q} \sum_{n=1}^{N} \hat{c}_n(\pi, Q) \geq -\epsilon \\
&\Rightarrow \min_{\pi,Q} \big[\frac{1}{N} \sum_{n=1}^{N} \hat{\mathcal{L}}(\pi, Q_n) - \frac{1}{N} \sum_{n=1}^{N} \hat{\mathcal{L}}(\pi_n, Q) \big] \geq -\epsilon \\
&\Rightarrow \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q}) - \sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) \geq -\epsilon \\
&\Rightarrow \hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) + \epsilon \geq \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q}) + \epsilon \geq \sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q)
\end{aligned}$$

where the first line holds because $\hat{c}_n(\pi_n, Q_n) = 0$, and so does the third line because of the definition of $\bar{Q} = \frac{1}{N} \sum_{n=1}^{N} Q_n$ and $\bar{\pi} = \frac{1}{N} \sum_{n=1}^{N} \pi_n$ and the fact that $\hat{\mathcal{L}}(\pi, .)$ is concave for all π (in fact we use its linearity in the future), and $\hat{\mathcal{L}}(., Q)$ is convex for all Q (this is linear too.) And then the last step is trivial, since $\hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) \geq \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q})$. **Note that in this proof it is not needed to assume that $\bar{Q} \in \mathcal{Q}$ or $\mathcal{Q}$ is convex**. Finally, since $\sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) \geq \hat{\mathcal{L}}(\bar{\pi}, \bar{Q})$, we can have,

$$\sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) \geq \hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) \geq \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q}) \geq \sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q) - \epsilon$$

□

The above lemma shows that any pair of the three quantities of $\sup_{Q \in \mathcal{Q}} \hat{\mathcal{L}}(\bar{\pi}, Q)$, $\min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q})$, and $\hat{\mathcal{L}}(\bar{\pi}, \bar{Q})$ are within an ϵ-neighborhood of each other.

The following lemma is crucial to show that $\hat{\mathcal{L}}(\bar{\pi}, \bar{Q})$ goes to zero with the growth of data and the horizon of the online learning problem.

Lemma B.2. *If* $\mathsf{Reg} \leq \epsilon$*, then*

$$\hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) \leq \frac{1}{N} \sum_{i=1}^{N} \Delta(\pi^E, Q_i) + \epsilon \tag{12}$$

Proof. From the definition of $\mathcal{L}$ we know that

$$\mathcal{L}(\pi^E, Q) = 0 \quad \forall Q \in \mathcal{Q}.$$

Also by definition of Δ we know, $\mathcal{L}(\pi^E, Q) + \Delta(\pi^E, Q) \geq \hat{\mathcal{L}}(\pi^E, Q)$ for all $Q \in \mathcal{Q}$. Thus, for all $Q_i \in \{Q_1, \ldots, Q_N\}$ we can write,

$$\mathcal{L}(\pi^E, Q_i) + \Delta(\pi^E, Q_i) \geq \hat{\mathcal{L}}(\pi^E, Q_i)$$

And by linearity of $\mathcal{L}$ and $\hat{\mathcal{L}}$ we have,

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\pi^E, Q_i) + \frac{1}{N} \sum_{i=1}^{N} \Delta(\pi^E, Q_i) &\geq \frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{L}}(\pi^E, Q_i) \\
&= \hat{\mathcal{L}}(\pi^E, \bar{Q}) \\
&\geq \min_{\pi} \hat{\mathcal{L}}(\pi, \bar{Q}) \\
&\geq \hat{\mathcal{L}}(\bar{\pi}, \bar{Q}) - \epsilon
\end{aligned}$$

Where the last inequality is from Lemma B.1. □

Note that for either of the above lemmas, strong duality is not needed.

B.2 Bounding the Δ terms and The Metric Entropy

Now we need to bound the Δ terms. First, we formally define the set $\tilde{\Pi}$ resulting from this choice of algorithm for the π-player, to be

$$\tilde{\Pi}_{\mathcal{Q},\eta} = \{\pi : \mathcal{S} \to \Delta_{\mathcal{A}} |\ \exists Q \in \mathcal{Q} : \forall s \in \mathcal{S};\ \pi(.|s) = \arg\min_{\pi \in \Delta_{\mathcal{A}}} \eta\langle \pi, -Q(s,.)\rangle + \frac{1}{2}\|\pi\|_2^2\} \cup \{\pi^E\}. \tag{13}$$

Note that we drop the subscript from $\tilde{\Pi}_{\mathcal{Q},\eta}$ since it is clear from the context. The following lemma is similar to Lemma 5 of Moulin et al. (2025). Define $\tilde{\Pi}$ be the set of policies that can be picked by the π-player together with the expert policy, π^E. Therefore $\pi^E \in \tilde{\Pi}$, and $\pi_n \in \tilde{\Pi}$ for all $n \in [N]$. In Lemma 4.2 we see the covering number of $\tilde{\Pi}$ is controlled by the covering number of $\mathcal{Q}$ because the π-player strategies are a simple function of the Q-player strategies (best responding). We need to obtain a uniform concentration inequality for all $\pi \in \tilde{\Pi}$ and all $Q \in \mathcal{Q}$.

Lemma B.3. *with probability $1-\delta$, for all $(\pi, Q) \in \tilde{\Pi} \times \mathcal{Q}$ we have*

$$\Delta(\pi, Q) = |\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)| \leq \frac{4\epsilon'}{(1-\gamma)} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')/\delta)}{\tau(1-\gamma)^2}} \tag{14}$$

Proof. For a fixed pair of $(\pi, Q) \in \tilde{\Pi} \times \mathcal{Q}$, Hoeffding's inequality gives us,

$$\mathbb{P}[|\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)| \geq t] \leq 2\exp(-\frac{2t^2\tau(1-\gamma)^2}{4})$$

and by choosing $t = \sqrt{\frac{2\log(2/\delta)}{\tau(1-\gamma)^2}}$, we equivalently obtain the following with probability $1-\delta$,

$$|\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)| \leq \sqrt{\frac{2\log(2/\delta)}{\tau(1-\gamma)^2}}.$$

Now suppose $\mathcal{C}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')$ is a minimal ϵ'-covering set on $\tilde{\Pi} \times \mathcal{Q}$ under the norm of $\|(\pi, Q)\|_{\infty,1,\infty} := \max_{s \in \mathcal{S}} \|\pi(.|s)\|_1 + \|Q\|_\infty$ and it has the size of $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon') := |\mathcal{C}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')|$. Then by applying union bound, we know that with probability $1-\delta$ for all $(\pi, Q) \in \mathcal{C}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')$

$$|\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)| \leq \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')/\delta)}{\tau(1-\gamma)^2}}$$

Now, to surpass the covering set and have a bound for all $(\pi, Q) \in \tilde{\Pi} \times \mathcal{Q}$, we need to bound $|\mathcal{L}(\pi, Q) - \mathcal{L}(\pi_{\epsilon'}, Q_{\epsilon'})|$ and $|\hat{\mathcal{L}}(\pi, Q) - \hat{\mathcal{L}}(\pi_{\epsilon'}, Q_{\epsilon'})|$ where $(\pi_{\epsilon'}, Q_{\epsilon'}) \in \mathcal{C}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')$ is the closest pair in the ϵ'-covering. From the definition of $\mathcal{L}$ and $\hat{\mathcal{L}}$ and using Holder inequality, we can get (similar to Moulin et al. (2025)) :

$$|\mathcal{L}(\pi, Q) - \mathcal{L}(\pi_{\epsilon'}, Q_{\epsilon'})| \leq \frac{2\epsilon'}{(1-\gamma)} \quad \text{and} \quad |\hat{\mathcal{L}}(\pi, Q) - \hat{\mathcal{L}}(\pi_{\epsilon'}, Q_{\epsilon'})| \leq \frac{2\epsilon'}{(1-\gamma)}$$

Finally, by the triangle inequality, we can obtain,

$$\begin{aligned}|\mathcal{L}(\pi, Q) - \hat{\mathcal{L}}(\pi, Q)| &\leq |\mathcal{L}(\pi, Q) - \mathcal{L}(\pi_{\epsilon'}, Q_{\epsilon'})| + |\hat{\mathcal{L}}(\pi_{\epsilon'}, Q_{\epsilon'}) - \mathcal{L}(\pi_{\epsilon'}, Q_{\epsilon'})| + |\hat{\mathcal{L}}(\pi, Q) - \hat{\mathcal{L}}(\pi_{\epsilon'}, Q_{\epsilon'})| \\ &\leq \frac{4\epsilon'}{(1-\gamma)} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon')/\delta)}{\tau(1-\gamma)^2}}\end{aligned}$$

with probability $1-\delta$ for all $(\pi, Q) \in \tilde{\Pi} \times \mathcal{Q}$. □

Now we can connect the metric entropy of $\tilde{\Pi}$ to the metric entropy of $\mathcal{Q}$,

Lemma B.4. *In the dynamic induced by the two players for the length of any $N \in \mathbb{N}$, where the π-player deploys the regularized best response (either delayed or not) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and the hyperparameter η, and the Q-player using any algorithm over $\mathcal{Q}$,*

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}}) + 1 \quad \textit{and} \quad \mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta \boldsymbol{A}})^2. \tag{15}$$

Proof. To upper bound the covering number of $\tilde{\Pi}$ with that of $\mathcal{Q}$, we need to take into account that each policy $\tilde{\pi} \in \tilde{\Pi}$ is a regularized best response to a $Q \in \mathcal{Q}$ (except π^E which is only one single policy), i.e. a solution to the following optimization problem,

$$\tilde{\pi}(.|s) = f_\eta(Q, s) = \arg\max_{\pi \in \Delta_A} \eta \langle \pi, Q(., s) \rangle - \frac{1}{2}\|\pi\|_2^2 \qquad \forall s \in \mathcal{S}.$$

From the above optimization, one can observe that the solutions of this optimization are equivalent to the solutions of projection optimizations, i.e. for a fixed $q \in \mathbb{R}^A$,

$$\arg\max_{\pi \in \Delta_A} \eta \langle \pi, q \rangle - \frac{1}{2}\|\pi\|_2^2 = \arg\min_{\pi \in \Delta_A} \|\pi - \eta q\|_2.$$

The proof is in Lemma B.5. Also, from Lemma B.14 we know that the projection on a convex set, such as Δ_A, is non-expansive. Therefore, $f_\eta(Q, s)$ is η-Lipschitz in $\|.\|_2$ with respect to Q.

Now we construct the covering ϵ-net over the set of policies $\tilde{\Pi}$. Consider $\mathcal{C}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}})$ to be a covering set in $\mathcal{Q}$. For any $s \in \mathcal{S}$ and for each element of this set $\tilde{Q} \in \mathcal{C}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}})$, pick the regularized best response which is the solutions of the following optimization problems,

$$\tilde{\pi}(.|s) := \arg\max_{\pi \in \Delta_A} \eta \langle \pi, \tilde{Q}(., s) \rangle - \frac{1}{2}\|\pi\|_2^2 = \arg\min_{\pi \in \Delta_A} \|\pi - \eta \tilde{Q}(., s)\|_2.$$

We argue that the set of all such best responses, which has at most $\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}})$ elements, is an ϵ-covering on $\tilde{\Pi}$. Let π' be a policy that belongs to $\tilde{\Pi}$. Thus there exists a $Q' \in \mathcal{Q}$, such that $\pi'(., s) = f_\eta(Q', s)$ for all $s \in \mathcal{S}$. Pick $\tilde{Q}' \in \mathcal{C}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta \boldsymbol{A}})$ as the closest element in the covering set to Q'. And also $\tilde{\pi}'$ be the regularized best response to $\tilde{Q}'$, i.e. $\tilde{\pi}'(., s) := f_\eta(\tilde{Q}', s)$. Then we can write,

$$\begin{aligned}
\|\pi'(.|s) - \tilde{\pi}'(.|s)\|_1 &\leq \sqrt{\boldsymbol{A}} \|\pi'(.|s) - \tilde{\pi}'(.|s)\| && (16)\\
&\leq \eta \sqrt{\boldsymbol{A}} \|Q'(.|s) - \tilde{Q}'(.|s)\| && (17)\\
&\leq \eta \boldsymbol{A} \|Q'(.|s) - \tilde{Q}'(.|s)\|_\infty. && (18)
\end{aligned}$$

The second inequality is due to the Lipschitzness of f_η. Since the above statement holds for all $s \in \mathcal{S}$, we can write,

$$\begin{aligned}
\max_{s \in \mathcal{S}} \|\pi'(.|s) - \tilde{\pi}'(.|s)\|_1 &\leq \eta \boldsymbol{A} \max_{s \in \mathcal{S}} \|Q'(.|s) - \tilde{Q}'(.|s)\|_\infty \\
&= \eta \boldsymbol{A} \|Q' - \tilde{Q}'\|_\infty \\
&\leq \epsilon.
\end{aligned}$$

Therefore, we have,

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{\eta A}) + 1. \tag{19}$$

where the "+1" is for covering π^E. Note that in $\mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon)$ the norm is defined over $\tilde{\Pi}$ and it is $\|\pi\|_{\infty,1} := \max_{s \in \mathcal{S}} \|\pi(.|s)\|_1$. And finally, for the set of $\tilde{\Pi} \times \mathcal{Q}$ under the norm of $\|(\pi, Q)\|_{\infty,1,\infty} := \max_{s \in \mathcal{S}} \|\pi(.|s)\|_1 + \|Q\|_\infty$ we can write,

$$\begin{aligned}
\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon) &\leq \mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon/2) . \mathcal{N}_\infty(\mathcal{Q}, \epsilon/2) && (20)\\
&\leq (\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta A}) + 1)\, \mathcal{N}_\infty(\mathcal{Q}, \epsilon/2) && (21)\\
&\leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon}{2\eta A})^2. && (22)
\end{aligned}$$

□

Lemma B.5. *For a fixed $q \in \mathbb{R}^A$, and any subset $\Pi \subseteq \mathbb{R}^A$, the following holds,*

$$\arg\max_{\pi\in\Pi} \eta\langle\pi, q\rangle - \frac{1}{2}\|\pi\|_2^2 = \arg\min_{\pi\in\Pi} \|\pi - \eta q\|_2.$$

Proof. for any π_1 and π_2 in Π that $\|\pi_1 - \eta q\| \leq \|\pi_2 - \eta q\|$, we can write,

$$\begin{aligned}
&\|\pi_1 - \eta q\|^2 \leq \|\pi_2 - \eta q\|^2 \Leftrightarrow \\
&\|\pi_1\|^2 + \eta^2\|q\|^2 - 2\eta\langle\pi_1, q\rangle \leq \|\pi_2\|^2 + \eta^2\|q\|^2 - 2\eta\langle\pi_2, q\rangle \Leftrightarrow \\
&\eta\langle\pi_1, q\rangle - \frac{1}{2}\|\pi_1\|^2 \geq \eta\langle\pi_2, q\rangle - \frac{1}{2}\|\pi_2\|^2.
\end{aligned}$$

□

B.3 Proof of Main Theorem under Q^π- Realizability

Theorem B.6. *Given a data set of size $\tau = \frac{48\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon^2(1-\gamma)^4}{2304A})/\delta)}{\epsilon^2(1-\gamma)^4}$, and running the two-player repeated game for $N \geq \frac{192^3}{(1-\gamma)^9\epsilon^3}\ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon(1-\gamma)^2/48)\delta^{-1})$ iterations with the π-player deploying the regularized delayed best response (RDBR) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \frac{24}{\epsilon(1-\gamma)^2}$, and the Q-player deploying Lazy Hedge over the expert set of $\mathcal{C}_\infty(\mathcal{Q}, \frac{\epsilon(1-\gamma)^2}{48})$, with the bucket length of $L = \lceil(\frac{5}{2})^{\frac{2}{3}}N^{\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon(1-\gamma)^2}{48})\delta^{-1})^{\frac{-1}{3}}\rceil$ and parameter$\lambda = (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon(1-\gamma)^2}{48}))/K}$, where $K = \lfloor(\frac{2}{5})^{\frac{2}{3}}N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q}, \frac{\epsilon(1-\gamma)^2}{48})\delta^{-1})^{\frac{1}{3}}\rfloor$, with probability of $1-2\delta$,*

$$V_{\pi^E} - V_{\bar{\pi}} \leq \epsilon,$$

where $\bar{\pi} = \frac{1}{N}\sum_{n=1}^N \pi_n$ is the average of the policies taken by the π-player.

Proof. when $\mathsf{Reg} \leq \psi$, from Proposition 2.2 and Proposition 4.1 in which we chose the covering precision being $\epsilon' = \psi/8$, we can write,

$$g_{\pi^E} - g_{\bar{\pi}} \overset{\text{w.p. } 1-2\delta}{\leq} 2\psi + \frac{8\frac{\psi}{8}}{(1-\gamma)} + \sqrt{\frac{2\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi}{8})/\delta)}{\tau(1-\gamma)^2}} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q}, \frac{\psi}{8})/\delta)}{\tau(1-\gamma)^2}},$$

where δ probability of failure is from Proposition 2.2, and δ probability of failure is from Proposition 4.1. Now from Lemma 4.2, and noting that $\eta = \frac{4}{\psi}$,we know

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q}, \frac{\psi}{8}) \leq \mathcal{N}_\infty(\mathcal{Q}, \frac{\psi}{16\eta A})^2 = \mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})^2.$$

Now we can use $\sqrt{log(A)} + \sqrt{\log(B)} \leq \sqrt{2\log(AB)}$, and have

$$\begin{aligned}
\sqrt{\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi}{8})/\delta)} + \sqrt{\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})^2/\delta)} &\leq \sqrt{2\log(4\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi}{8})\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})^2/\delta^2)} \\
&\leq \sqrt{2\log(8\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})^3/\delta^3)} \\
&\leq \sqrt{6\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})/\delta)}
\end{aligned}$$

Thus we obtain

$$g_{\pi^E} - g_{\bar{\pi}} \leq 2\psi + \frac{\psi}{(1-\gamma)} + \sqrt{\frac{12\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})/\delta)}{\tau(1-\gamma)^2}}.$$

Also since $\gamma < 1$ and $V_\pi = g_\pi/(1-\gamma)$, we can write,

$$V_{\pi^E} - V_{\bar{\pi}} \leq \frac{3\psi}{(1-\gamma)^2} + \sqrt{\frac{12\log(2\mathcal{N}_\infty(\mathcal{Q}, \frac{\psi^2}{64A})/\delta)}{\tau(1-\gamma)^4}}.$$

And now we can choose the first term, $\frac{3\psi}{(1-\gamma)^2} = \epsilon/2$ and the second term $\sqrt{\frac{12\log(2\mathcal{N}_\infty(\mathcal{Q},\frac{\psi^2}{64A})/\delta)}{\tau(1-\gamma)^4}} = \epsilon/2$. This gives us $\psi = \frac{\epsilon(1-\gamma)^2}{6}$, and $\tau = \frac{48\log(2\mathcal{N}_\infty(\mathcal{Q},\frac{\psi^2}{64A})/\delta)}{\epsilon^2(1-\gamma)^4} = \frac{48\log(2\mathcal{N}_\infty(\mathcal{Q},\frac{\epsilon^2(1-\gamma)^4}{2304A})/\delta)}{\epsilon^2(1-\gamma)^4}$. This is the sample complexity $\tau \in O(\frac{\log(\mathcal{N}_\infty(\mathcal{Q},\epsilon')/\delta)}{\epsilon^2(1-\gamma)^4})$. Also from Proposition 4.1 we know $N \geq \frac{32^3\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8)\delta^{-1})}{(1-\gamma)^3\psi^3}$ and with the above value of ψ, it implies that $N \geq \frac{192^3}{(1-\gamma)^9\epsilon^3}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon(1-\gamma)^2/48)\delta^{-1})$. □

B.4 No-Regret Learning Lemmas

Proposition B.7. *The regret of the online learning problem is upper-bounded by the sum of Q-player's and π-player's regret,*

$$\mathsf{Reg} \leq \mathsf{Reg}^{\mathcal{Q}} + \frac{1}{\tau}\sum_{i=1}^{\tau}\mathsf{Reg}_i^{\Pi} \tag{23}$$

where

$$\mathsf{Reg}^{\mathcal{Q}} := \frac{1}{N}\max_{Q\in\mathcal{Q}}\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)], \quad \textit{and} \quad \mathsf{Reg}_i^{\Pi} := \frac{1}{N}\max_{\pi\in\Pi}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)].$$

Proof. From the definitions, we can write,

$$\begin{aligned}
\mathsf{Reg} &:= \frac{1}{N}\big[\sum_{n=1}^{N}\hat{c}_n(\pi_n,Q_n) - \min_{\pi,Q}\sum_{n=1}^{N}\hat{c}_n(\pi,Q)\big] \\
&= \frac{1}{N}\big[\sum_{n=1}^{N}\hat{h}_n(\pi_n,Q_n) - \min_{\pi,Q}\sum_{n=1}^{N}\hat{h}_n(\pi,Q)\big] \\
&= \frac{1}{N}\big[\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q_n) + \frac{1}{\tau}\sum_{i=1}^{\tau}h_{n,i}^{\Pi}(\pi_n)] - \min_{\pi,Q}\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q) + \frac{1}{\tau}\sum_{i=1}^{\tau}h_{n,i}^{\Pi}(\pi)]\big] \\
&\leq \frac{1}{N}\max_{Q}\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)] + \frac{1}{N\tau}\sum_{i=1}^{\tau}\max_{\pi}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)]
\end{aligned}$$

where the last inequality holds because

$$\min_{\pi}\sum_{n=1}^{N}\sum_{i=1}^{\tau}h_{n,i}^{\Pi}(\pi) \geq \sum_{i=1}^{\tau}\min_{\pi}\sum_{n=1}^{N}h_{n,i}^{\Pi}(\pi).$$

The last expression shows the regret decomposition. The regret of online learning problem decomposes into two parts, the regret of the Q-player and the regret of the π-player, i.e.

$$\mathsf{Reg} \leq \mathsf{Reg}^{\mathcal{Q}} + \frac{1}{\tau}\sum_{i=1}^{\tau}\mathsf{Reg}_i^{\Pi}.$$

$$\mathsf{Reg} \leq \underbrace{\frac{1}{N}\max_{Q}\sum_{n=1}^{N}[h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)]}_{\mathsf{Reg}^{\mathcal{Q}}} + \frac{1}{\tau}\sum_{i=1}^{\tau}\underbrace{\max_{\pi}\frac{1}{N}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)]}_{\mathsf{Reg}_i^{\Pi}}.$$

□

In the following lemmas, we show that the two base online learning algorithms of the π-player and Q-player have a regret in the order of $\tilde{O}(\frac{1}{\sqrt{N}})$, where N is the number of iterations they play the repeated game. We will show that if the Q-player uses a prediction with expert advice algorithm that not only has a sublinear regret in the adversarial full information case, but also has sublinear switching cost, the later causes the π-player to have sublinear regret too.

Lemma B.8. *For all $i \in [\tau]$, the π-player by deploying the regularized delayed best response (RDBR) with the hyperparameter η and regularizer $\frac{1}{2}\|\pi\|_2^2$ incurs a regret with the following upper bound,*

$$\mathsf{Reg}_i^{\Pi} \leq \frac{1}{\eta} + \frac{2}{N}\sum_{n=1}^{N} \|Q_n(S_i,.) - Q_{n-1}(S_i,.)\|_\infty.$$

Proof. From the description above, we had the following regret definition for the π-player

$$\mathsf{Reg}_i^{\Pi} = \frac{1}{N}\max_{\pi}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)]$$

where the loss function is, $h_{n,i}^{\Pi}(\pi) := \langle \pi(.|S_i), -Q_n(S_i,.)\rangle$. The π-player chooses its action based on the following optimization,

$$\pi_n(.|S_i) = \arg\min_{\pi\in\Delta_A} \eta\langle \pi, -Q_{n-1}(S_i,.)\rangle + \frac{1}{2}\|\pi\|_2^2.$$

For an arbitrary initial action Q_0. The regularizer penalizes π when it goes towards deterministic policies (i.e. exact best-responding). Now we can bound the regret using the following proposition.

Proposition B.9. *(Bandit Algorithms Lattimore & Szepesvári (2020), Chap. 26) For any convex function $g(\pi)$, and its minimizer,*

$$\pi^* := \arg\min_{\pi} g(\pi),$$

we have,

$$\langle \pi^* - \pi, \nabla g(\pi^*)\rangle \leq 0.$$

Thus, by considering $g(\pi)$ to be $\eta\langle \pi, -Q_{n-1}(S_i,.)\rangle + \frac{1}{2}\|\pi\|_2^2$, for all $n \in [N]$ we can write,

$$\begin{aligned}
&\langle \pi_n - \pi, -\eta Q_{n-1}(S_i,.) + \pi_n\rangle \leq 0 \Rightarrow \\
&\langle \pi_n - \pi, -Q_{n-1}(S_i,.)\rangle \leq \frac{1}{\eta}\langle \pi - \pi_n, \pi_n\rangle \leq \frac{1}{\eta}\langle \pi, \pi_n\rangle \leq \frac{1}{\eta}.
\end{aligned}$$

where the last inequality is from the Holder inequality, $\langle \pi, \pi_n\rangle \leq \|\pi\|_1\|\pi_n\|_\infty$, and from the policies live in the simplex.

Now we can split the instantaneous regret and sum up over N iteration and obtain,

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}[h_{n,i}^{\Pi}(\pi_n) - h_{n,i}^{\Pi}(\pi)] &= \frac{1}{N}\sum_{n=1}^{N}\langle \pi_n - \pi, -Q_n(S_i,.)\rangle \\
&= \frac{1}{N}\sum_{n=1}^{N}\langle \pi_n - \pi, -Q_{n-1}(S_i,.) + Q_{n-1}(S_i,.) - Q_n(S_i,.)\rangle \\
&\leq \frac{1}{\eta} + \frac{1}{N}\sum_{n=1}^{N}\langle \pi_n - \pi, Q_{n-1}(S_i,.) - Q_n(S_i,.)\rangle \\
&\leq \frac{1}{\eta} + \frac{1}{N}\sum_{n=1}^{N}\|\pi_n - \pi\|_1\|Q_{n-1}(S_i,.) - Q_n(S_i,.)\|_\infty \\
&\leq \frac{1}{\eta} + \frac{2}{N}\sum_{n=1}^{N}\|Q_{n-1}(S_i,.) - Q_n(S_i,.)\|_\infty.
\end{aligned}$$

By noting that the above statement holds for all $\pi \in \Delta_{\boldsymbol{A}}$ we can recover the $max_{\pi\in\Delta_{\boldsymbol{A}}}$. The Q-player can choose an algorithm with a switching cost in $O(N^{\frac{2}{3}})$ to control the term $\sum_{n=1}^{N}\|Q_{n-1}(S_i,.) - Q_n(S_i,.)\|_\infty$, and the first term can be in the same order with $\eta = N^{\frac{1}{3}}$. Note that for the case that we have controlled the number of switches on the Q-player's action to be at most K, we can rewrite the above bound being,

$$\mathsf{Reg}_i^{\Pi} \leq \frac{1}{\eta} + \frac{4K}{N(1-\gamma)}.$$

□

Lemma B.10. *For all $N \geq 1$ and $\delta \in (0,1)$, **against the π-player who uses RDBR**, the Q-player by following Lazy Hedge algorithm, on the covering set of $\mathcal{C}_\infty(\mathcal{Q}, \epsilon')$ as the expert class, and with parameter $\lambda = (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon'))/K}$, with probability of at least $1-\delta$ incurs a regret with the following upper bound,*

$$\mathsf{Reg}^{\mathcal{Q}} \leq \frac{160^{\frac{1}{3}}}{1-\gamma} N^{-\frac{1}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon')\delta^{-1})^{\frac{1}{3}} + \frac{20^{\frac{2}{3}}}{1-\gamma} N^{-\frac{2}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon')\delta^{-1})^{\frac{-1}{3}} + 2\epsilon'.$$

where $K = \lfloor N/L \rfloor$ is the number of buckets and $L = \lceil (\frac{5}{2})^{\frac{2}{3}} N^{\frac{1}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon')\delta^{-1})^{\frac{-1}{3}} \rceil$ is the length of each bucket. Which in turn shows the number of switching experts by the Q-player is $K = \lfloor (\frac{2}{5})^{\frac{2}{3}} N^{\frac{2}{3}} \ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon')\delta^{-1})^{\frac{1}{3}} \rfloor$.

Proof. First of all, from the definition of $\mathsf{Reg}^{\mathcal{Q}}$, and by adding and subtracting the following terms, we can write,

$$\begin{aligned}
\mathsf{Reg}^{\mathcal{Q}} &= \frac{1}{N} \max_{Q} \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)] \\
&= \frac{1}{N} \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q^*_{net}) + h_n^{\mathcal{Q}}(Q^*_{net}) - h_n^{\mathcal{Q}}(Q^*)] \\
&\leq \frac{1}{N} \max_{Q_{net} \in \mathcal{C}_\infty(\mathcal{Q}, \epsilon')} \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q_{net})] + \frac{1}{N} \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q^*_{net}) - h_n^{\mathcal{Q}}(Q^*)],
\end{aligned}$$

where Q^* is in the $\arg\max$ of the regret defined over the original set $\mathcal{Q}$, and Q^*_{net} is the nearest point in the covering to Q^*, which is at most ϵ' far away under the infinity norm (on both actions and states). We note the first term of the last statement above as $\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}$. We then control the remaining term,

$$\begin{aligned}
\frac{1}{N} \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q^*_{net}) - h_n^{\mathcal{Q}}(Q^*)] &= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\tau} \sum_{i=1}^{\tau} \langle \pi_n(.|S_i) - \boldsymbol{\delta}(A_i), Q^*_{net}(S_i, .) - Q^*(S_i, .) \rangle \\
&\leq \frac{1}{N\tau} \sum_{n=1}^{N} \sum_{i=1}^{\tau} \|\pi_n(.|S_i) - \boldsymbol{\delta}(A_i)\|_1 \|Q^*_{net}(S_i, .) - Q^*(S_i, .)\|_\infty \\
&\leq \frac{1}{N\tau} \sum_{n=1}^{N} \sum_{i=1}^{\tau} 2\epsilon' \\
&\leq 2\epsilon'
\end{aligned}$$

Thus we know

$$\mathsf{Reg}^{\mathcal{Q}} \leq \mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})} + 2\epsilon'.$$

Now we move on to control $\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}$. Resulting from the Lazy Hedge algorithm, the Q-player keeps playing the same action within buckets of length L, and it changes its actions only going from the last time step of the bucket to the first time step of the new bucket. In addition, because the π-player best responds to the previous action of the Q-player, $h^{\mathcal{Q}}_{n+2}(Q) = h^{\mathcal{Q}}_{n+3}(Q) = \cdots = h^{\mathcal{Q}}_{n+L}(Q)$ for all n divisible by L, i.e. $n = k.L$ for some $k \in \mathbb{Z}$. Thus for all $Q \in \mathcal{C}_\infty(\mathcal{Q}, \epsilon')$, we can write the regret as follows,

$$\begin{aligned}
N.\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}(Q) &= \sum_{n=1}^{N} [h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)] \\
&= \sum_{k=1}^{K} (L-1)[h^{\mathcal{Q}}_{Lk}(Q_{Lk}) - h^{\mathcal{Q}}_{Lk}(Q)] + \sum_{k=1}^{K} [h^{\mathcal{Q}}_{L(k-1)+1}(Q_{Lk}) - h^{\mathcal{Q}}_{L(k-1)+1}(Q)] \\
&\quad + \sum_{n=LK+1}^{N} [h_n^{\mathcal{Q}}(Q_n) - h_n^{\mathcal{Q}}(Q)],
\end{aligned}$$

where $K = \lfloor \frac{N}{L} \rfloor$, and Q_{Lk} is the Q-player's action (expert) chosen in the k-th bucket. The second summation is for the K many transient time steps between the buckets that we bound them separately (the red time steps in Figure 1). And the third summation is for the remaining time steps at the end in case N is not dividable by L (The last red bucket that has less than L iterations in Figure 1). We know

$$\begin{aligned}|h_n^{\mathcal{Q}}(Q') - h_n^{\mathcal{Q}}(Q)| &\leq \frac{1}{N\tau}\sum_{n=1}^{N}\sum_{i=1}^{\tau}\|\pi_n(.|S_i) - \boldsymbol{\delta}(A_i)\|_1\|Q'(S_i,.) - Q(S_i,.)\|_\infty \\ &\leq \frac{4}{1-\gamma},\end{aligned}$$

since all of the $Q \in \mathcal{Q}$ has bounded norm of $\|Q\|_\infty \leq \frac{1}{1-\gamma}$. Thus, we can upper bound the two last summations in the previous statement as follows,

$$N.\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}(Q) \leq \sum_{k=1}^{K}(L-1)[h_{Lk}^{\mathcal{Q}}(Q_{Lk}) - h_{Lk}^{\mathcal{Q}}(Q)] + \frac{4K}{1-\gamma} + \frac{4(L-1)}{1-\gamma},$$

and the first summation can be controlled by the regret bound of regular Hedge (Cesa-Bianchi & Lugosi (2006) Corollary 4.2),

$$\begin{aligned}N.\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}(Q) &\leq (L-1)\sum_{k=1}^{K}[h_{Lk}^{\mathcal{Q}}(Q_{Lk}) - h_{Lk}^{\mathcal{Q}}(Q)] + \frac{4K}{(1-\gamma)} + \frac{4(L-1)}{(1-\gamma)} \\ &\overset{\text{w.p. } 1-\delta}{\leq} \frac{2L}{(1-\gamma)}\big[\sqrt{\frac{1}{2}K\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon'))} + \sqrt{\frac{1}{2}K\ln(1/\delta)}\big] + \frac{4K}{(1-\gamma)} + \frac{4L}{(1-\gamma)},\end{aligned}$$

Note that $\|Q\|_\infty \leq \frac{1}{1-\gamma}$ for all $Q \in \mathcal{Q}$. Also $K = \lfloor \frac{N}{L} \rfloor$ so we can replace K with N/L, and also with knowing that $\sqrt{\ln(A)} + \sqrt{\ln(B)} \leq \sqrt{2\ln(AB)}$ for all $A, B \geq 1$, we can write,

$$\begin{aligned}N(1-\gamma).\mathsf{Reg}^{\mathcal{C}_\infty(\mathcal{Q})}(Q) &\overset{\text{w.p. } 1-\delta}{\leq} 2\sqrt{NL\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})} + \frac{4N}{L} + 4L \\ &\leq 160^{\frac{1}{3}}N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} + 20^{\frac{2}{3}}N^{\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{-1}{3}}\end{aligned}$$

where the last inequality is by choosing the length of the buckets to be $L = \lceil(\frac{5}{2})^{\frac{2}{3}}N^{\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{-1}{3}}\rceil$ (And because of the ceiling we consider extra $K = N/L$) . This gives the number of buckets, or equivalently the number of switching experts for the algorithm to be $K = \lfloor(\frac{2}{5})^{\frac{2}{3}}N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}}\rfloor$. The proof is concluded by rearranging the $N(1-\gamma)$ term and considering the first part which gives us the extra term of $2\epsilon'$.

Proposition B.11. *For any $\delta \in (0,1)$, the dynamic with the horizon of $N \geq \frac{32^3}{(1-\gamma)^3\psi^3}\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8)\delta^{-1})$ induced by the two players, where the π-player deploys the regularized delayed best response (RDBR) with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \frac{4}{\psi}$, and the Q-player deploys the lazy Hedge over the expert set of $\mathcal{C}_\infty(\mathcal{Q},\psi/8)$ with bucket length of $L = \lceil(\frac{5}{2})^{\frac{2}{3}}N^{\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8)\delta^{-1})^{\frac{-1}{3}}\rceil$ and parameter $\lambda = (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8))/K}$, where $K = \lfloor(\frac{2}{5})^{\frac{2}{3}}N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8)\delta^{-1})^{\frac{1}{3}}\rfloor$, with probability of $1-\delta$ has the regret of $\mathsf{Reg} \leq \psi$.*

Proof. From the decomposition of Reg in 2.1, we can write

$$\mathsf{Reg} \leq \mathsf{Reg}^{\mathcal{Q}} + \frac{1}{\tau}\sum_{i=1}^{\tau}\mathsf{Reg}_i^{\Pi}.$$

Then we can use the results of regret bounds for the both players, i.e. Proposition B.8, Proposition B.10. For $C_1 = \frac{160^{\frac{1}{3}}}{1-\gamma}$, $C_2 = \frac{20^{\frac{2}{3}}}{1-\gamma}$ and $C_3 = (\frac{2}{5})^{\frac{2}{3}}$, we know,

$$\begin{aligned}\mathsf{Reg} \leq\ & C_1N^{-\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} + C_2N^{-\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{-1}{3}} + 2\epsilon' \\ &+ \frac{1}{\eta} + \frac{4}{N(1-\gamma)}C_3N^{\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}}.\end{aligned}$$

Thus we get,

$$\begin{aligned}\mathsf{Reg} &\leq (C_1 + \frac{4}{1-\gamma}C_3)N^{-\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} + C_2N^{-\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{-1}{3}} + 2\epsilon' + \frac{1}{\eta} \\ &\leq \frac{8}{1-\gamma}N^{-\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} + \frac{15}{2(1-\gamma)}N^{-\frac{2}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{-1}{3}} + 2\epsilon' + \frac{1}{\eta} \\ &\leq \frac{16}{1-\gamma}N^{-\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} + 2\epsilon' + \frac{1}{\eta}.\end{aligned}$$

where the last inequality is to eliminate the lower-order term. Now to have the whole expression less than ψ, we choose$\frac{1}{\eta} = \psi/4$, and $2\epsilon' = \psi/4$, and $\frac{16}{1-\gamma}N^{-\frac{1}{3}}\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon')\delta^{-1})^{\frac{1}{3}} \leq \psi/2$. This results in having $N \geq \frac{32^3}{(1-\gamma)^3\psi^3}\ln(\mathcal{N}_\infty(\mathcal{Q},\psi/8)\delta^{-1})$. □

The following lemma is for π-player's algorithm under linear Q^π-realizability, which is a regularized best-response algorithm (without delay).

Lemma B.12. *For all $i \in [\tau]$, the π-player by deploying the regularized best response (RBR) with the hyperparameter η incurs a regret with the following upper bound,*

$$\mathsf{Reg}_i^\Pi \leq \frac{1}{\eta}.$$

Proof. From the description above, we had the following regret definition for the π-player

$$\mathsf{Reg}_i^\Pi = \frac{1}{N}\max_\pi \sum_{n=1}^N [h_{n,i}^\Pi(\pi_n) - h_{n,i}^\Pi(\pi)]$$

where the loss function is, $h_{n,i}^\Pi(\pi) := \langle \pi(.|S_i), -Q_n(S_i,.)\rangle$. The π-player chooses its action based on the following optimization,

$$\pi_n(.|S_i) = \arg\min_{\pi\in\Delta_A} \eta\langle\pi, -Q_n(S_i,.)\rangle + \frac{1}{2}\|\pi\|_2^2.$$

For an arbitrary Q_0. The regularizer penalizes π when it is towards deterministic policies. Now we can bound the regret using the Proposition B.9, and considering $g(\pi)$ to be $\eta\langle\pi, -Q_n(S_i,.)\rangle + \frac{1}{2}\|\pi\|_2^2$, for all $n \in [N]$ we can write,

$$\begin{aligned}&\langle \pi_n - \pi, -\eta Q_n(S_i,.) + \pi_n\rangle \leq 0 \Rightarrow \\ &\langle \pi_n - \pi, -Q_n(S_i,.)\rangle \leq \frac{1}{\eta}\langle \pi - \pi_n, \pi_n\rangle \leq \frac{1}{\eta}\langle \pi, \pi_n\rangle \leq \frac{1}{\eta}.\end{aligned}$$

where the last inequality is from the Holder inequality, $\langle \pi, \pi_n\rangle \leq \|\pi\|_1\|\pi_n\|_\infty$, and from the policies live in the simplex.

Now we can sum up the instantaneous regret over N iterations and obtain,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N [h_{n,i}^\Pi(\pi_n) - h_{n,i}^\Pi(\pi)] &= \frac{1}{N}\sum_{n=1}^N \langle \pi_n - \pi, -Q_n(S_i,.)\rangle \\ &\leq \frac{1}{N}\sum_{n=1}^N \frac{1}{\eta} = \frac{1}{\eta}\end{aligned}$$

By noting that the above statement holds for all $\pi \in \Delta_{\boldsymbol{A}}$ we can recover the $max_{\pi\in\Delta_{\boldsymbol{A}}}$. □

B.5 Convex Optimization Lemmas

Lemma B.13. *In a Hilbert space $\mathcal{H}$ and norm $\|.\|$, suppose $\pi^* \in \arg\min_{\pi\in\mathcal{C}}\|q-\pi\|$, where $q \in \mathcal{H}$ is fixed and $\mathcal{C}$ a convex subset of $\mathcal{H}$. Then the following holds,*

$$\langle q - \pi^*, \pi - \pi^*\rangle \leq 0 \quad \forall \pi \in \mathcal{C}.$$

Proof. We prove the lemma through contradiction. Suppose that there exists a $\pi \in \mathcal{C}$, such that

$$\langle q - \pi^*, \pi - \pi^* \rangle > 0.$$

Since $\mathcal{C}$ is convex, we know there exist an $\alpha \in (0, 1)$, such that,

$$\langle q - \pi', \pi - \pi' \rangle > 0,$$

where $\pi' := \pi^* + \alpha(\pi - \pi^*)$. Note that since $\mathcal{C}$ is convex, $\pi' \in \mathcal{C}$. Now we can write,

$$\begin{aligned} \|q - \pi'\|^2 &= \|q - \pi^* - \alpha(\pi - \pi^*)\|^2 \\ &= \|q - \pi^*\|^2 + \alpha^2 \|\pi - \pi^*\|^2 - 2\alpha \langle q - \pi^*, \pi - \pi^* \rangle. \end{aligned}$$

And because we assumed $\langle q - \pi^*, \pi - \pi^* \rangle > 0$, there exist an small enough α such that

$$\alpha^2 \|\pi - \pi^*\|^2 - 2\alpha \langle q - \pi^*, \pi - \pi^* \rangle < 0.$$

This means $\|q - \pi'\| < \|q - \pi^*\|$, which gives us the contradiction. □

Lemma B.14. *Consider the Hilbert space $\mathcal{H}$ with norm $\|.\|$, and a convex subset $\mathcal{C}$ of it. Then for any q_1 and q_2 in $\mathcal{H}$, their projection under the norm on $\mathcal{C}$ is defined as,*

$$\pi_1 :\in \arg\min_{\pi \in \mathcal{C}} \|q_1 - \pi\| \quad \textit{and} \quad \pi_2 :\in \arg\min_{\pi \in \mathcal{C}} \|q_2 - \pi\|.$$

Then we have

$$\|\pi_1 - \pi_2\| \leq \|q_1 - q_2\|.$$

Proof. From Lemma B.13, we have

$$\langle q_1 - \pi_1, \pi_2 - \pi_1 \rangle \leq 0 \quad \text{and} \quad \langle q_2 - \pi_2, \pi_1 - \pi_2 \rangle \leq 0.$$

By adding the LHS of the inequalities, we obtain,

$$\begin{aligned} &\langle q_1 - \pi_1 + \pi_2 - q_2, \pi_2 - \pi_1 \rangle \\ &= \langle q_1 - q_2, \pi_2 - \pi_1 \rangle + \|\pi_2 - \pi_1\|^2 \leq 0. \end{aligned}$$

By moving the first part we get,

$$\|\pi_2 - \pi_1\|^2 \leq \langle q_2 - q_1, \pi_2 - \pi_1 \rangle \leq \|q_2 - q_1\| \, \|\pi_2 - \pi_1\|,$$

where the last Cauchy-Schwarz inequality concludes the proof. □

B.6 Linear Q^π Realizability

Regret Decomposition in Linear Case. The result of Proposition 2.1 is generic and holds for the linear Q^π-realizability as well. The linear Q^π-realizability assumption enables us to represent the utilities and the regret of Q-player regret in terms of features and the parameter θ. We can write the following in the linear case,

$$h_{n,i}^{\mathcal{B}_\theta}(\theta) := \langle \pi_n(.|S_i) - \boldsymbol{\delta}(A_i), \phi(S_i, .)\theta \rangle,$$

where $\phi(S_i, .)$ is treated as a matrix in $\mathbb{R}^{\boldsymbol{A} \times d}$. Therefore, by defining $y_n := \frac{1}{\tau} \sum_{i=1}^{\tau} (\pi_n(.|S_i) - \boldsymbol{\delta}(A_i))^\top \phi(S_i, .)$, as the loss vector for the Q-player (or θ-player in this case) at iteration n, we have, $h_n^{\mathcal{B}_\theta}(\theta) = \langle y_n, \theta \rangle$. Note that $y_n \in \mathbb{R}^d$. Now we can represent the regret of the Q-player in the linear setting as,

$$\mathsf{Reg}^{\mathcal{B}_\theta} := \frac{1}{N} \max_{\theta \in \mathcal{B}_\theta} \sum_{n=1}^{N} [h_n^{\mathcal{B}_\theta}(\theta_n) - h_n^{\mathcal{B}_\theta}(\theta)]. \tag{24}$$

For the π-player the regret Reg_i^Π and the cost function $h_{n,i}^\Pi(\pi)$ will remain unchanged in the linear case.

B.6.1 ILENA UNDER LINEAR Q^π-REALIZABILITY ASSUMPTION

The π-player's No-regret Algorithm (Alg^π). For every game corresponding to state S_i, the π-player picks its policy at iteration $n \in [N]$ by following a regularized best responding algorithm with parameter $\eta = \sqrt{N}$,

$$\pi_n(.|S_i) = \arg\min_{\pi \in \Delta_A} \eta \langle \pi, -Q_n(S_i, .)\rangle + \frac{1}{2}\|\pi\|_2^2.$$

Note that the π-player chooses its action (π_n) *based on* the action played by the leader, Q_n. In Lemma B.12 we show that the regret of this algorithm $\mathsf{Reg}_i^\Pi \leq \frac{1}{\sqrt{N}}$. Note that there is no randomness in this π-player's algorithm, and this upper bound on Reg_i^Π holds for any sequence of Q_n.

Algorithm 2: ILENA (linear Q^π realizability)

Input: $\mathcal{D} = \{(S_i, A_i)\}_{i=1}^\tau, \mathcal{B}_\theta, N, \mathsf{Alg}^Q, \mathsf{Alg}^\pi$

for $n \in \{1, 2, \ldots, N\}$ **do**
- Q-player takes θ_n suggested by Alg^Q ▷ Mirror Descent on $\mathcal{B}_\theta$
- **for** $i \in \{1, 2, \ldots, \tau\}$ **do**
  - π-player takes $\pi_n(.|S_i)$ suggested by Alg_i^π ▷ Regularized Best Response (RBR)
- $h_n^{\mathcal{B}_\theta}(.)$ is revealed to the Q-player, and $h_{n,i}^\Pi(.)$ to the π-player ▷ Full info. feedback

Result $\bar{\pi}(.|s) = \frac{1}{N}\sum_{n=1}^N \pi_n(.|s)$ *for all* $s \in \mathcal{S}$

Lemma B.15. *(Similar to Lattimore & Szepesvári (2020), Proposition 28.6) For all $N \geq 1$, the Q-player by following Mirror Descent on the ball of $\mathcal{B}_\theta$ with potential function $F(\theta) = \frac{1}{2}\|\theta\|_2^2$ and parameter $\lambda = \frac{B_\theta}{2B_\phi\sqrt{N}}$, the regret will be bounded as follows,*

$$\mathsf{Reg}^{\mathcal{B}_\theta} \leq \frac{2B_\theta B_\phi}{\sqrt{N}}$$

Remark B.16. *Note that there is no randomness in the Q-player algorithm making a deterministic regret guarantee for it. Also MD algorithm has the above regret guarantee even for the adaptive adversary that best responds to the current action θ_n. Thus we do not need to incorporate the delayed best responding idea in the linear case.*

Proof.

$$\mathsf{Reg}^{\mathcal{B}_\theta} := \frac{1}{N}\max_{\theta \in \mathcal{B}_\theta} \sum_{n=1}^N [h_n^{\mathcal{B}_\theta}(\theta_n) - h_n^{\mathcal{B}_\theta}(\theta)]$$

By Equation 28.9 in Lattimore & Szepesvári (2020), we have $\tilde{\theta}_{n+1} = \theta_n - \lambda y_n$. Recall that $y_n = \frac{1}{\tau}\sum_{i=1}^\tau (\pi_n(.|S_i) - \delta(A_i))^\top \phi(S_i, .)$. Thus,

$$D_F(\theta_n, \tilde{\theta}_{n+1}) = \frac{1}{2}\|\tilde{\theta}_{n+1} - \theta_n\|_2^2 = \frac{\lambda}{2}\|y_n\|_2^2.$$

Using the triangle inequality, we can write, $\|y_n\| \leq \frac{1}{\tau}\sum_{i=1}^\tau (\|\pi(.|S_i)^\top \phi(S_i, .)\| + \|\delta(A_i)^\top \phi(S_i, .)\|)$ and since π and δ are probability distributions, and again the triangle inequality, we have, $\|y_n\| \leq \frac{1}{\tau}\sum_{i=1}^\tau (\sum_a (\pi(a|S_i) + \mathbf{1}(a = A_i))\|\phi(S_i, a)\|) \leq 2B_\phi$, for all n. Also, $Diam_F(\mathcal{B}_\theta) = \frac{1}{2}B_\theta^2$. Note that the last statements follow from Assumption 0.2. Thus,

$$\begin{aligned}\mathsf{Reg}^{\mathcal{B}_\theta} &\leq \frac{Diam_F(\mathcal{B}_\theta)}{N\lambda} + \frac{\lambda}{2N}\sum_{n=1}^N \|y_n\|_2^2 \\ &\leq \frac{B_\theta^2}{2N\lambda} + 2\lambda B_\phi^2 \leq \frac{2B_\theta B_\phi}{\sqrt{N}},\end{aligned}$$

where the last inequality is from the choice of $\lambda = \frac{B_\theta}{2B_\phi\sqrt{N}}$. □

Proposition B.17. *The dynamic induced by the two players after $N = \frac{(2B_\theta B_\phi+1)^2}{\epsilon^2}$ iterations, where the π-player deploys the regularized best response with the regularizer of $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter $\eta = \sqrt{N}$, and Q-player deploys mirror descent on the ball of $\mathcal{B}_\theta$ with potential function $F(\theta) = \frac{1}{2}\|\theta\|_2^2$ and parameter $\lambda = \frac{B_\theta}{2B_\phi\sqrt{N}}$, deterministically, has the regret of* $\mathsf{Reg} \leq \epsilon$.

Proof. The proof can be easily obtained from Lemma B.12 and Lemma B.15 , and the regret decomposition 2.1. With the choice of $N = \frac{(2B_\theta B_\phi+1)^2}{\epsilon^2}$, can write,

$$\mathsf{Reg} \leq \mathsf{Reg}^{\mathcal{B}_\theta} + \frac{1}{\tau}\sum_{i=1}^{\tau}\mathsf{Reg}_i^{\Pi} \leq \frac{2B_\theta B_\phi}{\sqrt{N}} + \sqrt{\frac{1}{N}} \leq \epsilon.$$

□

We define Q_{lin} be the set of all state-action value functions that can be represented by the inner product of a vector $\theta \in \mathcal{B}_\theta$ and the features, i.e.

$$\mathcal{Q}_{lin} = \{Q \in \mathbb{R}^{S\times A} |\ \exists\theta \in \mathcal{B}_\theta : \forall(s,a) \in \mathcal{S}\times\mathcal{A}\ \ Q(s,a) = \langle\phi(s,a),\theta\rangle\}.$$

Now we can state the following lemma that contributes in controlling the statistical gap for linear case.

Lemma B.18. *Under the linear Q^π-realizability assumption (Assumption 0.2), and any interaction between the θ-player and π-player, where the π-player deploys the regularized best response with regularizer $\frac{1}{2}\|\pi\|_2^2$ and hyperparameter η, we have*

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q}_{lin},\epsilon) \leq (1+\frac{4B_\theta B_\phi \eta A}{\epsilon})^{2d},$$

and

$$\mathcal{N}_\infty(\mathcal{Q}_{lin},\epsilon) \leq \mathcal{N}_2(\mathcal{B}_\theta,\frac{\epsilon}{B_\phi}) \leq (1+\frac{2B_\theta B_\phi}{\epsilon})^d.$$

Proof. In the linear case, we are interested in having an upper bound $\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q}_{lin},\epsilon)$. The proof is similar to Lemma 4.2, and by the same initial steps (Inequality 16) we reach

$$\begin{aligned}\|\pi'(.|s)-\tilde{\pi}'(.|s)\|_1 &\leq \sqrt{\boldsymbol{A}}\|\pi'(.|s)-\tilde{\pi}'(.|s)\| \\ &\leq \eta\sqrt{\boldsymbol{A}}\|Q'(.|s)-\tilde{Q}'(.|s)\|.\end{aligned}$$

Then we continue by using linear Q^π-realizability assumption 0.2,

$$\begin{aligned}\|\pi'(.|s)-\tilde{\pi}'(.|s)\|_1 &\leq \eta\sqrt{\boldsymbol{A}}\ \|Q'(s,.)-\tilde{Q}'(s,.)\|_2 \\ &= \eta\sqrt{\boldsymbol{A}\sum_{a\in\mathcal{A}}\langle\phi(s,a),\theta'-\tilde{\theta}'\rangle^2} \\ &\leq \eta\sqrt{\boldsymbol{A}\sum_{a\in\mathcal{A}}\|\phi(s,a)\|_2^2\|\theta'-\tilde{\theta}'\|_2^2} \\ &\leq \eta\sqrt{\boldsymbol{A}^2B_\phi^2\|\theta'-\tilde{\theta}'\|^2} \\ &= \eta\boldsymbol{A}B_\phi\|\theta'-\tilde{\theta}'\|.\end{aligned}$$

Since the above inequality holds for all $s\in\mathcal{S}$, we have

$$\sup_{s\in\mathcal{S}}\|\pi'(.|s)-\tilde{\pi}'(.|s)\|_1 = \|\pi'-\tilde{\pi}'\|_{\infty,1} \leq \eta\boldsymbol{A}B_\phi\|\theta'-\tilde{\theta}'\|.$$

This means that if $\Theta_N (\subseteq \mathcal{B}_\theta)$ be the set of all of parameter vectors that is possible to appear in the interaction between Q-player and π-player under the aforementioned algorithms and over N rounds, then any minimal covering with precision $\frac{\epsilon}{\eta A B_\phi}$ over Θ_N induces a covering over $\tilde{\Pi}$ with precision ϵ. This is because the above inequality implies that the set of regularized best responses to the points of covering in Θ_N makes an ϵ-net over $\tilde{\Pi}$. Therefore, since $\Theta_N \subseteq \mathcal{B}_\theta$ we can say,

$$\mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon) \leq \mathcal{N}_2(\mathcal{B}_\theta, \frac{\epsilon}{\eta A B_\phi}) + 1,$$

where "+ 1" is for considering π_E in the covering set too.

Now we move on to the part corresponding to $\mathcal{Q}_{lin}$. For every Q_1 and Q_2 in $\mathcal{Q}_{lin}$, we can write,

$$\begin{aligned}
\|Q_1 - Q_2\|_\infty &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\langle \phi(s,a), \theta_1 - \theta_2 \rangle| \\
&\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\| \|\theta_1 - \theta_2\| \\
&\leq B_\phi \|\theta_1 - \theta_2\|
\end{aligned}$$

where the first inequality is Cauchy-Schwarz and the second is by Assumption 0.2. With the same argument as before, we can write,

$$\mathcal{N}_\infty(\mathcal{Q}_{lin}, \epsilon) \leq \mathcal{N}_2(\mathcal{B}_\theta, \frac{\epsilon}{B_\phi}).$$

Now, from the definition of $\|.\|_{\infty,1,\infty}$, we can write,

$$\begin{aligned}
\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}_{lin}, \epsilon) &\leq \mathcal{N}_{\infty,1}(\tilde{\Pi}, \epsilon/2) . \mathcal{N}_\infty(\mathcal{Q}_{lin}, \epsilon/2) \\
&\leq (\mathcal{N}_2(\mathcal{B}_\theta, \frac{\epsilon}{2\eta A B_\phi}) + 1) \ . \mathcal{N}_2(\mathcal{B}_\theta, \frac{\epsilon}{2B_\phi}) \\
&\leq \mathcal{N}_2(\mathcal{B}_\theta, \frac{\epsilon}{2\eta A B_\phi})^2 \\
&\leq (1 + \frac{4B_\theta B_\phi \eta A}{\epsilon})^{2d}.
\end{aligned}$$

The last inequality is from Lemma B.20. □

B.6.2 POOF OF THEOREM 4.6, THE MAIN THEOREM IN THE LINEAR CASE:

The backbone of the proof consists of two main parts: (i) Controlling the optimization gap, in which the difference is using Mirror Descent instead of Lazy Hedge for Q-player and regularized best response (RBR) for the π player. (refer to Proposition B.17) (ii) Analyzing how the linear structure and the chosen base algorithms affect the joint set of $\tilde{\Pi} \times \mathcal{Q}$ and its covering number $\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \epsilon)$, to control the statistical gap (refer to B.18).

Theorem B.19. *Given a data set of size* $\tau = \frac{64d \log(\frac{2}{\delta} + \frac{3200 B_\phi B_\theta A(1+2B_\theta B_\phi)}{\delta(1-\gamma)^4\epsilon^2})}{(1-\gamma)^4\epsilon^2}$*, and running two player repeated game for* $N = \frac{400(1+2B_\phi B_\theta)^2}{(1-\gamma)^4\epsilon^2}$ *iterations with the* π*-player deploying the regularized best response with the regularizer of* $\frac{1}{2}\|\pi\|_2^2$ *and hyperparameter* $\eta = \sqrt{N}$*, and the* Q*-player deploying Mirror Descent on the ball* $\mathcal{B}_\theta$ *with potential function* $F(\theta) = \frac{1}{2}\|\theta\|_2^2$ *and parameter*$\lambda = \frac{B_\theta}{2B_\phi\sqrt{N}}$*, with probability of* $1 - \delta$*,*

$$V_{\pi^E} - V_{\bar{\pi}} \leq \epsilon,$$

where $\bar{\pi} = \frac{1}{N}\sum_{n=1}^N \pi_n$ *is the average of the policies taken by the* π*-player.*

Proof. From Proposition 2.2, we know for any dynamic in the game between Q-player and π-player (including our choice mentioned in the theorem) that provides $\mathsf{Reg} \leq \psi$, with probability $1 - \delta$,

$$g_{\pi^E} - g_{\bar{\pi}} \leq 2\psi + \frac{8\psi}{(1-\gamma)} + \sqrt{\frac{2\log(2\mathcal{N}_\infty(\mathcal{Q}, \psi)/\delta)}{\tau(1-\gamma)^2}} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi} \times \mathcal{Q}, \psi)/\delta)}{\tau(1-\gamma)^2}}.$$

We choose $\epsilon = \epsilon' = \psi$ and we will choose a proper value for ψ shortly. From Lemma B.18 and doing some algebra, we can write,

$$\begin{aligned}
g_{\pi^E} - g_{\bar{\pi}} &\leq 2\psi + \frac{8\psi}{(1-\gamma)} + \sqrt{\frac{2\log(2\mathcal{N}_\infty(\mathcal{Q},\psi)/\delta)}{\tau(1-\gamma)^2}} + \sqrt{\frac{2\log(2\mathcal{N}_{\infty,1}(\tilde{\Pi}\times\mathcal{Q},\psi)/\delta)}{\tau(1-\gamma)^2}} \\
&\leq \frac{10\psi}{1-\gamma} + \sqrt{\frac{2\log(2\mathcal{N}_2(\mathcal{B}_\theta,\frac{\psi}{B_\phi})/\delta)}{\tau(1-\gamma)^2}} + \sqrt{\frac{2\log(2\mathcal{N}_2(\mathcal{B}_\theta,\frac{\psi}{2\eta A B_\phi})^2/\delta)}{\tau(1-\gamma)^2}} \\
&\leq \frac{10\psi}{1-\gamma} + 2\sqrt{\frac{2\log(2\mathcal{N}_2(\mathcal{B}_\theta,\frac{\psi}{2\eta A B_\phi})^2/\delta)}{\tau(1-\gamma)^2}} \\
&\leq \frac{10\psi}{1-\gamma} + \frac{2}{1-\gamma}\sqrt{\frac{2\log(2(1+\frac{4B_\theta B_\phi \eta A}{\psi})^{2d}/\delta)}{\tau}} \\
&\leq \frac{10\psi}{1-\gamma} + \frac{4}{1-\gamma}\sqrt{\frac{d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})}{\tau}}.
\end{aligned}$$

Which in turn implies,

$$V_{\pi^E} - V_{\bar{\pi}} \leq \frac{10\psi}{(1-\gamma)^2} + \frac{4}{(1-\gamma)^2}\sqrt{\frac{d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})}{\tau}}.$$

Now, we choose ψ and τ such that $\frac{10\psi}{(1-\gamma)^2} = \frac{\epsilon}{2}$ and also the other term $\frac{4}{(1-\gamma)^2}\sqrt{\frac{d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})}{\tau}} = \frac{\epsilon}{2}$, in this way the right-hand side of the inequality will be ϵ and we will conclude the proof. The first one implies that $\psi = \frac{(1-\gamma)^2\epsilon}{20}$, and from Proposition B.17 we know that the horizon of the dynamic should be $N = \frac{(1+2B_\phi B_\theta)^2}{\psi^2} = \frac{400(1+2B_\phi B_\theta)^2}{(1-\gamma)^4\epsilon^2}$.

For the second term, we write,

$$\begin{aligned}
\frac{4}{(1-\gamma)^2}\sqrt{\frac{d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})}{\tau}} = \frac{\epsilon}{2} &\Rightarrow \frac{8}{(1-\gamma)^2\epsilon}\sqrt{d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})} = \sqrt{\tau} \\
&\Rightarrow \tau = \frac{64\, d\log(\frac{2}{\delta}+\frac{8B_\theta B_\phi A\eta}{\psi\delta})}{(1-\gamma)^4\epsilon^2}.
\end{aligned}$$

Now we just need to substitute the values of $\psi = \frac{(1-\gamma)^2\epsilon}{20}$, and $\eta = \sqrt{N} = \frac{20(1+2B_\phi B_\theta)}{(1-\gamma)^2\epsilon}$ in the numerator, which gives us,

$$\tau = \frac{64d\log(\frac{2}{\delta}+\frac{3200B_\phi B_\theta A(1+2B_\theta B_\phi)}{\delta(1-\gamma)^4\epsilon^2})}{(1-\gamma)^4\epsilon^2}.$$

□

Lemma B.20. *(Vershynin, 2018) The covering number of the Euclidean ball of radius R in $\mathbb{R}^d$, $\mathcal{B}(R)$, is*

$$\mathcal{N}_2(\mathcal{B}(R),\epsilon) \leq (1+\frac{2R}{\epsilon})^d.$$

C EXPERIMENT

Confirming the theoretical results, we also show that ILENA is able to learn in the setting that the expert belongs to a complex class of policies. In these settings, value-based methods such as ILENA and SPOIL , outperform other offline imitation learning algorithms algorithms, such as behavior cloning (Foster et al., 2024), IQLearn (Garg et al., 2021), and ValueDICE(Kostrikov et al., 2019).

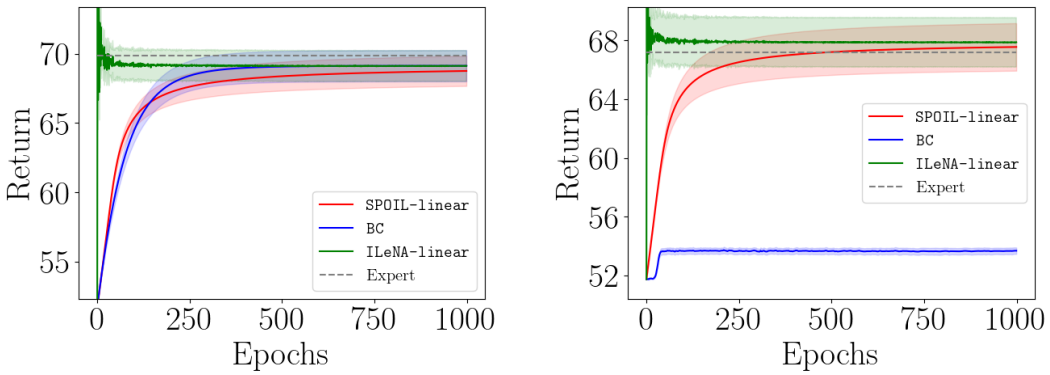


Figure 2: Linear MDP, $\tau = 100$: return of each method's output policy versus the number of epochs, for the simple (softmax-linear, left) and the complex (neural, right) expert. ILENA -linear and SPOIL -linear show the value of their state-wise average $\bar{\pi}_n$, and BC the value of its current policy. Curves show the mean over 10 seeds and shaded regions one standard error of the mean.

C.1 Random Linear MDP (Similar to Moulin et al. (2025) Environment)

Linear MDP. To study the performance of algorithms under linear Q^π-realizability, for the first experiment, we work with the same environment as (Moulin et al., 2025) for a fair comparison. Following Moulin et al. (2025), we use a randomly generated linear MDP with $|\mathcal{S}| = 1000$, $|\mathcal{A}| = 500$, $d = 7$, $\gamma = 0.99$ and uniform ρ_0. Features are $\phi(s,a) = \mathrm{softmax}(g_{s,a}) \in \Delta_d$ with $g_{s,a} \sim \mathcal{N}(0, I_d)$, the reward is $r(s,a) = \langle \phi(s,a), w_r \rangle$ with $w_r \sim \mathcal{N}(0, I_d)$, and the transitions are $P(\cdot|s,a) = \sum_j \phi_j(s,a) W_j$ with $W_j = \mathrm{softmax}(h_j)$, $h_j \sim \mathcal{N}(0, I_{|\mathcal{S}|})$. For every policy the Bellman equation gives $Q^\pi = \langle \phi, w_r + \gamma W V^\pi \rangle$, so the MDP is linearly Q^π-realizable (Assumption 0.2) with the known features ϕ.

Experts and data. The *simple* expert is the soft-optimal policy $\pi^E(\cdot|s) = \mathrm{softmax}(Q^*_{\mathrm{soft}}(s,\cdot))$, computed by soft value iteration; it is a softmax-linear policy with 7 parameters. The *complex* expert is a three-layer MLP (1000 hidden units per layer, one-hot state input) trained to match the simple expert in KL divergence, so it has a similar value but belongs to a much larger policy class. The simple expert belongs to a much smaller set of policies, and and the complex expert to a larger one. Accordingly, we expect the algorithms like BC that search on the set of experts, will perform worse in the complex policy class. For each expert we sample $\tau = 100$ state-action pairs from μ_{π^E} and repeat over 10 seeds.

Methods. ILENA uses Algorithm 3: Mirror Descent on the ball of radius $B_\theta = 5$ for the Q-player and the regularized best response with $\eta = \sqrt{N}$ for the π-player, for $N = 1000$ rounds, and outputs the state-wise average $\bar{\pi}_N(\cdot|s) = \frac{1}{N}\sum_n \pi_n(\cdot|s)$. SPOIL uses the implementation of Moulin et al. (2025) with the same features. BC maximizes the log-likelihood of the expert actions, over softmax-linear policies for the simple expert and over the same MLP architecture as the expert for the complex one. The return is $V^\pi(\rho_0)$.

Results. Since the environment is linear we use the linear versions of ILENA and SPOIL . One epoch is one pass over the $\tau = 100$ expert samples: one iteration of the two-player game for ILENA -linear and SPOIL -linear, one full-batch log-likelihood gradient step for linear BC, and four minibatch Adam steps for neural BC. The x-axis thus measures computation rather than data, and the speed of each curve mainly reflects its step sizes. With the simple expert class, BC is given the correct softmax-linear policy class, and all three methods end within one unit of the expert's value (69.9), with differences between them smaller than the variation across seeds. With the complex expert class, BC's neural policy cannot generalize from 100 samples to the 1000 states and stays close to the uniform policy (53.7), whereas ILENA -linear (67.9) and SPOIL -linear (67.6), which rely only on the linear structure of Q^π, are unaffected by the complexity of the expert's policy class and reach the expert's value (67.3).

C.2 Correct Answer MDP: Non-convex Set of Q^π

To evaluate ILENA under general (non-convex) Q^π-realizability, we keep the low-rank dynamics of the linear experiment, $P(\cdot \mid s,a) = \sum_{j=1}^{d} \phi_j(s,a)\, W_j(\cdot)$, where $\phi(s,a) \in \Delta_d$ and each row $W_j \in \Delta_{\mathcal{S}}$, with $|\mathcal{S}| = 300$, $|\mathcal{A}| = 50$, $d = 4$, $\gamma = 0.9$ and ρ_0 uniform. However, to incorporate a non-convex set of Q, we replace the linear reward with a sparse binary one. The learner is given a hypothesis class $\{f_1, \dots, f_M\}$, where each hypothesis f marks a set $C_f(s) \subset \mathcal{A}$ of 5 correct actions in every state, and

$$r(s,a) = r_{f^*}(s,a) = \mathbf{1}\big[a \in C_{f^*}(s)\big]$$

for an unknown f^*. The correct sets are drawn uniformly at random and independently across hypotheses and states. Any policy supported on $C_{f^*}(s)$ in every state collects a reward of one at every step and is therefore optimal, with $V^\pi = 1/(1-\gamma)$. The expert is such a policy with arbitrary (Dirichlet) weights over the correct actions: it is optimal, and it belongs to an enormous policy class.

The set of action-value functions. Because the dynamics are low-rank, the Bellman equation separates the reward from the future value. For every policy π,

$$Q^\pi(s,a) = r_{f^*}(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\pi(s') = r_{f^*}(s,a) + \big\langle \phi(s,a), \theta^\pi \big\rangle, \qquad \theta^\pi := \gamma\, W V^\pi \in \mathbb{R}^d.$$

The policy enters only through the d numbers θ^π. Unlike in the linear experiment, the reward is not linear in ϕ, so it remains a separate term, and since f^* is unknown the learner must consider every hypothesis.

Two observations let us restrict θ to a small box. First, since $\phi(s,a) \in \Delta_d$, shifting θ by $c\mathbf{1}$ adds the same constant c to every $Q(s,a)$, and this cancels in L and $\hat{L}$ because $\sum_a \pi(a|s) = 1$; only the span of $\max_j \theta_j - \min_j \theta_j$ matters. Second, let $TV_W := \max_{j,k} \mathrm{TV}(W_j, W_k)$. For every π, $\mathrm{spread}(\theta^\pi) \le \gamma\, TV_W \,\mathrm{span}(V^\pi)$, and the Bellman equation gives $\mathrm{span}(V^\pi) \le 1 + \gamma\, TV_W \,\mathrm{span}(V^\pi)$. Hence

$$\theta^\pi \in c^\pi \mathbf{1} + [0, \delta_\theta]^d, \qquad \delta_\theta := \frac{\gamma\, TV_W}{1 - \gamma\, TV_W},$$

for some $c^\pi \in \mathbb{R}$. In our constructed MDP and for the sampled W we compute $TV_W = \max_{j,k} \mathrm{TV}(W_j, W_k) = 0.300$, which gives $\delta_\theta = 0.370$. Therefore, the class $\mathcal{Q}$ contains Q^π for every policy of this instance. Thus the class we give to the algorithms is therefore

$$\mathcal{Q} = \bigcup_{f=1}^{M} \mathcal{Q}_f, \qquad \mathcal{Q}_f := \Big\{ r_f + \langle \phi, \theta \rangle :\ \theta \in [0, \delta_\theta]^d \Big\},$$

which contains Q^π (up to the irrelevant constant) for every π. This set of action value functions is a union of M affine d-dimensional pieces, which is a non-convex set, with metric entropy $\log \mathcal{N}_\infty(\mathcal{Q}, \epsilon) \le \log M + d \log\big(1 + \delta_\theta/(2\epsilon)\big)$, whereas its linear span has dimension $M + d$. Varying M thus controls the gap between the class and its convex hull.

Covering Set and Lazy Hedge. The hypothesis part of $\mathcal{Q}$ is finite and needs no covering. For the θ part we use a product grid Θ_{grid} on $[0, \delta_\theta]^d$ with spacing 2ϵ. Since $\phi(s,a) \in \Delta_d$, $|\langle \phi(s,a), \theta - \theta' \rangle| \le \|\theta - \theta'\|_\infty$, so

$$\mathcal{C}_\infty(\mathcal{Q}, \epsilon) = \Big\{ r_f + \langle \phi, \theta \rangle :\ f \in [M],\ \theta \in \Theta_{\mathrm{grid}} \Big\}, \qquad |\mathcal{C}_\infty(\mathcal{Q}, \epsilon)| \le M \big(\lceil \delta_\theta/(2\epsilon) \rceil + 1 \big)^d,$$

is an ϵ-cover of $\mathcal{Q}$ in $\|\cdot\|_\infty$. In our experiments we use 4 grid points per coordinate, i.e. $|\Theta_{\mathrm{grid}}| = 256$ which implies $\epsilon = \frac{1}{2}\frac{\delta_\theta}{3} \approx 0.06$. The Q-player's cost is separable over this cover,

$$h_n^Q\big(r_f + \langle \phi, \theta \rangle\big) = \underbrace{\frac{1}{\tau} \sum_{i=1}^{\tau} \Big[\big\langle \pi_n(\cdot|S_i), r_f(S_i, \cdot) \big\rangle - r_f(S_i, A_i) \Big]}_{c_n(f)} + \langle y_n, \theta \rangle,$$

so the exponential weights of Lazy Hedge factorize exactly, $p_k(f, \theta) \propto e^{-\lambda \sum_{\kappa<k} c_\kappa(f)}\, e^{-\lambda \sum_{\kappa<k} \langle y_\kappa, \theta \rangle}$. Sampling the realized action of a bucket therefore amounts to sampling f and θ independently, at a cost of $O(M + |\Theta_{\mathrm{grid}}|)$ per bucket rather than $O(M|\Theta_{\mathrm{grid}}|)$.

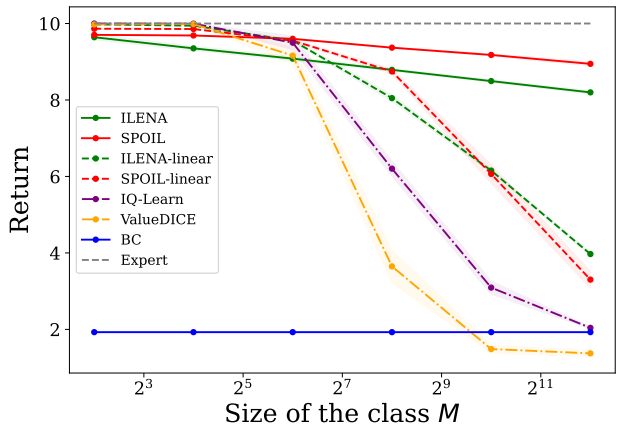


Figure 3: Correct-answer MDP: return of the output policy versus the size M of the hypothesis class, with $\tau = 30$ expert samples and $N = 1000$ iterations. Curves show the mean and shaded regions show standard error over 10 seeds. The expert is optimal ($V^{\pi^E} = 1/(1-\gamma) = 10$). Solid ILENA and SPOIL use the non-convex class $\mathcal{Q} = \bigcup_f \mathcal{Q}_f$.

Convexified baselines. A method that can only use linear realizability must work with the span of $\mathcal{Q}$, i.e. the features $\psi(s,a) = \big[r_1(s,a), \dots, r_M(s,a),\ \delta_\theta\, \phi(s,a)\big] \in \mathbb{R}^{M+d}$, with $Q_w = \langle \psi, w \rangle$ and $\|w\|_2 \le \sqrt{1+d}$ (the true Q^π corresponds to $w = (e_{f^*}, \theta^\pi/\delta_\theta)$). Its metric entropy grows like $(M+d)\log(1/\epsilon)$ instead of $\log M + d\log(1/\epsilon)$. We run ILENA-linear, SPOIL-linear, IQ-Learn and ValueDICE on these features, and ILENA (Lazy Hedge + RDBR) and SPOIL (best response + FTRL) on the non-convex class $\mathcal{Q}$ itself.

Results on the Second Environment. Figure 3 isolates the effect of non-convexity. We fix the data ($\tau = 30$ expert state-action pairs) and increase the number of hypotheses $M \in \{4, 16, \dots, 4096\}$, which enlarges the class of action-value functions by only $\log M$ in metric entropy but its linear span by M dimensions. ILENA (Lazy Hedge over $\mathcal{C}_\infty(\mathcal{Q}, \epsilon)$ with $\lambda = 10$, and RDBR with $\eta = 5$) and SPOIL (exact best response over $\mathcal{Q}$ and softmax-FTRL with $\eta = 0.08$) run on $\mathcal{Q}$ itself for $N = 1000$ iterations and output the state-wise average $\bar{\pi}_N$. ILENA -linear (Algorithm 2) and SPOIL -linear run for the same number of iterations on the features $\psi = [r_1, \dots, r_M, \delta_\theta \phi]$. IQ-Learn and ValueDICE run for 1000 Adam steps on the same features, and additionally receive the next state and the initial state of each sample (this is an additional information for these two algorithms than ILENA , SPOIL , and BC which only nead the state-action pairs). BC is an MLP trained with the log loss for 300 epochs. The methods that exploit the non-convex structure, ILENA and SPOIL , degrade slowly. BC does not use the action value class $\mathcal{Q}$, so its return (about 1.9) is independent of M; its policy class cannot generalize to the many states unseen in the data.

Figure 4 tracks the value of the state-wise average $\bar{\pi}_N = \frac{1}{N}\sum_{n \le N} \pi_n$ over a single run of 2500 iterations at a fixed dataset size ($\tau = 100$). With a small class ($M = 64$), all four methods approach the expert: after 2500 iterations ILENA-linear reaches 9.98, SPOIL-linear 9.94, SPOIL 9.87 and ILENA 9.51. As shown in Figure 4, ILENA , especially its linear version, reaches its converging return faster than SPOIL . It is worth mentioning that comparing ILENA and SPOIL in N (number of iterations) captures the number of times their actor chooses a new policy and their critic chooses a new action-value function. With a large class ($M = 4096$), the linear-feature methods plateau far below the expert: ILENA-linear stays at 6.6 from $N \approx 250$ onwards, and SPOIL-linear rises only slowly, to 6.0. Their gap is therefore statistical rather than an optimization error, and more iterations do not close it. The methods that use the non-convex class keep their performance: ILENA reaches 9.05, and SPOIL peaks at 9.53 around $N = 1500$–2000. SPOIL then decreases slightly, to 9.46 at $N = 2500$. This decrease exceeds the standard error, but it is small compared with the gap to the linear-feature methods.

With $M = 1024$ (Figure 5, left), the methods that use the non-convex class are much more sample-efficient: with $\tau = 5$, SPOIL and ILENA reach 6.8 and 5.6, while all linear-feature methods stay at

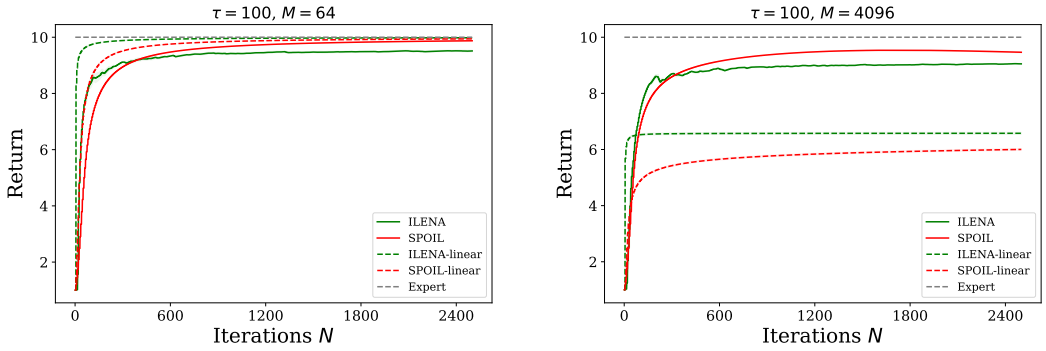


Figure 4: Correct-answer MDP: return of the output policy $\bar{\pi}_n$ versus the number of iterations n of the two-player game, with $\tau = 100$ expert samples, for a small ($M = 64$, left) and a large ($M = 4096$, right) hypothesis class. Curves show the mean and shaded regions show standard error over 10 seeds.

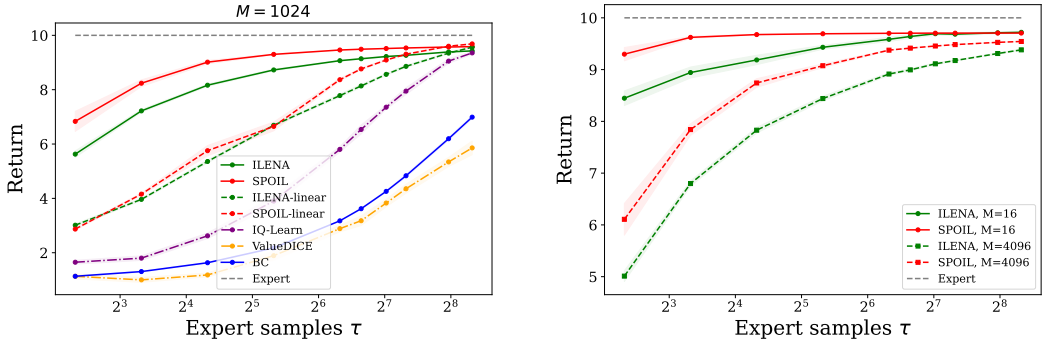


Figure 5: Correct-answer MDP: return of the output policy versus the number of expert samples τ, after $N = 1000$ iterations (or gradient steps). Left: all methods, $M = 1024$. Right: ILENA and SPOIL with $M = 16$ and $M = 4096$. Curves show the mean over 10 seeds and shaded regions one standard error of the mean.

or below 3.0. A return of 9 takes about 20 samples for SPOIL and 80 for ILENA , compared with about 130 for SPOIL-linear, between 160 and 250 for ILENA -linear, and 250 for IQ-Learn. With enough data the linear-feature methods catch up (9.35–9.69 at $\tau = 320$), while BC and ValueDICE reach only 7.0 and 5.9. The right panel shows that a larger class mainly costs samples at small τ: from $\tau = 5$ to $\tau = 320$, the gap between $M = 16$ and $M = 4096$ shrinks from 3.4 to 0.3 for ILENA and from 3.2 to 0.2 for SPOIL. For $M = 16$, ILENA matches SPOIL from $\tau \approx 130$ onwards. Overall, this suggests that SPOIL is also performing well empirically, suggesting that their proved sub-optimality gap of order ϵ^{-4} is loose.

Finally, Figure 6 summarizes Figure 5 (left) as a sample complexity: the number of expert samples each method needs to come within ϵ of the expert. For every $\epsilon \geq 1$, the methods that use the non-convex class need the fewest samples. For $\epsilon = 1$, SPOIL needs about 20 samples and ILENA about 70, compared with about 120 for SPOIL-linear, 180 for ILENA-linear and 245 for IQ-Learn, while ValueDICE and BC do not reach this level within $\tau \leq 320$ (they reach only $\epsilon = 6$ and $\epsilon = 4$, respectively). At the strictest level reached, $\epsilon = 0.5$, SPOIL needs about 110 samples, SPOIL-linear 215 and ILENA-linear 300; ILENA and IQ-Learn do not reach it within the sweep, as their returns plateau at about 9.4. Note that reference slopes are a visual guide only. The theory provides upper bounds, and the sweep is too coarse to estimate empirical rates. However, we suspect that the ϵ^{-4} sample complexity give in Moulin et al. (2025) is a loose result for SPOIL , since its empirical performance, at least in this environment, is more suggesting a ϵ^{-2} sample complexity.

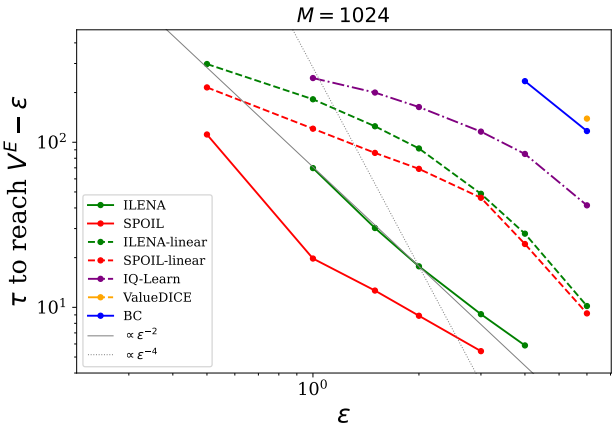


Figure 6: Correct-answer MDP, $M = 1024$: number of expert samples τ needed for the mean return (over 10 seeds, from the sweep of Figure 5) to reach $V^{\pi^E} - \epsilon$. The sweep uses $\tau \in \{5, 10, 20, 40, 80, 100, 130, 160, 250, 320\}$, and values between sweep points are interpolated linearly in $\log \tau$. Levels already reached at $\tau = 5$ or not reached by $\tau = 320$ are omitted. Gray lines are reference slopes ϵ^{-2} and ϵ^{-4} anchored at ILENA's point at $\epsilon = 2$.

D Further Explanation and Discussion

We have put some further explanations regarding this paper here in appendix.

D.1 Lazy Hedge Algorithm Pseudocode

Algorithm 3 illustrates the pseudocode of Lazy Hedge Algorithm which was use as the no-regret algorithm of the critic (Q-player) under general Q^π-realizability.

Algorithm 3: Lazy Hedge

Input: N, L, $\mathcal{C}(= \mathcal{C}_\infty(\mathcal{Q}, \epsilon_N))$, $\lambda(= (1-\gamma)\sqrt{2\ln(\mathcal{N}_\infty(\mathcal{Q}, \epsilon_N))/K}$

for $k \in \{1, 2, \ldots, \lceil \frac{N}{L} \rceil\}$ **do**

  Q-player puts the distribution over the experts in $\mathcal{C}$ as follows, ▷ Hedge

$$p_{q,k} = \frac{\exp(-\lambda \sum_{\kappa=1}^{k-1} h_\kappa(q))}{\sum_{q' \in \mathcal{C}} \exp(-\lambda \sum_{\kappa=1}^{k-1} h_\kappa(q'))} \qquad \forall q \in \mathcal{C}$$

  Pick the realized action $q_k \sim p_k$. (p_1 is uniform distribution over $\mathcal{C}$)

  for $n \in \{(k-1)L + 1, \ldots, kL\}$ **do**

    if $n \leq N$ **then**

      Keep playing the sampled action for the k-th bucket, i.e. $Q_n = q_k$

  Set $h_k(.)$ to be the last cost vector of the bucket, $h^Q_{kL}(.)$ ▷ Full info. feedback

D.2 Exponential Difference Between Covering Numbers of $\mathcal{Q}$ and Π

In this section, we support the argument that there are many cases in which having an assumption on different achievable value functions (Q^π realizability) might be notably better than the expert realizability assumption. Consider an MDP with $2^d - 1$ states, and two actions of right and left. The transition kernel is deterministic, such that it makes a binary tree structure on the MDP. The rewards of taking actions in all states are zero, except for the last state $s_{2^d-2} = s^*$ for which $r(s^*, right)$ is 1. The states at level d are all absorbing states. Figure 7 depicts the MDP.

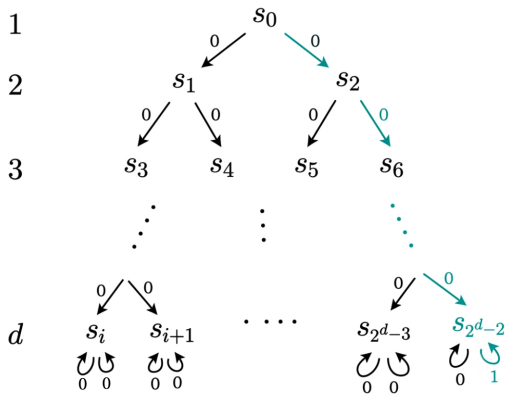


Figure 7: The tree MDP, with large $\mathcal{N}(\Pi, \epsilon)$ and small $\mathcal{N}(\mathcal{Q}, \epsilon)$

To begin with, we only limit ourselves to deterministic policies. We have $2^{\boldsymbol{S}} = 2^{2^d-1}$ deterministic policies. However, under deterministic policies, there is only $d+1$ different Q functions such that the action value function of any deterministic policy π is one of those. This depends on the highest level that the policy π deviates from the green path, denoted by $l^* \in \{0, 1, \dots, d\}$. Until l^*, all states have 0 value, and for the states after and including l^*, such as s_h which are at layer $h \in \{l^*, l^*+1, \dots, d\}$, the policy must stay on the green path and have $Q^\pi(s, right) = \gamma^{d-h}$. This means that there is a *super exponential* difference between the size of the set of deterministic policies, and their corresponding set of action value functions, i.e. $|\Pi_{det}| = 2^{2^d-1}$ and $|\mathcal{Q}^\pi_{det}| = d+1$.

Now we move on to consider stochastic policies. We describe the state-action value function of a given policy π in this MDP. For any π, action $a \in \{right, left\}$, and state $s \in \mathcal{S}$, $Q^\pi(s_h, a) = \alpha$ where α is the discounted probability of reaching to $(s^*, right)$ from (s, a). α is non-zero only for $(s, right)$ pairs where s is on the green path, which are d many states. This means that $\mathcal{Q}^\pi$, the set of Q functions of all possible policies, can be equivalent to a subset of $[0,1]^d$. This means $\mathcal{N}_\infty(\mathcal{Q}^\pi, \epsilon) \in O((\frac{1}{\epsilon})^d)$.

However, the covering number of the set of policies is super exponential in d under different choices of norms. For instant, under TV distance $\mathcal{N}_{\infty,TV}(\Pi, \epsilon) \in O((\frac{1}{\epsilon})^{2^d})$. Remember $\|\pi - \pi'\|_{\infty,TV} = \sup_{s\in\mathcal{S}} TV(\pi(.|s), \pi'(.|s))$. Foster et al. (2024) use other norm of $\|\pi - \pi'\|_{\log} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\log(\frac{\pi(a|s)}{\pi'(a|s)})|$ which makes the covering number even worse since $\log$ function has infinite slope near zero probabilities. In that case, we need to put a threshold p^- as the smallest probability of taking action by any policy, and in turn deal with the misspecification error since the expert policy might violate this threshold.

In conclusion, we observed that for some MDPs the $\log \mathcal{N}(\mathcal{Q}, \epsilon)$ can be exponentially smaller than $\log \mathcal{N}(\Pi, \epsilon)$, which means the sample complexity results based on the former can be notably better. Roughly speaking, the argument is similar to the data processing inequality. The set of different state-action value functions is a function of a set of different policies, so one can expect that its size should be smaller than the size of the main source.

D.3 Computational Complexity

ILENA has two main loops, one over the iterations of the underlying two-player games, and the other for the π player over the data points. Under the general Q^π-realizability assumption, we have K updates for the lazy hedge algorithm, and each optimization query of hedge takes $O(\boldsymbol{A}\mathcal{N}_\infty(\mathcal{Q}, \epsilon_N))$ per sample. So the overall computational complexity will be $O(K\tau\boldsymbol{A}\mathcal{N}_\infty(\mathcal{Q}, \epsilon_N))$ taking the max. For RDBR, we have $O(K\tau)$ updates, and each optimization takes $O(\boldsymbol{A}\log(\boldsymbol{A}))$ computation since it is a Euclidean projection optimization. Thus Lazy Hedge complexity dominates, and putting the values from Theorem 4.4 gives us, $\tilde{O}(\boldsymbol{A}\epsilon^{-4}(1-\gamma)^{-10}\mathcal{N}_\infty(\mathcal{Q}, \epsilon(1-\gamma)^2))$.

Under linear Q^π realizability, $\tau \in \tilde{O}(d/(1-\gamma)^4\epsilon^2)$ and $N \in O((B_\theta B\phi)^2/(1-\gamma)^4\epsilon^2)$. Since the computation cost of mirror descent on $\mathcal{B}_\theta$ is in $O(\tau + d)$ per iteration, the final computation

	Sample Complexity (τ)	Repeated Game iterations (N)	Computational complexity
Q^π realizability	$O(\frac{\log(\mathcal{N}_\infty(\mathcal{Q},\epsilon_\tau)/\delta)}{\epsilon^2(1-\gamma)^4})$	$O(\frac{\ln(\mathcal{N}_\infty(\mathcal{Q},\epsilon_N)/\delta)}{(1-\gamma)^9\epsilon^3})$	$\tilde{O}(\frac{\boldsymbol{A}\mathcal{N}_\infty(\mathcal{Q},\epsilon_N)}{\epsilon^4(1-\gamma)^{10}})$
Linear Q^π Realizability	$\tilde{O}(\frac{d}{(1-\gamma)^4\epsilon^2})$	$O(\frac{B_\phi^2 B_\theta^2}{(1-\gamma)^4\epsilon^2})$	$\tilde{O}(\frac{d^2 B_\theta^2 B_\phi^2 \boldsymbol{A}}{\epsilon^4(1-\gamma)^8})$

Table 2: Sample and computational complexity of ILENA under the two different assumptions. Note $\epsilon_\tau = O(\frac{\epsilon^2(1-\gamma)^4}{\boldsymbol{A}})$ and $\epsilon_N = O(\epsilon(1-\gamma)^2)$.

complexity will be in $\tilde{O}(N\tau\boldsymbol{A}d)$. This is because the RBR computation for the π-player dominates, which has Ad operations for matrix multiplication and $O(A\log(A))$ operations for the Euclidean projection, per iteration. The final order will be $\tilde{O}(\frac{d^2 B_\theta^2 B_\phi^2 \boldsymbol{A}}{\epsilon^4(1-\gamma)^8})$.

D.4 Future Work

There are many future directions for this work. Firstly, there might be other algorithms with better sample or computational complexities. This paper can point a direction to find such algorithms, potentially through a better choice of no-regret algorithms. For instance, there might exist a better choice for the pair of no-regret algorithms, that gives us fast convergence of $\tilde{O}(\frac{1}{N})$, as well as having the statistical gap of $\tilde{O}(\sqrt{\frac{\log(\mathcal{N}_\infty(\mathcal{Q},\frac{\epsilon}{2\eta A}))}{\tau}})$, for the general Q^π-realizability. Although it would not improve the sample complexity of $\tilde{O}(\epsilon^{-2})$, it could save some computational cost. Also, it can be a guide for replacing the expert realizability assumption in online imitation learning. It is also interesting to work on other forms of access to expert data, such as a state-only dataset, or having access only to $\phi(S_i, A_i)$, without knowing S_i and A_i in the linear case. Having data from multiple experts can also be an exciting extension of this work. Last but not least, is studying the lower bound results under the Q^π-realizability assumption in imitation learning.

D.5 The No Regret Algorithms which did not work

To shave off the aforementioned factor of N in the numerator of the fraction in the statistical gap, we need to have a π player that does not depend on multiple Q-player's actions in past. So it rules out the algorithms such as MD or FTRL for the π-player. We can not use an instance of FTPL either, since we have τ many games in which the π-player should play. The good thing is that for each instance, as the result of FTPL, the actions are going to be the vertices of the simplex $\Delta_{\mathcal{A}}$, and only $\boldsymbol{A}$ many actions will be played by the π-player, but we have many states and the number of points we should consider as the covering for the π-players strategies, is $A^{|\mathcal{S}|}$. This does not depend on the covering number of the action value function $\mathcal{Q}$ and which is not desired.

One might also think the π-player can do the regularized best response, and the guarantee for the Q-player can be achieved via an RVU analysis. Although the RVU analysis works, but to connect the regret have in the RVU to the actual regret, we need an additional Martingale concentration which breaks due to the dependence of the π-players action to the current action Q_t. This is the point that leads us to have best responding to the previous action and guarantee the π-players regret through a less switching in the Q-players actions.

However, there is still another choice of algorithm for the π-player that might work. What if we have a lazy best responder for the π-player that guarantees its own external regret, but it also sometimes favors the Q-player so that it also can achieve no-external regret?! This might replace our choice of algorithms. We leave these alternative design of base no-regret algorithms under the framework of ILENA for the future work.