# Rate optimal learning of equilibria from data

Till Freihaut*
freihaut@ifi.uzh.ch
University of Zurich

Luca Viano*
luca.viano@epfl.ch
EPFL

Emanuele Nevali
emanuele.nevali@epfl.ch
EPFL

Volkan Cevher
volkan.cevher@epfl.ch
EPFL

Matthieu Geist
matthieu@earthspecies.org
Earth Species Project

Giorgia Ramponi
ramponi@ifi.uzh.ch
University of Zurich

**Abstract**

We close open theoretical gaps in Multi-Agent Imitation Learning (MAIL) by characterizing the limits of non-interactive MAIL and presenting the first interactive algorithm with near-optimal sample complexity. In the non-interactive setting, we prove a statistical lower bound that identifies the *all-policy deviation concentrability coefficient* as the fundamental complexity measure, and we show that Behavior Cloning (BC) is rate-optimal. For the interactive setting, we introduce a framework that combines reward-free reinforcement learning with interactive MAIL and instantiate it with an algorithm, *MAIL-WARM*. It improves the best previously known sample complexity from $\mathcal{O}(\varepsilon^{-8})$ to $\mathcal{O}(\varepsilon^{-2})$, matching the dependence on $\varepsilon$ implied by our lower bound. Finally, we provide numerical results that support our theory and illustrate, in environments such as grid worlds, where Behavior Cloning fails to learn.

## 1 Introduction

More and more AI systems are deployed in real-world scenarios. This naturally leads to AI systems interacting and adapting their behavior to each other. Importantly, this interaction can be captured as a multi-agent system [Hammond et al., 2025] , and since reward functions are often inaccessible, learning directly from expert demonstrations via Imitation Learning (IL) becomes especially compelling. To capture expert behavior without knowing a reward function, IL serves as a great framework, showcasing impressive empirical success in single-agent settings [Torabi et al., 2019, Jain et al., 2025, Foster et al., 2024] and strong theoretical guarantees [Foster et al., 2024, Viano et al., 2024, Rajaraman et al., 2020]. However, applying IL to multi-agent systems remains largely underexplored. In particular, previous works [Yu et al., 2019, Song et al., 2018, Bui et al., 2024] are mostly empirical and they lack theoretical guarantees. Moreover, they often fail to capture the potentially strategic behavior of agents acting in multi-agent systems since they are not designed to learn a Nash equilibrium profile.

Closer to our work, Tang et al. [2024] showed that optimizing the objective that captures these strategic behaviors, namely the *Nash Gap*, is hard. In particular, they provide guarantees for BC, assuming that the equilibrium profile generating the data assigns strictly positive probability to every state. More recently, Freihaut et al. [2025] dropped this assumption and proved a tighter BC guarantee involving the *all policy* deviation concentrability coefficient $\mathcal{C}_{\max}$. On an intuitive level, $\mathcal{C}_{\max}$ quantifies the coverage only of the states that can be reached by a best response against an arbitrary policy. Their work considers two settings. In the non-interactive setting, the learner

[*] Equal contribution, alphabetical order.

receives a fixed dataset of expert trajectories and cannot query the expert further. In the interactive setting, the learner is allowed to query the expert at states encountered during training. In their work and also this work, the experts are assumed to be playing according to a Nash equilibrium strategy $(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$. For the non-interactive case, they showed that dependence on $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is unavoidable. Informally, $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ describes the coverage of states reachable under all potential Nash equilibria. If $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is infinite, then no algorithm can succeed, even with unlimited data. They removed the $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ dependence in the interactive setting, introducing MURMAIL that comes with a sample complexity of order $\mathcal{O}(\varepsilon^{-8})$ independent of $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$.

However, the authors left open several questions, which we address in this work. The first concerns the non-interactive setting and is presented next.

**Open Question 1** *Does there exist a non-interactive MAIL algorithm with guarantees featuring only $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ and not $\mathcal{C}_{\max}$ ?*

Answering this question is crucial for both theoreticians and practitioners, as it clarifies when applying BC as an algorithm is appropriate and when, instead, an interactive expert is required. Moreover, in the interactive setting, the guarantees for MURMAIL scales suboptimally with the game parameters and precision $\varepsilon$. Therefore, it is natural to ask:

**Open Question 2** *Can we design an algorithm which outputs an $\varepsilon$-approximate Nash equilibrium with the optimal order of expert queries, which is $\mathcal{O}(\varepsilon^{-2})$ ?*

In this work, we answer these questions with the following contributions:

1. We construct a Markov game where, even if $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is bounded, no non-interactive Imitation Learning algorithm can learn an $\varepsilon$-Nash equilibrium from data. Surprisingly, the construction is a striking simple Markov Games with 3 states only.

2. Additionally, with the same construction, we show that the *all policy deviation concentrability coefficient* $\mathcal{C}_{\max}$, which upper bounds $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$, is the fundamental quantity of non-interactive MAIL by proving a statistical lower bound of order $\Omega(\mathcal{C}_{\max}\varepsilon^{-2})$ on the sample complexity of any non-interactive MAIL algorithm. This construction answers **Open Question 1** in the negative. Moreover, since Freihaut et al. [2025] proved an upper bound for BC of order $\mathcal{O}\left(\mathcal{C}_{\max}^{2}\varepsilon^{2}\right)$, we conclude that BC is rate optimal in the non-interactive setting. Indeed, BC matches the optimal $\varepsilon$-dependence, and the gap between the BC upper bound and the information-theoretic lower bound is only polynomial in the concentrability and in the parameters of the game.

3. For the interactive setting, we provide a new framework, that combines reward-free exploration with (interactive) Multi-Agent Imitation Learning. This allows us to derive a new algorithm, namely *MAIL-WARM*, that improves the best currently known sample complexity of MURMAIL from $\mathcal{O}(\varepsilon^{-8})$ to $\mathcal{O}(\varepsilon^{-2})$, matching the lower bound in $\varepsilon$ in this setting. This provides a positive answer to **Open Question 2**.

4. We empirically demonstrate the effectiveness of our algorithm, showing that it outperforms other interactive Multi-Agent Imitation Learning algorithms in settings where BC fails to recover an $\varepsilon$-Nash equilibrium.

A complete summary of our results can be found in Table 1.

Table 1: For simplicity, we report results for the two-player zero-sum setting with horizon $H$, finite state space $\mathcal{S}$, finite action spaces $\mathcal{A}$, $\mathcal{B}$. Let $A_{\max} = \max\left(|\mathcal{A}|, |\mathcal{B}|\right)$. For a fair comparison, we restate the results of Freihaut et al. [2025] in the finite-horizon setting. Indeed, we have verified that the results prove therein transfer to the finite-horizon setting. The column **Expert Data** reports the number of data collected in either the interactive or non-interactive setting to attain a Nash gap bound of order $\mathcal{O}(\varepsilon)$.

| Algorithm | Expert Data | Queriable Expert |
|---|---|---|
| BC (Analysis in Freihaut et al. [2025]) | $\widetilde{\mathcal{O}}\left(\frac{H^4 S A_{\max} \mathcal{C}^2_{\max}}{\varepsilon^2}\right)$ | ✗ |
| **Lower Bound (This work)** | $\Omega\left(\frac{\mathcal{C}_{\max}}{\varepsilon^2}\right)$ | ✗ |
| MURMAIL (Freihaut et al. [2025]) | $\widetilde{\mathcal{O}}\left(\frac{H^{12} S^4 A^5_{\max}}{\varepsilon^8}\right)$ | ✓ |
| **MAIL-WARM(This work)** | $\widetilde{\mathcal{O}}\left(\frac{H^7 S^3 A^3_{\max}}{\varepsilon^2}\right)$ | ✓ |
| **Lower bound (This work)** | $\Omega\left(\varepsilon^{-2}\right)$ | ✓ |

## 2 Preliminaries

We start by formalizing the concept of two-player zero-sum Markov games.

**Two-player zero-sum Markov game** A finite-horizon two-player zero-sum Markov game is described by the tuple $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, d_0)$, where $\mathcal{S}$ is a finite state space of cardinality $S := |\mathcal{S}|$, $\mathcal{A}$ and $\mathcal{B}$ are the finite action spaces of cardinality $A := |\mathcal{A}|$ and $B := |\mathcal{B}|$ for player 1 and player 2, respectively. Moreover, let $A_{\max} = \max\{A, B\}$ be the cardinality of the largest action space. The transition dynamics at each time step $h \in \{1, \ldots, H\} := [H]$ are governed by an (unknown) transition kernel $P_h \in \mathbb{R}^{SAB \times S}$, and the reward function $r_h \in [-1, 1]^{SAB}$ assigns a scalar payoff to each state-action triplet. The game starts from an initial state $S_0 \sim d_0$, where $d_0$ is a distribution over $\mathcal{S}$.

In this setup, player 1 seeks to maximize the total reward, while player 2 aims to minimize it, leading to the zero-sum property: for any $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and all $h \in [H]$, it holds that $r^1_h(s, a, b) = -r^2_h(s, a, b)$. Hence, we omit superscripts and refer to the reward as $r$. A (stochastic) Markov policy for player 1 is denoted by $\mu_h : \mathcal{S} \times H \to \Delta_{\mathcal{A}}$, and for player 2 by $\nu_h : \mathcal{S} \times H \to \Delta_{\mathcal{B}}$, where $\Delta_{\mathcal{A}}$ and $\Delta_{\mathcal{B}}$ denote probability simplices over actions. Moreover, we will use $\Pi_\mu$, $\Pi_\nu$ to denote the set of all Markov policies for the $\mu$ and $\nu$ player respectively. We denote a policy $\mu$ of a game as a set of policies $\mu := \{\mu_h\}_{h=0}^H$ and similar for $\nu := \{\nu_h\}_{h=0}^H$.

Let $\{(S_h, A_h, B_h)\}_{h=0}^H$ denote the stochastic process induced by a policy pair $(\mu, \nu)$ acting in a Markov game $\mathcal{G}$. Then, the value function and state-action value function are defined as:

$$V_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}\left[\sum_{t=h}^H r_t(S_t, A_t, B_t) \,\middle|\, S_h = s\right],$$

$$Q_h^{\mu,\nu}(s, a, b) := \mathbb{E}_{\mu,\nu}\left[\sum_{t=h}^H r_t(S_t, A_t, B_t) \,\middle|\, S_h = s, A_h = a, B_h = b\right].$$

We also define the (unnormalized) state visitation distribution at stage $h \in [H]$ induced by the policy pair $(\mu, \nu)$ as follows:

$$d_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}\left[\mathbb{1}_{\{S_h = s\}} \,\middle|\, S_0 \sim d_0\right].$$

Fixing the policy of one agent reduces the Markov game to a Markov decision process (MDP). For example, if player 2 follows policy $\nu$, then the effective transition dynamics under $(s, a)$ are:

$$P_h^{\nu}(s' \mid s, a) := \sum_{b \in \mathcal{B}} \nu_h(b \mid s) P_h(s' \mid s, a, b).$$

A similar expression holds when $\mu$ is fixed. The best-response set for each player against a fixed policy of the other is defined as:

$$\mathrm{br}(\nu) := \arg\max_{\mu \in \Pi_\mu} \langle d_0, V_0^{\mu,\nu} \rangle.$$

Equivalently for the second player, we have $\mathrm{br}(\mu) := \arg\min_{\nu \in \Pi_\nu} \langle d_0, V_0^{\mu,\nu} \rangle$. Notice, that the best-response may not be unique, therefore $\mathrm{br}(\mu)$ and $\mathrm{br}(\nu)$ are sets in general. On the contrary, the corresponding value in zero-sum games is unique. A pair of policies $(\mu^\star, \nu^\star)$ forms a Nash equilibrium (NE) if they are best responses to each other.

To quantify how far a policy pair is from an equilibrium, we define the *Nash gap*:

$$\mathrm{Nash{-}Gap}(\mu, \nu) := \langle d_0, V_0^{\mu^\star,\nu} - V_0^{\mu,\nu^\star} \rangle.$$

This measure satisfies $\mathrm{Nash{-}Gap}(\mu, \nu) = 0$ if and only if $(\mu, \nu)$ is a NE, and is strictly positive otherwise. Another important property of zero-sum games is that the set of Nash equilibria is convex.

## 3 Setting and Main Results

In Multi-Agent Imitation Learning (MAIL), the objective is to design an algorithm Alg that, after accessing $\mathrm{poly}(\varepsilon^{-1})$ actions sampled from the expert policies, returns a pair of policies $(\widehat{\mu}, \widehat{\nu})$ such that the expected Nash gap is bounded:

$$\mathbb{E}_{\mathrm{Alg}}\left[\mathrm{Nash{-}Gap}(\widehat{\mu}, \widehat{\nu})\right] < \varepsilon. \tag{1}$$

This formulation captures the goal of learning approximately optimal behavior in competitive settings through selective expert guidance.

Furthermore, we differentiate the two following settings:

- We refer to **non-interactive** Multi-Agent Imitation Learning as the setting where a dataset is precollected from Nash equilibrium policies $(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$. In particular, states are sampled as $\{s_h^i\}_{i=1}^N \sim \rho_h \in \Delta_{\mathcal{S}}$, while actions are sampled as $\{a_h^i\}_{i=1}^N \sim \mu^{\mathrm{E}}(\cdot|s_h^i)$ and $\{b_h^i\}_{i=1}^N \sim \nu^{\mathrm{E}}(\cdot|s_h^i)$. Once the dataset is received, the learner can no further interact with the expert policies during the learning process.
- In **interactive** Multi-Agent Imitation Learning, the learner can query the expert on demand. The learning process unfolds over multiple rounds of interaction with the environment. During each round, the learner deploys a policy pair to collect a trajectory and may query the expert at any visited state.

Non-interactive MAIL has the advantage of avoiding a queriable expert, but in that setting, the theoretical bounds are worse. In contrast, interactive Imitation Learning algorithms can achieve statistical bounds independent of $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$, but this setting captures a smaller subset of real-world scenarios, as a queriable expert might not always be available.

### 3.1 Main result in non-interactive MAIL

In this section, we present our main result that solves a question left open by [Freihaut et al., 2025]. We consider the finite-horizon setting, which can be related to the discounted case through the effective horizon $H \approx \frac{1}{1-\gamma}$. In order to state the theoretical gap we need to give some context which we introduce next. We adapt the concentrability coefficient from Freihaut et al. [2025] to the finite-horizon setting. For a policy pair $\mu, \nu$ and dataset state distribution $\rho := \{\rho_h\}_{h=1}^{H}$ it is defined as

$$C(\mu, \nu) := \max \left\{ \max_{\nu^\star \in \mathrm{br}(\mu)} \max_{h \in [H]} \left\| \frac{d_h^{\mu^{\mathrm{E}}, \nu^\star}}{\rho_h} \right\|_\infty , \max_{\mu^\star \in \mathrm{br}(\nu)} \max_{h \in [H]} \left\| \frac{d_h^{\mu^\star, \nu^{\mathrm{E}}}}{\rho_h} \right\|_\infty \right\}.$$

In most common non-interactive situations, we have $\rho_h = d_h^{\mu^{\mathrm{E}}, \nu^{\mathrm{E}}}$ where $\mu^{\mathrm{E}}, \nu^{\mathrm{E}}$ is a possible Nash equilibrium profile. They showed that if $C(\mu^{\mathrm{E}}, \nu^{\mathrm{E}}) = \infty$ any non-interactive Multi-Agent Learning algorithm suffers from a Nash gap of the order of the horizon $H$. Defining $\mathcal{C}_{\max} = \max_{\mu,\nu} \mathcal{C}(\mu, \nu)$ , Freihaut et al. [2025] also presented a behavioral cloning analysis showing that $\widetilde{\mathcal{O}}\left(\mathcal{C}_{\max}^2 \varepsilon^{-2}\right)$ samples are needed to learn a policy pair $\hat{\mu}, \hat{\nu}$ achieving the learning goal given in (1). Clearly, we have that $\mathcal{C}_{\max} \geq \mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$, therefore there is a gap between the upper and the lower bound. In particular, given only the results of Freihaut et al. [2025] it is not clear if BC or another non-interactive MAIL algorithm can learn when $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is finite while $\mathcal{C}_{\max}$ is infinite. The following result closes the gap in the negative excluding the possibility that a non-interactive MAIL algorithm can avoid the dependence on $\mathcal{C}_{\max}$ in its sample complexity.

**Theorem 3.1.** *Let $\hat{\mu}, \hat{\nu}$ be the output of a non-interactive MAIL algorithm* Alg*. Then, for any* Alg*, there exists a Markov game such that satisfying $\mathbb{E}_{\mathrm{Alg}}\left[\left\langle d_0, V^{\mu^\star, \widehat{\nu}} - V^{\widehat{\mu}, \nu^\star}\right\rangle\right] \leq \mathcal{O}(\varepsilon)$ requires an expert dataset of size $N = \Omega(\frac{\mathcal{C}_{\max}}{\varepsilon^2})$.*

The same construction gives the following corollary for unbounded $\mathcal{C}_{\max}$.

**Corollary 3.1.** *For any non-interactive MAIL algorithm* Alg*, there exists a Markov game $\mathcal{G}$ with $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}}) < \infty$ and $\mathcal{C}_{\max} = \infty$ where $\mathbb{E}_{\mathrm{Alg}}\left[\left\langle d_0, V^{\mu^\star, \widehat{\nu}} - V^{\widehat{\mu}, \nu^\star}\right\rangle\right] \geq \frac{H-1}{60}$.*

The lower bound therefore excludes the existence of non-interactive MAIL algorithms improving the dependence on $\varepsilon$ on the sample complexity bound for BC and that avoids the dependence on $\mathcal{C}_{\max}$. Moreover, Corollary 3.1 shows that it is possible to construct games with small $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ but unbounded $\mathcal{C}_{\max}$. In this regime, the results of Freihaut et al. [2025] cannot characterize the behavior of BC, whereas our results predict that learning in a non-interactive setting with unbounded $\mathcal{C}_{\max}$ is not possible.

### 3.2 Main result in interactive MAIL

This section presents our main result for the interactive MAIL setting, focusing on our new algorithm MAIL-WARM (see Algorithm 1). To provide context, Freihaut et al. [2025] introduced MURMAIL, the first interactive MAIL algorithm with theoretical guarantees. Their analysis shows that avoiding the concentrability coefficient $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ requires a queriable expert. While this enables effective minimization of the Nash Gap, it comes at the cost of $\mathcal{O}(\varepsilon^{-8})$ expert queries.

This raises a gap: when the concentrability coefficient is small and bounded, non-interactive imitation learning algorithms can outperform existing interactive methods. To close this gap, we introduce MAIL-WARM (Algorithm 1), which reduces the required number of expert queries from $\mathcal{O}(\varepsilon^{-8})$ to $\mathcal{O}(\varepsilon^{-2})$ by leveraging a reward-free warm-up phase. As a result, the sample complexity of interactive MAIL becomes comparable to the Behavior Cloning upper bound, while entirely removing dependence on the concentrability coefficient.

Since each trajectory collection queries the expert $H$ times, the total number of expert queries is bounded by $\mathcal{O}(NH) = \mathcal{O}\Big(\frac{H^7 S^3 A^2 B \log(S/\delta_{\text{fail}})}{\varepsilon^2}\Big)$. In contrast to MAIL-BRO [Freihaut et al., 2025], our approach requires no additional assumptions such as access to a best-response oracle, which takes an input the policy of one player, and outputs a best responding policy for the other player. Up to problem-dependent factors in $H, S, A$, the rate matches that of Behavior Cloning but crucially avoids reliance on the concentrability coefficient, which can be unbounded and cause non-interactive methods to fail. Thus, MAIL-WARM improves the best known guarantees by an order of $\varepsilon^{-6}$. Moreover, by adapting the construction from Theorem 3.1, we show that the rate $\mathcal{O}(\varepsilon^{-2})$ is optimal. The following theorem states our main guarantee for MAIL-WARM.

**Theorem 3.2.** *For any $\varepsilon > 0$ and $\delta_{\text{fail}} \in (0,1)$, if we execute Algorithm 1 and choose the parameters according to $N = \mathcal{O}(\frac{H^6 S^3 A^3_{\max} \log(S/\delta_{\text{fail}})}{\varepsilon^2})$ and $N_0 \geq \mathcal{O}\left(S^3 A^2_{\max} H^6 \iota_0^3/\varepsilon\right)$, we get with probability $1 - \delta_{\text{fail}}$ for the policies $(\widehat{\mu}, \widehat{\nu})$ that*

$$\text{Nash}-\text{Gap}(\widehat{\mu}, \widehat{\nu}) \leq \mathcal{O}(\varepsilon).$$

This result establishes a sample complexity guarantee of $\mathcal{O}(\varepsilon^{-2})$, which matches our lower bound and is therefore rate-optimal in the sense that the dependence on $\varepsilon$ cannot be improved. Consequently, MAIL-WARM achieves the optimal sample complexity in the interactive setting, closing the remaining theoretical gap in interactive MAIL.

# 4 Lower bound

We now provide intuition for our lower bound result. The key insight is that the boundedness of the concentrability coefficient $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is not sufficient to guarantee that a non-interactive imitation learning algorithm outputs an $\varepsilon$-Nash equilibrium.

The reason is that bounded $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ only guarantees that all states that can be visited by *rational* deviations have positive probability under the dataset distribution $\rho$. By rational deviation, we mean that the opponent might choose to act only according to policies which are guaranteed not to be exploited by a Nash equilibrium. However, we show that the opponent could exploit the output of a non-interactive MAIL algorithm deviating *irrationaly*, in the sense that such deviation would be exploitable by a Nash profile. Bounded $\mathcal{C}_{\max}$ imposes that also the states reachable by irrational deviations are covered by the sampling distribution $\rho$.

To illustrate this, consider the Markov game instance from Theorem 3.1, depicted in Figure 1. Both agents start in state $s_1$, where agent 2 controls the transition: action $b_1$ leads to $s_2$, while action $b_2$ leads to $s_3$, regardless of agent 1's choice. The rewards are $r(s_1) = r(s_2) = 0$, while in $s_3$ the agents play a normal-form game with Nash value of 1. Importantly, in equilibrium, both players visit only $s_1$ and $s_2$ when acting according to a NE, which ensures that $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is bounded. However, since the learner never observes equilibrium play in $s_3$, it cannot recover the correct Nash strategy for agent 1 in that state. This allows agent 2 to deviate *irrationaly* by choosing $b_2$ and exploit the learner in $s_3$.

The lower bound is established with well-established information techniques (see e.g. [Lattimore and Szepesvári, 2020]) by constructing a family of two Markov games in which the payoff structure in $s_3$ corresponds to different versions of Matching Pennies, with the unique Nash value differing by $\varepsilon$. Distinguishing these two games requires $\Omega(\varepsilon^{-2})$ samples. Since $s_3$ is visited only $\mathcal{C}_{\max}$ many times under the expert equilibrium, a non-interactive learner cannot collect enough data to resolve this ambiguity, which yields the lower bound in Theorem 3.1. A detailed proof is provided in Appendix B.

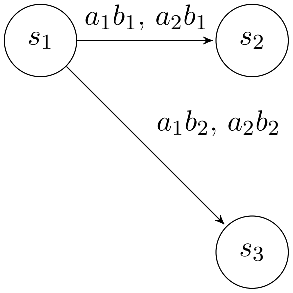


Figure 1: Markov game instance used for the Lower bound.

# 5 Interactive MAIL algorithm

In this section, we describe our proposed algorithm MAIL-WARM in detail. The algorithmic pseudocode can be found in Algorithm 1. In the first phase, we loop over states and stages. For each $s, h \in \mathcal{S} \times [H]$, we instantiate a reward-maximization problem in which the only nonzero, positive reward is obtained by reaching state $s$ after $h$ stages in the MDP induced by fixing the $\nu$-player to follow the Nash equilibrium policy $\nu^{\mathrm{E}}$. At line 6, this RL problem is solved via EULER [Zanette and Brunskill, 2019], which outputs a sequence of $N_0$ policies satisfying a first-order regret bound. At line 8, we set the policies at all stages after $h$ to be uniform. This choice is arbitrary, the effective reward instance has an horizon of length $h$ and the policy can be arbitrary set after step $h$ without changing the regret properties. At the end, of the initial for loop we obtain a policy set $\Psi^{\nu^{\mathrm{E}}}$ containing $SHN_0$ policies. Next, we can exploit this set to construct an exploratory dataset $\mathcal{D}^{\nu^{\mathrm{E}}}$ of $N$ state action sequences in the Markov game. A trajectory is collected fixing the $\nu$-player to the strategy $\nu^{\mathrm{E}}$ while the $\mu$ player acts according to a policy sampled uniformly at random from the set $\Psi^{\nu^{\mathrm{E}}}$. Repeating the sampling process for choosing the policy of the player $\mu$, $N$ times, gives the desired dataset. Switching the player that remains fixed and reapplying the same procedure, we obtain the exploratory dataset $\mathcal{D}^{\mu^{\mathrm{E}}}$ for the single player MDP induced by fixing the $\mu$-player to

**Algorithm 1** Multi-Agent Imitation Learning with reward-free warm-up (MAIL-WARM)

```
1: **Input:** iteration number $N_0$, $N$, queriable expert $\nu^{\mathrm{E}}$.
2: **Reward-free warm-up phase:**
3: set policy class $\Psi^{\nu^{\mathrm{E}}} \leftarrow \emptyset$, and dataset $\mathcal{D} \leftarrow \emptyset$.
4: **for** all $(s, h) \in \mathcal{S} \times [H]$ **do**
5:     $r_h^{\nu^{\mathrm{E}}}(s', a') \leftarrow \mathbf{1}[s' = s \text{ and } h' = h]$ for all $(s', a', h') \in \mathcal{S} \times \mathcal{A} \times [H]$.
6:     $\left\{\pi_i^{(s,h)}\right\}_{i=1}^{N_0} \leftarrow \mathrm{EULER}(r^{\nu^{\mathrm{E}}}, N_0, P^{\nu^{\mathrm{E}}})$.
7:     Let $\Phi^{(s,h)} \leftarrow \left\{\pi_i^{(s,h)}\right\}_{i=1}^{N_0}$
8:     $\mu_{h'}(\cdot|s) \leftarrow \mathrm{Unif}(\mathcal{A}), \ \forall \mu \in \Phi^{(s,h)}, \forall h' \geq h$.
9:     $\Psi^{\nu^{\mathrm{E}}} \leftarrow \Psi^{\nu^{\mathrm{E}}} \cup \Phi^{(s,h)}$.
10: **end for**
11: **for** $n = 1 \ldots N$ **do**
12:     sample policy $\mu \sim \mathrm{Unif}(\Psi^{\nu^{\mathrm{E}}})$.
13:     Collect $z_n = (s_1, a_1, b_1, \ldots, s_{H+1}) \sim \mu, \nu^{\mathrm{E}}$.
14:     $\mathcal{D}^{\nu^{\mathrm{E}}} \leftarrow \mathcal{D}^{\nu^{\mathrm{E}}} \cup \{z_n\}$
15: **end for**
16: **Repeat**: the reward-free warm-up phase fixing $\mu^{\mathrm{E}}$ for fixed $\mu^{\mathrm{E}}$ generating the policy set $\Psi^{\mu^{\mathrm{E}}}$
    and the dataset $\mathcal{D}^{\mu^{\mathrm{E}}}$.
17: **Receive:** datasets $(\mathcal{D}^{\mu^{\mathrm{E}}}, \mathcal{D}^{\nu^{\mathrm{E}}})$.
18: **Imitation Learning**
19: Use dataset $\mathcal{D}^{\nu^{\mathrm{E}}}$ to compute

        $\hat{\nu} = \operatorname*{argmin}_{\nu \in \Pi_\nu} \sum_{s,b \in \mathcal{D}^{\nu^{\mathrm{E}}}} -\log \nu(b|s)$

20: Use dataset $\mathcal{D}^{\mu^{\mathrm{E}}}$ to compute

        $\hat{\mu} = \operatorname*{argmin}_{\mu \in \Pi_\mu} \sum_{s,a \in \mathcal{D}^{\mu^{\mathrm{E}}}} -\log \mu(a|s)$

21: **Return** Nash estimate $(\widehat{\mu}, \widehat{\nu})$.
```

the policy $\mu^{\mathrm{E}}$. Finally, the last step is to use the datasets $\mathcal{D}^{\mu^{\mathrm{E}}}$ and $\mathcal{D}^{\nu^{\mathrm{E}}}$ in lines 19 and 20 to apply behavioral cloning over the state-action pairs in these datasets.

All in all, our new algorithm is based on the intuition of using a reward-free routine to design an exploratory dataset over which we can prove benign bounds for behavioral cloning which are actually independent of the concentrability factor. This is without contradiction with our lower bound since in the dataset construction phase we used interaction with the expert.

Compared to *MURMAIL* and *MAIL-BRO* by Freihaut et al. [2025], our new algorithm requires neither a best response oracle nor to compute the maximum uncertainty response by solving a RL inner-loop at every iteration. In particular, the maximum uncertainty response in MURMAIL required to compute at each iteration $k \in [K]$ an RL problem with reward $\left\|\nu_k(\cdot|s) - \nu^{\mathrm{E}}(\cdot|s)\right\|^2$. In comparison, MAIL-WARM solves only $SH$ RL problems with rewards independent of a particular policy $\nu_k$. Since the number of RL problems no longer depends on $K$, MAIL-WARM achieves better guarantees.

This intuition is formalized in the next section, which presents the key steps of the proof of Theorem 3.2.

# 6 Analysis

The goal of this section is to prove our main result stated in Theorem 3.2. We begin by presenting general results on the reward-free algorithm of Jin et al. [2020], applied to the single-player MDPs that arise when one of the two players is fixed to the queriable expert policy. Then, we show how analyzing Behavior Cloning on the dataset generated during the reward-free warm-up phase can be used to effectively minimize the Nash gap.

## 6.1 Phase 1: Reward-free Warm-up

As evident from the algorithm presentation, the core concept of the analysis is the single-agent MDP induced by fixing one of the two players to the policy of the queriable expert.

**Definition 6.1** (Expert Induced MDP). *Let $\mathcal{G}$ be a Markov game and $\mu^{\mathrm{E}}, \nu^{\mathrm{E}}$ the expert policies, then $\mathcal{M}^{\nu^{\mathrm{E}}} := (\mathcal{S}, \mathcal{A}, P^{\nu^{\mathrm{E}}}, r^{\nu^{\mathrm{E}}}, H)$ is the MDP induced by the expert $\nu^{\mathrm{E}}$ with the transition model $P^{\nu^{\mathrm{E}}} = \left\{P_h^{\nu^{\mathrm{E}}}\right\}_{h=1}^{H}$ with $P_h^{\nu^{\mathrm{E}}}(s' \mid s, a) := \sum_{b \in \mathcal{B}} \nu_h^{\mathrm{E}}(b \mid s) P_h(s' \mid s, a, b)$ and an arbitrary reward function $r^{\nu^{\mathrm{E}}} = \left\{r_h^{\nu^{\mathrm{E}}}\right\}_{h=1}^{H}$, such that $r_h^{\nu^{\mathrm{E}}}(s,a) \in \{0,1\} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Analogously we define $\mathcal{M}^{\mu^{\mathrm{E}}}$ as the MDP induced by the expert policy $\mu^{\mathrm{E}}$.*

Notice that the reward function in the induced expert MDP is not related to the true unknown reward function of the original Markov game, instead it is used for exploration purposes as shown in Algorithm 1. With this definition, we can state a fundamental result about the reward-free warm-up phase. In particular, Theorem 6.1 states that using the distributions $p_h^{\mu^{\mathrm{E}}}(s,b) := (N_0 SH)^{-1} \sum_{\nu \in \Psi^{\mu^{\mathrm{E}}}} d_h^{\mu^{\mathrm{E}},\nu}(s)\nu(b|s)$ and $p_h^{\nu^{\mathrm{E}}}(s,a) := (N_0 SH)^{-1} \sum_{\mu \in \Psi^{\nu^{\mathrm{E}}}} d_h^{\mu,\nu^{\mathrm{E}}}(s)\mu(a|s)$ to generate the dataset (that is using them as distribution $\rho$) guarantees a benign bound on the coverage over the set of $\delta$-reachable states of any possible occupancy measure in the induced MDPs. By coverage, we mean the ratio $\max_{a,h} d_h^{\mu,\nu^{\mathrm{E}}}(s,a)/p_h^{\nu^{\mathrm{E}}}(s,a)$ which at an intuitive level can be thought at the average number of times one needs to sample from $p^{\nu^{\mathrm{E}}}$ before hitting a state action pair which has probability $d^{\mu,\nu^{\mathrm{E}}}(s,a)$ under the policy $\mu$.

Technically, the result is obtained applying Theorem 3.3 of Jin et al. [2020] twice in the induced MDPs $\mathcal{M}^{\nu^{\mathrm{E}}}$ and $\mathcal{M}^{\mu^{\mathrm{E}}}$.

**Theorem 6.1.** *Let $\mathcal{M}^{\nu^{\mathrm{E}}}$ be the induced MDP defined in Definition 6.1 and the policy $\Psi^{\nu^{\mathrm{E}}}$ set generated according to Algorithm 1. Then there exists an absolute constant $c > 0$ such that for any $\varepsilon > 0$ and $p \in (0,1)$, if we set $N_0 \geq cS^2AH^4\iota_0^3/\delta$, where $\iota_0 := \log(SAH/p\delta)$, then with probability*

$1-p$, *the reward free exploration returns a sampling distribution* $p^{\nu^{\mathrm{E}}}$ *such that for all* $\mu \in \Pi_\mu$

$$\forall \delta - \mathrm{significant}(s,h), \quad \max_{a,h} \frac{d_h^{\mu,\nu^{\mathrm{E}}}(s,a)}{p_h^{\nu^{\mathrm{E}}}(s,a)} \leq 2SAH,$$

*where* $(s,h)$ *is* $\delta - \mathrm{significant}$, *if the probability to reach a state* $s$ *at step* $h$ *in the induced MDP* $\mathcal{M}^{\nu^{\mathrm{E}}}$ *is lower bounded by* $\delta$*:*

$$\max_{\mu \in \Pi_\mu} d_h^{\mu,\nu^{\mathrm{E}}}(s) \geq \delta.$$

*Analogously, we get the same result for* $\delta$*-reachable states in* $\mathcal{M}^{\mu^{\mathrm{E}}}$ *with sampling distribution* $p^{\mu^{\mathrm{E}}}$*.*

For convenience, we will denote the set of $\delta$-significant states in the MDP $\mathcal{M}^{\nu^{\mathrm{E}}}$ as $\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}$. For the other player, $\mathcal{S}_{\delta,h}^{\mu^{\mathrm{E}}}$ is defined analogously.

In reward-free RL this dataset is taken to enable learning against any reward function. Here, instead we find a new interesting use case of the warm-up phase which is to use the generated datasets $\mathcal{D}^{\mu^{\mathrm{E}}}$ and $\mathcal{D}^{\nu^{\mathrm{E}}}$ for running behavioral cloning.

In particular, we can perform a change of measure to avoid the need of sampling states according to the occupancy measure of the best response policy which cannot be computed. Given the good coverage properties of the distributions $p^{\nu^{\mathrm{E}}}$ and $p^{\mu^{\mathrm{E}}}$, the change of measure creates an increase in the number of needed samples of at most $SHA_{\max}$. Since the increase is independent on the desired accuracy $\varepsilon$ we can retain the optimal $\mathcal{O}(\varepsilon^{-2})$ rate.

Next, we can see how the coverage property in Theorem 6.1 enables to prove Theorem 3.2.

### 6.2 Phase 2: BC over the datasets generated in Phase 1

In the last section, we have seen how one can construct datasets $\mathcal{D}^{\nu^{\mathrm{E}}}$ and $\mathcal{D}^{\mu^{\mathrm{E}}}$ using ideas from reward-free RL. We will use the coverage property of these datasets on top of the following exploitability decomposition.

**Lemma 6.1 (Exploitability decomposition).** *For any policy pair* $\nu, \nu'$*, we define their total variation at state* $s$ *as* $\mathrm{TV}(\nu,\nu')(s) = \sum_{b\in\mathcal{B}} |\nu(b|s) - \nu'(b|s)|$ *. It holds that*

$$\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\widehat{\mu},\nu^\star} \right\rangle \leq 2H \sum_{h=1}^{H} \sum_{\pi\in\{\hat{\mu},\hat{\nu}\}} \mathrm{Err}_h(\pi).$$

*where we have defined* $\mathrm{Err}_h(\hat{\nu}) := \max_{\mu_h \in \mathrm{br}(\widehat{\nu}_h)} \mathbb{E}_{s\sim d_h^{\mu,\nu^{\mathrm{E}}}} \left[\mathrm{TV}\left(\nu_h^{\mathrm{E}}, \widehat{\nu}_h\right)(s)\right]$ *for player 1 and additionally* $\mathrm{Err}_h(\hat{\mu}) := \max_{\nu_h \in \mathrm{br}(\widehat{\mu}_h)} \mathbb{E}_{s\sim d_h^{\mu^{\mathrm{E}},\nu}} \left[\mathrm{TV}\left(\mu_h^{\mathrm{E}}, \widehat{\nu}_h\right)(s)\right]$ *for player 2.*

Now, we can see that the expectation in the definition of $\mathrm{Err}(\hat{\nu})$ depends on $\mu, \nu^{\mathrm{E}}$, where $\mu \in \mathrm{br}(\widehat{\nu})$ is one best response to the estimated expert. Bounding this term is challenging because the reward function is unknown, so $\mathrm{br}(\widehat{\nu})$ can neither be computed nor bounded using optimistic estimators. Equivalent considerations hold for $\mathrm{Err}(\hat{\mu})$. However, we can here make use of the reward-free warm-up phase. In particular, we can use the constructed distribution $p_h^{\nu^{\mathrm{E}}}$ to perform the following change of measure, and rewrite $\mathrm{Err}_h(\hat{\nu})$ in the following way

$$\mathrm{Err}_h(\hat{\nu}) = \underbrace{\sum_{s\in\mathcal{S}_\delta}\sum_{a\in\mathcal{A}} w_h^{\widehat{\nu}}(s,a)p_h^{\nu^{\mathrm{E}}}(s,a)\mathrm{TV}\left(\nu_h^{\mathrm{E}},\widehat{\nu}_h\right)(s)}_{:=\mathrm{Err}_h(\hat{\nu};\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}})} + \underbrace{\sum_{s\notin\mathcal{S}_\delta}\sum_{a\in\mathcal{A}} \max_{\mu\in\mathrm{br}(\hat{\nu})} d_h^{\mu,\nu^{\mathrm{E}}}(s,a)\mathrm{TV}\left(\nu_h^{\mathrm{E}},\widehat{\nu}_h\right)(s)}_{:=\mathrm{Err}_h(\hat{\nu};\bar{\mathcal{S}}_{\delta,h}^{\nu^{\mathrm{E}}})},$$

where we introduce the importance weight correction $w_h^{\widehat{\nu}}(s,a) := \max_{\mu\in\mathrm{br}(\widehat{\nu})} \frac{d_h^{\mu,\nu^{\mathrm{E}}}(s,a)}{p_h^{\nu^{\mathrm{E}}}(s,a)}$ and we split the sum over the $\delta$-significant states set $\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}$ from the sum over the not significant states set denoted by $\bar{\mathcal{S}}_{\delta,h}^{\nu^{\mathrm{E}}} := \mathcal{S}\setminus\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}$. At this point, we can bound $\mathrm{Err}_h(\hat{\nu};\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}})$ using Theorem 6.1 which implies that for any $s\in\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}$ it holds that $\max_{h\in[H],a\in\mathcal{A}} w_h^{\hat{\nu}}(s,a)\leq 2SAH$. In this way, we obtain

$$\mathrm{Err}_h(\hat{\nu};\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}) \leq 2SAH\mathbb{E}_{s\sim p_h^{\nu^{\mathrm{E}}}}\left[\mathrm{TV}\left(\nu_h^{\mathrm{E}},\widehat{\nu}_h\right)(s)\right].$$

This term is now easy to bound since $\hat{\nu}$ is computed as the minimizer of the log loss over the dataset $\mathcal{D}^{\nu^{\mathrm{E}}}$ which is sampled from the distribution $p_h^{\nu^{\mathrm{E}}}$. A standard concentration inequality for the log loss minimizer (see Lemma C.2) guarantees that with probability at least $1-\delta_{\mathrm{fail}}/2$

$$\mathrm{Err}_h(\hat{\nu};\mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}) \leq 8SAH\sqrt{\frac{SB\log(4S/\delta_{\mathrm{fail}})}{N}}.$$

Using the same analysis switching the roles of the players ensure that

$$\mathrm{Err}_h(\hat{\mu};\mathcal{S}_{\delta,h}^{\mu^{\mathrm{E}}}) \leq 8SBH\sqrt{\frac{SA\log(4S/\delta_{\mathrm{fail}})}{N}} \quad \text{w.p.} \quad 1-\frac{\delta_{\mathrm{fail}}}{2}.$$

Finally, by definition of the set $S_\delta$, it follows that the error term contribution coming from the non significant states can be bounded as

$$\mathrm{Err}_h(\hat{\mu};\bar{\mathcal{S}}_{\delta,h}^{\mu^{\mathrm{E}}}) \leq SA\delta, \quad \mathrm{Err}_h(\hat{\nu};\bar{\mathcal{S}}_{\delta,h}^{\nu^{\mathrm{E}}}) \leq SB\delta.$$

All in all, we obtain that with probability at least $1-\delta_{\mathrm{fail}}$ it holds that $\left\langle d_0, V^{\mu^\star,\hat{\nu}} - V^{\hat{\mu},\nu^\star}\right\rangle$ can be bounded by

$$\mathcal{O}\left(\sqrt{\frac{S^3A_{\max}^3H^6\log\left(4S/\delta_{\mathrm{fail}}\right)}{N}} + SA_{\max}\delta\right).$$

Therefore, selecting $\delta = \mathcal{O}(\varepsilon/(SA_{\max}H^2))$ and $N = \mathcal{O}\left(\frac{S^3A_{\max}^3H^6\log(4S/\delta_{\mathrm{fail}})}{\varepsilon^2}\right)$ proves Theorem 3.2. In Appendix C.4 we show that our approach translates to the n-player general-sum setting smoothly without incurring a sample complexity exponential in the number of players.

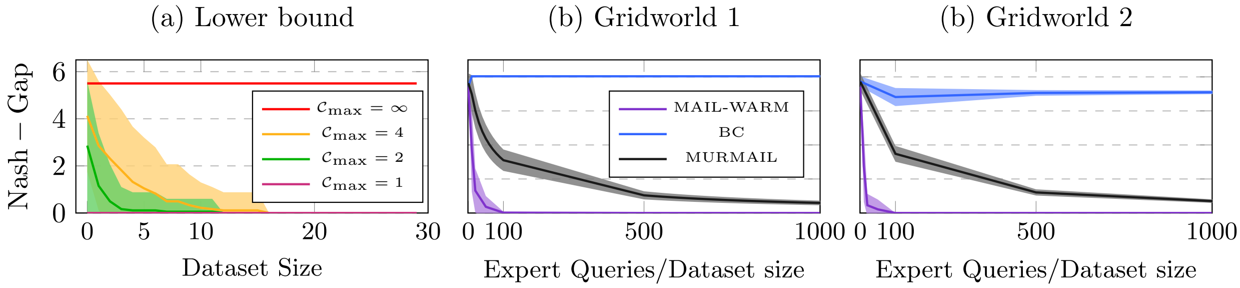
**On the rate optimality for MAIL-WARM** The rate optimality of MAIL-WARM follows from an even simplier construction compared to Figure 1. In particular, consider a single state Markov game where the payoff matrices are given by the perturbed Matching Pennies game used in $s_3$ of Figure 1. We can establish with the exact same proof of Theorem 3.1 that $\Omega(\varepsilon^{-2})$ expert queries are absolutely necessary even in interactive MAIL.

# 7 Numerical verifications

In this section, we provide some numerical verifications supporting our theoretical analysis. First, we consider the Markov game used for the lower bound (see Figure 1). Additionally, we compare BC, *MURMAIL* and *MAIL-WARM* in a zero-sum Gridworld. The 3x3 Gridworld, visualized in Figure 4, has 72 states, where each state indicates the position of both players. Both agents cannot be at the same position. In the top right corner of the Gridworld, there is a goal state. If an agent reaches the goal before the other agent does, this agent receives a reward of 1, and the other agent gets a reward of $-1$. Gridworld 1 and 2 refer to the same environment described above, the only difference is that in Gridworld 2 we consider observing data from a mixed Nash created taking convex combination of multiple Nash equilibria. Therefore, the dataset state coverage in Gridworld 2 is better. For these created games we run a Nash equilibrium solver, zero-sum Value Iteration [Perolat et al., 2015], to obtain the Nash expert policies $(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$. We repeat this for every game 10 times across varying seeds. To plot the performance, we compute the exploitability of the estimated experts for different sample sizes of the sampled states from the obtained dataset, i.e. the expert queries. We plot the average obtained exploitability across games and trials as well as the standard deviation for the different sizes. The code used for the experiments is available at https://github.com/emanuelenevali/MAIL_WARM.

In Figure 2, we quantify the expected behavior for BC in the hard instance used for our lower bound. We observe that BC's performance degrades as the state coverage decreases and that BC is not able to learn if $\mathcal{C}_{\max}$ is infinite, even though $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is constant in all considered environments. This verifies that $\mathcal{C}_{\max}$ and does not $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ predict the performance of BC and any other non-interactive MAIL algorithm.

Interestingly, we observe that BC suffers from a constant Nash Gap in both described Gridworld set-ups. In picture $(b)$, the Nash Gap of BC is higher compared to picture $(c)$ as the expert data has better coverage. In contrast to BC, we observe that for MURMAIL and MAIL-WARM, the exploitability decreases with the amount of expert queries and is independent of the expert coverage. However, MAIL-WARM outperforms MURMAIL requiring significantly fewer expert queries to minimize the Nash Gap.


(a) Lower bound

(b) Gridworld 1

(b) Gridworld 2

Figure 2: Exploitability of BC in the lower-bound Markov game and comparison of imitation learning algorithms in Gridworlds for a pure NE expert (Gridworld 1) and a mixed one (Gridworld 2).

# 8 Conclusion and future directions

In this work, we resolved **Open Question 1** for the non-interactive setting and **Open Question 2** for interactive MAIL. For both cases, we established rate-optimal results, showing that the dependence on $\varepsilon$ cannot be improved.

Our analysis leaves several important directions open. First, while we have closed the gap in $\varepsilon$-dependence, optimal guarantees with respect to other problem parameters $S$, $A$, and $H$ remain unknown.

Second, our results are developed in the finite-horizon setting. Extending them to the infinite-horizon case is non-trivial, since the reward-free warm-up phase relies on regret guarantees of the EULER algorithm that are specific to finite horizons (see Section 8 in Zanette and Brunskill [2019]). Whether analogous results can be obtained in the infinite-horizon regime remains an open challenge, and progress in this direction could be of interest independent of MAIL.

Third, we provided a Markov game (Figure 1), where $\mathcal{C}_{\max}$ can be computed explicitly. However, even in simple environments such as Gridworld, this calculation becomes challenging. As $\mathcal{C}_{\max}$ is the key fundamental quantity describing the hardness of non-interactive MAIL, developing efficient methods for calculating it would be highly valuable. In particular this would help to clarify the practical feasibility of behavioral cloning in standard MARL benchmarks.

Finally, a next step is to extend our results to the non-tabular case and to provide deep MAIL algorithms building on our framework.

# References

D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18, 10 2012. doi: 10.1214/ECP.v18-2359.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

T. V. Bui, T. Mai, and T. H. Nguyen. Mimicking to dominate: Imitation learning strategies for success in multiagent competitive games, 2023. URL `https://arxiv.org/abs/2308.10188`.

T. V. Bui, T. Mai, and T. H. Nguyen. Inverse factorized soft q-learning for cooperative multi-agent imitation learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 27178–27206. Curran Associates, Inc., 2024.

T. V. Bui, T. Mai, and H. T. Nguyen. Misodice: Multi-agent imitation from unlabeled mixed-quality demonstrations, 2025. URL `https://arxiv.org/abs/2505.18595`.

Q. Cui and S. S. Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances* *in Neural Information Processing Systems*, volume 35, pages 11739–11751. Curran Associates, Inc., 2022.

D. J. Foster, A. Block, and D. Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *arXiv preprint arXiv:2407.15007*, 2024.

T. Freihaut, L. Viano, V. Cevher, M. Geist, and G. Ramponi. Learning equilibria from data: Provably efficient multi-agent imitation learning, 2025. URL `https://arxiv.org/abs/2505.17610`.

L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak, T. A. Han, E. Hughes, V. Kovařík, J. Kulveit, J. Z. Leibo, C. Oesterheld, C. S. de Witt, N. Shah, M. Wellman, P. Bova, T. Cimpeanu, C. Ezell, Q. Feuillade-Montixi, M. Franklin, E. Kran, I. Krawczuk, M. Lamparth, N. Lauffer, A. Meinke, S. Motwani, A. Reuel, V. Conitzer, M. Dennis, I. Gabriel, A. Gleave, G. Hadfield, N. Haghtalab, A. Kasirzadeh, S. Krier, K. Larson, J. Lehman, D. C. Parkes, G. Piliouras, and I. Rahwan. Multi-agent risks from advanced ai. Technical Report 1, Cooperative AI Foundation, 2025.

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

A. K. Jain, V. Mohta, S. Kim, A. Bhardwaj, J. Ren, Y. Feng, S. Choudhury, and G. Swamy. A smooth sea never made a skilled `SAILOR`: Robust imitation via learning to search, 2025. URL `https://arxiv.org/abs/2506.05294`.

C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

S. M. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002. URL `https://api.semanticscholar.org/CorpusID:31442909`.

E. Kaufmann, P. Ménard, O. Darwiche Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 16–19 Mar 2021. URL `https://proceedings.mlr.press/v132/kaufmann21a.html`.

T. Lattimore and C. Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

H. M. Le, Y. Yue, P. Carr, and P. Lucey. Coordinated multi-agent imitation learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1995–2003. JMLR.org, 2017.

G. Li, Y. Yan, Y. Chen, and J. Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning, 2024. URL `https://arxiv.org/abs/2304.07278`.

P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko. Fast active learning for pure exploration in reinforcement learning, 2020. URL `https://arxiv.org/abs/2007.13442`.

M. J. Osborne and A. Rubinstein. *A course in game theory.* MIT press, 1994.

J. Perolat, B. Scherrer, B. Piot, and O. Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1321–1329, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/perolat15.html.

N. Rajaraman, L. Yang, J. Jiao, and K. Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

G. Ramponi, P. Kolev, O. Pietquin, N. He, M. Lauriere, and M. Geist. On imitation in mean-field games. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 40426–40437. Curran Associates, Inc., 2023.

A. Rubinstein. Settling the complexity of computing approximate two-player nash equilibria, 2016. URL https://arxiv.org/abs/1606.04550.

J. Song, H. Ren, D. Sadigh, and S. Ermon. Multi-agent generative adversarial imitation learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/240c945bb72980130446fc2b40fbb8e0-Paper.pdf.

J. Tang, G. Swamy, F. Fang, and Z. S. Wu. Multi-agent imitation learning: Value is easy, regret is hard. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 27790–27816. Curran Associates, Inc., 2024.

F. Torabi, G. Warnell, and P. Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019.

L. Viano, S. Skoulakis, and V. Cevher. Imitation learning in discounted linear MDPs without exploration assumptions. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DChQpB4AJy.

A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22430–22456. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wagenmaker22b.html.

T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai. Policy finetuning: bridging sample-efficient offline and online reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

L. Yu, J. Song, and S. Ermon. Multi-agent adversarial inverse reinforcement learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7194–7201. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/yu19e.html`.

A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds, 2019. URL `https://arxiv.org/abs/1901.00210`.

J. Zhang, W. Zhang, and Q. Gu. Optimal horizon-free reward-free exploration for linear mixture mdps, 2024. URL `https://arxiv.org/abs/2303.10165`.

# A Related Work

The related works can be split into into Multi-Agent Imitation Learning and reward-free Reinforcement Learning.

**Multi-Agent Imitation Learning** Most of the works in Multi-Agent Imitation Learning are on the empirical side. Some works consider the cooperative setting, meaning that each agents maximizes the same reward function [Bui et al., 2024, Le et al., 2017, Bui et al., 2025, 2023]. This line of work does not necessitate any equilibrium solution framework, which is essential for our work. Yu et al. [2019] or Song et al. [2018] extend the framework of adversarial imitation learning to the multi-agent setting. In particular, Song et al. [2018] extend GAIL [Ho and Ermon, 2016] to the multi-agent setting. While the authors consider Nash equilibrium experts, their theoretical results require a unique Nash equilibrium solution, which rarely holds in Markov games. Yu et al. [2019] consider inverse reinforcement learning and introduce regularization to make the Nash equilibrium, technically not a Nash equilibrium anymore, unique. In this work, we consider Nash equilibrium experts, the most common solution concept in Markov games, without any additional assumption.
In the context of mean-field games, Ramponi et al. [2023] were the first to study imitation learning through the lens of the Nash Gap. For general $n$-player games, the seminal work of Tang et al. [2024] established fundamental hardness results for minimizing the Nash Gap. They provided guarantees for behavioral cloning (BC) under the assumption that the Nash equilibrium policy profile generating the data visits every state with positive probability. The closest line of work to ours is that of Freihaut et al. [2025], who extended these results in several directions. First, they showed that the strict coverage assumption of Tang et al. [2024] can be dropped, and instead provided a tighter BC guarantee in terms of the all-policy deviation concentrability coefficient $\mathcal{C}_{\max}$. Their analysis also established a fundamental separation between two settings. In the non-interactive setting, where the learner only has access to a fixed dataset of expert trajectories, they proved that the dependence on $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is unavoidable. In particular, if $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is infinite, then no non-interactive algorithm can succeed, even with unlimited data and known transitions. In the interactive setting, they introduced the first algorithm with sample complexity guarantees, *MURMAIL*, which achieves $\tilde{\mathcal{O}}(\varepsilon^{-8})$ queries, independent of $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$. Our work sharpens these results in both regimes. In the non-interactive case, we close the theoretical gap by identifying the precise dependence on $\mathcal{C}_{\max}$ and proving that behavioral cloning is rate-optimal. In the interactive case, we propose a new framework that reduces the query complexity to $\tilde{\mathcal{O}}(\varepsilon^{-2})$, matching the dependence on $\varepsilon$ implied by our lower bound.

**Reward-free Reinforcement Learning** In their seminal work, Jin et al. [2020] introduced the framework of reward-free reinforcement learning in the single-agent setting. The central idea is to construct a dataset without knowledge of the reward function that provides sufficient coverage of the environment so that an optimal policy can later be learned for any reward specified afterward. The framework naturally decomposes into two phases: an exploration phase, where trajectories are collected to ensure broad coverage, and a planning phase, where the collected data is used to compute an $\varepsilon$-optimal policy once a reward is revealed. Their proposed algorithm achieves this goal with sample complexity of order $\tilde{\mathcal{O}}(\frac{H^5 S^2 A}{\varepsilon^2})$, and is accompanied by a nearly matching lower bound of $\Omega(\frac{H^2 S^2 A}{\varepsilon^2})$.
This gap has since been closed progressively. Kaufmann et al. [2021] improved the upper bound to $\tilde{\mathcal{O}}(\frac{H^4 S^2 A}{\varepsilon^2})$, and Ménard et al. [2020] further refined it to $\tilde{\mathcal{O}}(\frac{H^3 S^2 A}{\varepsilon^2})$, which matches the lower bound

for non-stationary transition dynamics. The same bound has also been achieved independently by Li et al. [2024]. Beyond tabular MDPs, the reward-free framework has been extended to the linear MDP setting [Wagenmaker et al., 2022, Zhang et al., 2024].
Here, we focus primarily on the exploration phase of reward-free RL. Specifically, we leverage the idea of constructing datasets for induced expert MDPs, each of which produces a state distribution that provides sufficient coverage of the relevant state space. This principle serves as the key ingredient for deriving our sample-efficient algorithm.

# B Proof of Lower bound

In this section, we provide the detailed proof of Theorem 3.1 and Corollary 3.1. Both proofs require the same Markov game construction, which is again illustrated in Figure 3 to provide further intuition for the proof.

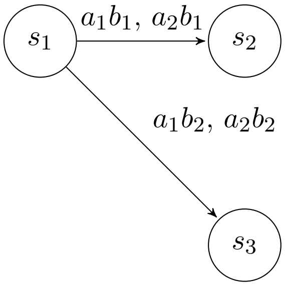


Figure 3: Markov game instance used for the Lower bound.

The main idea of the proof is twofold. In a first step, we show that even if $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}})$ is bounded, no non-interactive Multi-Agent Imitation Learning algorithm can learn an equilibrium from data as long as $\mathcal{C}_{\max} = \infty$. Then, we construct a state with two different reward functions such that the unique Nash equilibrium value for these games only differ by $\varepsilon$. This construction allows to show, that at least an expert dataset of size $\Omega(\varepsilon^{-2})$ is required to learn an Nash equilibrium from data. Additionally, noting that the expected visits of this state are given by $\mathcal{C}_{\max}$ completes the proof. Let us first restate both results (respectively Thm. 3.1 and Cor. 3.1).

**Theorem B.1.** *Let $\hat{\mu}, \hat{\nu}$ be the output of a non-interactive MAIL algorithm* Alg. *Then, for any* Alg*, there exists a Markov game such that satisfying $\mathbb{E}_{\mathrm{Alg}}\left[\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\widehat{\mu},\nu^\star}\rangle\right] \leq \mathcal{O}(\varepsilon)$ requires an expert dataset of size $N = \Omega(\frac{\mathcal{C}_{\max}}{\varepsilon^2})$.*

The same construction gives the following corollary for unbounded $\mathcal{C}_{\max}$.

**Corollary B.1.** *For any non-interactive MAIL algorithm, there exists a Markov game $\mathcal{G}$ with $\mathcal{C}(\mu^{\mathrm{E}}, \nu^{\mathrm{E}}) < \infty$ and $\mathcal{C}_{\max} = \infty$ where $\mathbb{E}_{\mathrm{Alg}}\left[\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\widehat{\mu},\nu^\star}\rangle\right] \geq \frac{H-1}{60}$.*

Next, we proceed with the proof of Theorem 3.1.

*Proof.* We start the proof with the setup of the Markov game (see Figure 3). Let us consider a family of two Markov games $\mathcal{H} = \{\mathcal{G}_1, \mathcal{G}_2, \}$ with states structure illustrated in Figure 1. The state space is $\mathcal{S} = \{s_1, s_2, s_3\}$, the action set for the $\mu$-player is $\mathcal{A} = \{a_1, a_2\}$ and for the $\nu$ player is $\mathcal{B} = \{b_1, b_2\}$.

The initial joint state is $s_1$, then the next state is $s_2$ when the $\nu$ player plays the action $b_1$ and $s_3$ if $b_2$ is played instead. That is, the next state is fully decided by the $\nu$ player no matter which is the action sampled by the $\mu$ player. The reward function is given by $r(s_1, a, b) = 0$ and $r(s_2, a, b) = 0$ for all Markov games in the family $\mathcal{H}$. On the contrary the reward function in the state $s_3$ differs among the members of the family $\mathcal{H}$. For any game $\mathcal{G} \in \mathcal{H}$, there exist a scalar parameter $\Delta_{\mathcal{G}} > 0$ which parameterize the payoff matrix as follows

$$r_{\mathcal{G}}(s_3, a, b) = \mathbf{e}_a^T \underbrace{\begin{pmatrix} 1 + \Delta_{\mathcal{G}} & -1 \\ -1 & 1 \end{pmatrix}}_{:= R_{\mathcal{G}}} \mathbf{e}_b,$$

where $\mathbf{e}_a$ and $\mathbf{e}_b$ are indicator vectors for the actions $a$ and $b$ respectively.
Note that for $\Delta_{\mathcal{G}} = 0$ this game is known as *Matching Pennies* and it is well known that the unique Nash equilibrium corresponds to a uniform policy for both players. In particular, we next consider $\mathcal{H} = \{\mathcal{G}_1, \mathcal{G}_2\}$ with $\Delta_{\mathcal{G}_1} = 2\varepsilon$ and $\Delta_{\mathcal{G}_2} = \varepsilon$. Therefore, both members in the family $\mathcal{H}$ can be seen as $\Delta$-perturbations of *Matching Pennies.* For a mixed Nash equilibrium it holds true that a player must be indifferent between both actions (see e.g.Osborne and Rubinstein [1994, Section 3.2.2]). Therefore, for the two arbitrary strategies $\mu = (p, 1-p)$ and $\nu = (q, 1-q)$, we get

$$(1 + \Delta_{\mathcal{G}})p - (1-p) = -p + 1 - p \Leftrightarrow p = \frac{1}{2 + \frac{\Delta_{\mathcal{G}}}{2}}.$$

Therefore, the Nash strategy is given by

$$\mu_{\text{Nash},\mathcal{G}} = \left( \frac{1}{2 + \frac{\Delta_{\mathcal{G}}}{2}}, 1 - \frac{1}{2 + \frac{\Delta_{\mathcal{G}}}{2}} \right)^\top = \left( \frac{1}{2} - \frac{\Delta_{\mathcal{G}}}{8 + 2\Delta_{\mathcal{G}}}, \frac{1}{2} + \frac{\Delta_{\mathcal{G}}}{8 + 2\Delta_{\mathcal{G}}} \right)^\top.$$

Note that for $\Delta_{\mathcal{G}} = 0$ we recover the Nash equilibrium strategies of the standard version of *Matching Pennies.* Because of the symmetry of the game $q = \frac{1}{2 + \frac{\Delta_{\mathcal{G}}}{2}}$ and can denote analogously $\nu_{\text{Nash},\mathcal{G}}$. Having defined the Nash strategies, we can compute the unique value of the game for player one as

$$\mu_{\text{Nash},\mathcal{G}}^\top R_{\mathcal{G}} \nu_{\text{Nash},\mathcal{G}} = \frac{\Delta_{\mathcal{G}}/2}{2 + \Delta_{\mathcal{G}}/2} := v_{\mathcal{G}}$$

Next, we compute the Nash Gap as a function of $\Delta_{\mathcal{G}} \in \mathbb{R}, p, q \in [0,1]$ when they play according to policies

$$\hat{\mu}(s_3) = [p, 1-p] \quad \hat{\nu}(s_3) = [q, 1-q]$$

In particular, we have that

$$\begin{aligned}
\text{Exploitability}_{\mathcal{G}}^{\nu}(\hat{\nu}) &:= \max_{\mu} \mu(s_3)^\top R_{\mathcal{G}} \hat{\nu}(s_3) - v_{\mathcal{G}} = \max_{p \in [0,1]} [p, 1-p] \begin{pmatrix} (2 + \Delta_{\mathcal{G}})q - 1 \\ 1 - 2q \end{pmatrix} - v_{\mathcal{G}} \\
&= \max \{(2 + \Delta_{\mathcal{G}})q - 1, 1 - 2q\} - v_{\mathcal{G}}, \\
\text{Exploitability}_{\mathcal{G}}^{\mu}(\hat{\mu}) &:= v_{\mathcal{G}} - \min_{\nu} \hat{\mu}(s_3)^\top R_{\mathcal{G}} \nu(s_3) = v_{\mathcal{G}} - \min_{q \in [0,1]} [q, 1-q] \begin{pmatrix} (2 + \Delta_{\mathcal{G}})p - 1 \\ 1 - 2p \end{pmatrix} \\
&= v_{\mathcal{G}} - \min \{(2 + \Delta_{\mathcal{G}})p - 1, 1 - 2p\},
\end{aligned}$$

Combining both exploitabilities, we can derive

$$\text{Nash}-\text{Gap}_{\mathcal{G}}(\hat{\mu},\hat{\nu}) = \max\left\{(2+\Delta_{\mathcal{G}})q - 1, 1-2q\right\} - \min\left\{(2+\Delta_{\mathcal{G}})p - 1, 1-2p\right\}.$$

Observe that for the Nash strategies $\mu_{\text{Nash},\mathcal{G}}$ and $\nu_{\text{Nash},\mathcal{G}}$ we have that the exploitability equals 0. Therefore, if both players play according to the Nash equilibrium policies in the state $s_3$ the first player gain reward $v_{\mathcal{G}} = \frac{\Delta_{\mathcal{G}}}{4+\Delta_{\mathcal{G}}} > 0$ which is larger than what can be gained in the state $s_2$. Vice-versa, the $\nu$-player can get at most $-v_{\mathcal{G}} = -\frac{\Delta_{\mathcal{G}}}{4+\Delta_{\mathcal{G}}} < 0$ playing against the Nash profile for the $\mu$ player. It follows that the $\nu$-player can always get a higher reward in the state $s_2$. Recalling the dynamics in the state $s_1$, we notice that the $\nu$-player can always ensure that the next visited state is $s_2$ by playing the action $b_1$. It follows that in the state $s_1$, the Nash equilibrium policies are $\mu^{\mathrm{E}}(s_1)$ arbitrary and $\nu^{\mathrm{E}}(b_1|s_1) = 1$ and $\nu^{\mathrm{E}}(b_2|s_1) = 0$. Therefore, for any arbitrary choice of $\mu^{\mathrm{E}}$, we have that $\nu^{\mathrm{E}}$ is the unique best response. Therefore, we have that for any choice of $\mu^{\mathrm{E}}$ it holds that the occupancy measure equals

$$d_0^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_1) = 1 \quad d_1^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_2) = 1 \quad d_1^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_3) = 0.$$

Therefore, for any offline state sampling distribution $\rho = \{\rho_h\}_{h=0}^{1}$ with $\rho_0(s_1) = 1$, we have that the single policy deviation coefficient for the $\mu$-player is given by

$$\mathcal{C}(\mu^{\mathrm{E}},\nu^{\mathrm{E}}) = \max\left\{\frac{d_0^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_1)}{\rho_0(s_1)}, \frac{d_1^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_2)}{\rho_1(s_1)}\right\} = \max\left\{\frac{1}{\rho_0(s_1)}, \frac{1}{\rho_1(s_2)}\right\} = \rho_1^{-1}(s_2),$$

where the equality follows from the fact that $\rho_0(s_1) = 1$. Therefore, to ensure a bounded value of $\mathcal{C}(\mu^{\mathrm{E}},\nu^{\mathrm{E}})$ it is enough to choose $\rho$ to have support including the states $s_1, s_2$ but not necessary $s_3$. However, let us now assume that the non-interactive MAIL algorithm at hand outputs a policy $\hat{\mu}$, no matter how this policy is produced. The $\nu$-player can choose to play according the policy $\nu_{\text{exploit}}$ such that $\nu_{\text{exploit}}(b_2|s_1) = 1$ to ensure that $d_1^{\mu^{\mathrm{E}},\nu_{\text{exploit}}}(s_3) = 1$. In words, the $\nu$-player can now choose an action outside the support of the Nash equilibrium to ensure that the next state is $s_3$. This would be irrational if $\hat{\mu}$ would coincide with the Nash profile in the state $s_3$ because in such situation the $\mu$-player could gain reward $v_{\mathcal{G}} > 0$ while the $\nu$ player would get $-v_{\mathcal{G}}$ at most. However, if $\rho_1(s_3) = 0$ the state is never visited and there exists at least one game in $\mathcal{H}$ where its sub-optimality is constant. On the one hand, $\mathrm{C}(\mu^{\mathrm{E}},\nu^{\mathrm{E}})$ is finite in this case. On the other hand, we have that the all policy deviations concentrability coefficient for the $\mu$-player is given for any arbitrary $\mu^{\mathrm{E}}$ by

$$\begin{aligned}\mathcal{C}_{\max} = \max_{\nu\in\nu^{\mathrm{E}},\nu_{\text{exploit}}}\left\{\frac{d_1^{\mu^{\mathrm{E}},\nu}(s_2)}{\rho_1(s_2)}, \frac{d_1^{\mu^{\mathrm{E}},\nu}(s_3)}{\rho_1(s_3)}\right\} &= \max\left\{\frac{d_1^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}(s_2)}{\rho_1(s_2)}, \frac{d_1^{\mu^{\mathrm{E}},\nu_{\text{exploit}}}(s_3)}{\rho_1(s_3)}\right\} \\ &= \max\left\{\frac{1}{\rho_1(s_2)}, \frac{1}{\rho_1(s_3)}\right\} = \frac{1}{\rho_1(s_3)},\end{aligned}$$

where the last equality follows assuming that $\rho_1(s_3) \le \rho_1(s_2)$. Therefore, we showed that learning an equilibrium from data is not possible if $\max_{\mu,\nu} C(\mu,\nu)$ is unbounded.

With the same construction, we can quantify a finite time statistical rate, bounding the number of times expert actions should be seen in $s_3$ in order to learn an $\varepsilon$-approximate Nash equilibrium.

Now, let us consider the defined $\hat{\mu}(s_3) = [p, 1-p]$, parametrized as the following function of $\alpha \in [0,1]$,

$$p = \frac{1}{2} - \left(\frac{2\alpha\varepsilon}{8+4\varepsilon} + \frac{(1-\alpha)\varepsilon}{8+2\varepsilon}\right),$$

and $q$ as a function of $\beta \in [0,1]$ as follows,

$$q = \frac{1}{2} - \left(\frac{2\beta\varepsilon}{8+4\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right).$$

Therefore, setting $\alpha = \beta = 1$, we have that $\hat{\mu}, \hat{\nu}$ are equilibrium policies in $\mathcal{G}_1$. Vice-versa, setting $\alpha = \beta = 0$ we have that $\hat{\mu}, \hat{\nu}$ equals $\mu_{\mathcal{G}_2}, \nu_{\mathcal{G}_2}$. Indeed, we can interpret $\alpha, \beta$ has the probability of choosing either the equilibrium profile in $\mathcal{G}_1$ or the one in $\mathcal{G}_2$. We now proceed proving a lower bound only for the non-interactive MAIL algorithm which outputs policies $\hat{\mu}, \hat{\nu}$ parameterized by values of $\alpha, \beta \in [0,1]$. In Lemma B.1 we prove that it is enough to consider this restricted class of policies. The intuition is that, an algorithm that only considers these policies has an advantage compared to any other algorithm for this lower bound, as the considered policies interpolate between the Nash equilibria of the two games. In particular, Lemma B.1 shows that the worst case expected exploitability can only increase for values $\alpha, \beta \notin [0,1]$.
To proceed, we write the Nash Gap in the game $\mathcal{G}_1$ as a function of $\alpha, \beta$. In particular, we have that

$$\max\left\{(2+\Delta_{\mathcal{G}_1})q - 1, 1-2q\right\} - v_{\mathcal{G}_1} = \max\left\{\frac{2\varepsilon(1+\varepsilon)(1-\beta)}{(4+\varepsilon)(2+\varepsilon)}, -\frac{2\varepsilon(1-\beta)}{(2+\varepsilon)(4+\varepsilon)}\right\}. \tag{2}$$

Next, we provide the detailed calculation how this was derived. As a first step, we consider the first term of the maximum expression. Let us start with the definition of $\Delta_{\mathcal{G}_1}$ and $q$ to get

$$(2+\Delta_{\mathcal{G}_1})q - 1 - v_{\mathcal{G}_1} = 2(1+\varepsilon)\left(\frac{1}{2} - \left(\frac{2\beta\varepsilon}{8+4\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)\right) - 1 - \frac{2\varepsilon}{4+2\varepsilon}.$$

Next, refactoring the first part of the equation and summarizing gives

$$\begin{aligned}
&2(1+\varepsilon)\left(\frac{1}{2} - \left(\frac{2\beta\varepsilon}{8+4\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)\right) - 1 - \frac{2\varepsilon}{4+2\varepsilon} \\
&= (1+\varepsilon) - 2(1+\varepsilon)\left(\frac{\beta\varepsilon}{4+2\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right) - 1 - \frac{2\varepsilon}{4+2\varepsilon} \\
&= \varepsilon - 2(1+\varepsilon)\left(\frac{\beta\varepsilon}{4+2\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right) - \frac{2\varepsilon}{4+2\varepsilon}.
\end{aligned}$$

Next, by simplifying the expression and bringing all parts on the same denominator, we get

$$\begin{aligned}
&\varepsilon - 2(1+\varepsilon)\left(\frac{\beta\varepsilon}{4+2\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right) - \frac{2\varepsilon}{4+2\varepsilon} \\
&= \varepsilon - \frac{\varepsilon(1+\varepsilon)}{4+\varepsilon} - \frac{2\varepsilon}{4+2\varepsilon} + \beta(1+\varepsilon)\left(\frac{\varepsilon}{4+\varepsilon} - \frac{\varepsilon}{2+\varepsilon}\right) \\
&= \varepsilon - \frac{\varepsilon(1+\varepsilon)}{4+\varepsilon} - \frac{\varepsilon}{2+\varepsilon} + \beta(1+\varepsilon)\left(\frac{\varepsilon(2+\varepsilon) - \varepsilon(4+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\right) \\
&= \left(\varepsilon - \frac{\varepsilon(1+\varepsilon)}{4+\varepsilon} - \frac{\varepsilon}{2+\varepsilon}\right) + \beta(1+\varepsilon)\left(\frac{2\varepsilon + \varepsilon^2 - 4\varepsilon - \varepsilon^2}{(4+\varepsilon)(2+\varepsilon)}\right) \\
&= \left(\frac{\varepsilon(4+\varepsilon)(2+\varepsilon) - \varepsilon(1+\varepsilon)(2+\varepsilon) - \varepsilon(4+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\right) + \beta(1+\varepsilon)\left(\frac{-2\varepsilon}{(4+\varepsilon)(2+\varepsilon)}\right).
\end{aligned}$$

As a last step, we simplify the numerator and get

$$\begin{aligned}
&\left(\frac{\varepsilon(4+\varepsilon)(2+\varepsilon)-\varepsilon(1+\varepsilon)(2+\varepsilon)-\varepsilon(4+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\right)+\beta(1+\varepsilon)\left(\frac{-2\varepsilon}{(4+\varepsilon)(2+\varepsilon)}\right)\\
&=\frac{\varepsilon\left[(4+\varepsilon)(2+\varepsilon)-(1+\varepsilon)(2+\varepsilon)-(4+\varepsilon)\right]}{(4+\varepsilon)(2+\varepsilon)}-\frac{2\beta\varepsilon(1+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\\
&=\frac{\varepsilon\left[(8+6\varepsilon+\varepsilon^2)-(2+3\varepsilon+\varepsilon^2)-(4+\varepsilon)\right]-2\beta\varepsilon(1+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\\
&=\frac{\varepsilon[2+2\varepsilon]-2\beta\varepsilon(1+\varepsilon)}{(4+\varepsilon)(2+\varepsilon)}\\
&=\frac{2\varepsilon(1+\varepsilon)(1-\beta)}{(4+\varepsilon)(2+\varepsilon)}.
\end{aligned}$$

For the second part of the maximum, we again start with the definition of $q$ and the Nash value of the first game $v_{\mathcal{G}_1}$ and receive

$$1-2q-v_{\mathcal{G}_1}=1-1+2\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)-\frac{2\varepsilon}{4+2\varepsilon}.$$

Simplifying now gives

$$\begin{aligned}
&1-1+2\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)-\frac{2\varepsilon}{4+2\varepsilon}\\
&=\left(\frac{\beta\varepsilon}{2+\varepsilon}+\frac{(1-\beta)\varepsilon}{4+\varepsilon}\right)-\frac{\varepsilon}{2+\varepsilon}\\
&=\left(\frac{\beta\varepsilon}{2+\varepsilon}-\frac{\varepsilon}{2+\varepsilon}\right)+\frac{(1-\beta)\varepsilon}{4+\varepsilon}\\
&=\frac{\beta\varepsilon-\varepsilon}{2+\varepsilon}+\frac{(1-\beta)\varepsilon}{4+\varepsilon}\\
&=-\frac{\varepsilon(1-\beta)}{2+\varepsilon}+\frac{\varepsilon(1-\beta)}{4+\varepsilon}.
\end{aligned}$$

As a last step, we again simplify and bring both terms on the same denominator to receive

$$\begin{aligned}
&-\frac{\varepsilon(1-\beta)}{2+\varepsilon}+\frac{\varepsilon(1-\beta)}{4+\varepsilon}\\
&=\varepsilon(1-\beta)\left(-\frac{1}{2+\varepsilon}+\frac{1}{4+\varepsilon}\right)\\
&=\varepsilon(1-\beta)\left(\frac{-(4+\varepsilon)+(2+\varepsilon)}{(2+\varepsilon)(4+\varepsilon)}\right)\\
&=\varepsilon(1-\beta)\left(\frac{-2}{(2+\varepsilon)(4+\varepsilon)}\right)\\
&=-\frac{2\varepsilon(1-\beta)}{(2+\varepsilon)(4+\varepsilon)}.
\end{aligned}$$

Putting both final expressions together gives (2).

Similarly, in the environment $\mathcal{G}_2$, we can compute that

$$\max\{(2+\Delta_{\mathcal{G}_2})q-1, 1-2q\} - v_{\mathcal{G}_2} = \max\left\{-\frac{\beta\varepsilon}{4+\varepsilon}, \frac{2\varepsilon\beta}{(2+\varepsilon)(4+\varepsilon)}\right\}. \tag{3}$$

For completeness, we also provide the detailed calculation next. First, we plug in the Nash value of the second game $v_{\mathcal{G}_2}$ as well as the considered strategy $q$ and the definition of $\Delta_{\mathcal{G}_2}$ and receive

$$2+\Delta_{\mathcal{G}_2}q-1-v_{\mathcal{G}_2} = (2+\varepsilon)\left(\frac{1}{2}-\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)\right)-1-\frac{\varepsilon}{4+\varepsilon}.$$

Next, we bring everything on the same denominator $2(4+\epsilon)$ and get

$$\begin{aligned}
&(2+\varepsilon)\left(\frac{1}{2}-\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)\right)-1-\frac{\varepsilon}{4+\varepsilon}\\
&=\frac{\varepsilon}{2}-\frac{\beta\varepsilon}{2}-\frac{\varepsilon(1-\beta)(2+\varepsilon)}{2(4+\varepsilon)}-\frac{\varepsilon}{4+\varepsilon}\\
&=\frac{\varepsilon(4+\varepsilon)-\beta\varepsilon(4+\varepsilon)-\varepsilon(1-\beta)(2+\varepsilon)-2\varepsilon}{2(4+\varepsilon)}
\end{aligned}$$

In the last step, we simplify the numerator, giving

$$\begin{aligned}
&\frac{\varepsilon(4+\varepsilon)-\beta\varepsilon(4+\varepsilon)-\varepsilon(1-\beta)(2+\varepsilon)-2\varepsilon}{2(4+\varepsilon)}\\
&=\frac{(4\varepsilon+\varepsilon^2)-(4\beta\varepsilon+\beta\varepsilon^2)-(\varepsilon(2+\varepsilon-2\beta-\beta\varepsilon))-2\varepsilon}{2(4+\varepsilon)}\\
&=\frac{4\varepsilon+\varepsilon^2-4\beta\varepsilon-\beta\varepsilon^2-2\varepsilon-\varepsilon^2+2\beta\varepsilon+\beta\varepsilon^2-2\varepsilon}{2(4+\varepsilon)}\\
&=-\frac{\beta\varepsilon}{4+\varepsilon}.
\end{aligned}$$

For the second part of the maximum expression, we again plug in the definitions of $v_{\mathcal{G}_2}$ and $q$ to get

$$1-2q-v_{\mathcal{G}_2} = 1-1+2\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)-\frac{\varepsilon}{4+\varepsilon}$$

Simplifying this, we directly get

$$\begin{aligned}
&1-1+2\left(\frac{2\beta\varepsilon}{8+4\varepsilon}+\frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right)-\frac{\varepsilon}{4+\varepsilon}\\
&=\left(\frac{\beta\varepsilon}{2+\varepsilon}-\frac{\beta\varepsilon}{4+\varepsilon}\right)\\
&=\frac{2\varepsilon\beta}{(2+\varepsilon)(4+\varepsilon)}.
\end{aligned}$$

Combining both derived expressions, we receive (3).

Let us now consider that $\hat{\nu}$ is the output of a non-interactive Multi-Agent Imitation Learning algorithm Alg (we denote this with $\hat{\nu} = \mathrm{Alg}(\mathcal{D}_{\mathcal{G}})$ ) which takes as input the pre-collected dataset sampled from the Nash profile which we denote $\mathcal{D}_{\mathcal{G}} = \left\{ B^i_{\mathcal{G}} \right\}_{i=1}^N$ where for each $i \in [N]$, $B^i_{\mathcal{G}} \sim \nu_{\mathrm{Nash},\mathcal{G}}$. We now have

$$\begin{aligned}
\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\mathrm{Exploitability}^{\nu}_{\mathcal{G}}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}}))\right] &\geq \frac{1}{2}\sum_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\mathrm{Exploitability}^{\nu}_{\mathcal{G}}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}}))\right] \\
&= \frac{1}{2}\Bigg( \frac{2\varepsilon(1+\varepsilon)\mathbb{P}_{\mathcal{G}_1}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_1}) = \nu_{\mathrm{Nash},\mathcal{G}_2})}{(4+\varepsilon)(2+\varepsilon)} \\
&\qquad + \frac{2\varepsilon\mathbb{P}_{\mathcal{G}_2}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_2}) = \nu_{\mathrm{Nash},\mathcal{G}_1})}{(2+\varepsilon)(4+\varepsilon)}\Bigg) \\
&\geq \frac{\varepsilon}{15}\left(\mathbb{P}_{\mathcal{G}_1}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_1}) = \nu_{\mathrm{Nash},\mathcal{G}_2}) + \mathbb{P}_{\mathcal{G}_2}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_2}) = \nu_{\mathrm{Nash},\mathcal{G}_1})\right) \\
&= \frac{\varepsilon}{15}\left(\mathbb{P}_{\mathcal{G}_1}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_1}) = \nu_{\mathrm{Nash},\mathcal{G}_2}) + \mathbb{P}_{\mathcal{G}_2}(\mathrm{Alg}(\mathcal{D}_{\mathcal{G}_2}) \neq \nu_{\mathrm{Nash},\mathcal{G}_2})\right) \\
&\geq \frac{\varepsilon}{30}\exp\left(-\mathrm{KL}\left(\mathbb{P}_{\mathcal{G}_1}, \mathbb{P}_{\mathcal{G}_2}\right)\right) \\
&\geq \frac{\varepsilon}{30}\exp\left(-N\mathrm{KL}\left(\nu_{\mathrm{Nash},\mathcal{G}_1}, \nu_{\mathrm{Nash},\mathcal{G}_2}\right)\right),
\end{aligned}$$

where we used $\beta \in [0,1]$ in the first equality and $\varepsilon \leq 1$ by assumption in the second inequality. In the second last inequality, we used the Bretagnolle-Huber inequality [Bretagnolle and Huber, 1979] which gives that for any distributions $P$ and $Q$ and event $A$ and its complementary $A^C$ it holds that

$$P(A) + Q(A^C) \geq \frac{1}{2}\exp\left(-\mathrm{KL}\left(P, Q\right)\right).$$

In the last inequality, we used that via the chain rule for the KL divergence and the identically independent sampling of the datasets $\mathcal{D}_{\mathcal{G}_1}$,$\mathcal{D}_{\mathcal{G}_2}$, we can rewrite $\mathrm{KL}\left(\mathbb{P}_{\mathcal{G}_1}, \mathbb{P}_{\mathcal{G}_2}\right)$ as $N\mathrm{KL}\left(\nu_{\mathrm{Nash},\mathcal{G}_1}, \nu_{\mathrm{Nash},\mathcal{G}_2}\right)$. Next, note that this can be seen as a Bernoulli random variable indicating whether we are in $\mathcal{G}_1$ or $\mathcal{G}_2$. At this point, we can treat $\nu_{\mathrm{Nash},\mathcal{G}_1}, \nu_{\mathrm{Nash},\mathcal{G}_2}$ as two Bernoulli random variables with mean $\frac{1}{2} - \frac{2\alpha\varepsilon}{8+4\varepsilon}$ and $\frac{1}{2} - \frac{\alpha\varepsilon}{8+2\varepsilon}$. Therefore, defining $\mathrm{kl} : [0,1]^2 \to \mathbb{R}$ as $\mathrm{kl}(r,s) = r\log(r/s) + (1-r)\log((1-r)/(1-s))$, recalling that the $\chi^2$ divergence between Bernoulli random variables with mean $r, s$ is given by $\chi^2(r,s) = \frac{(r-s)^2}{s(1-s)}$ and, finally, by the fact that the $\chi^2$-divergence upper bounds the $KL$-divergence we have that.

$$\begin{aligned}
N\mathrm{KL}\left(\nu_{\mathrm{Nash},\mathcal{G}_1}, \nu_{\mathrm{Nash},\mathcal{G}_2}\right) &= N\mathrm{kl}\left(\frac{1}{2} - \frac{2\alpha\varepsilon}{8+4\varepsilon}, \frac{1}{2} - \frac{\alpha\varepsilon}{8+2\varepsilon}\right) \\
&\leq N\chi^2\left(\frac{1}{2} - \frac{2\alpha\varepsilon}{8+4\varepsilon}, \frac{1}{2} - \frac{\alpha\varepsilon}{8+2\varepsilon}\right) \\
&= N\frac{\varepsilon^2(\varepsilon-4)^2}{(8+2\varepsilon)(4+\varepsilon)(4+2\varepsilon)}
\end{aligned}$$

Next, let us consider small $\varepsilon \in (0,1)$. Then, we have $9 \leq (\varepsilon - 4)^2 \leq 4^2 = 16$. For the denominator, we get that

$$8 \leq (8+2\varepsilon) \leq 10, \quad 4 \leq (4+\varepsilon) \leq 5, \quad 4 \leq 4+2\varepsilon \leq 6.$$

Putting this together gives

$$128 \leq (8+2\varepsilon)(4+\varepsilon)(4+2\varepsilon) \leq 300$$

Combining these we can bound the $\chi^2$ distance between

$$\frac{9\varepsilon^2}{300} \leq \chi^2\left(\nu_{\text{Nash},\mathcal{G}_1}, \nu_{\text{Nash},\mathcal{G}_2}\right) \leq \frac{16\varepsilon^2}{128}.$$

Plugging this into the expected exploitability gives

$$\begin{aligned}\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Exploitability}^{\nu}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))\right] &\geq \frac{\varepsilon}{30}\exp\left(-N\text{KL}\left(\nu_{\text{Nash},\mathcal{G}_1}, \nu_{\text{Nash},\mathcal{G}_2}\right)\right) \\ &\geq \frac{\varepsilon}{30}\exp\left(-N\chi^2\left(\frac{1}{2} - \frac{2\alpha\varepsilon}{8+4\varepsilon}, \frac{1}{2} - \frac{\alpha\varepsilon}{8+2\varepsilon}\right)\right) \\ &\geq \frac{\varepsilon}{30}\exp\left(-N\frac{1}{8}\varepsilon^2\right).\end{aligned}$$

To complete this step we need to set the number of samples $N$ to achieve a Nash Gap of $\mathcal{O}(\varepsilon)$. It follows that $N = \Omega(\frac{1}{\varepsilon^2})$. Therefore, it requires $N = \Omega(\frac{1}{\varepsilon^2})$ to learn a $\mathcal{O}(\varepsilon)$ Nash equilibrium in state $s_3$.

We remind ourselves that the expected number of times any non-interactive algorithm visits state $s_3$ is given by $\frac{1}{\rho_1(s_3)} = \mathcal{C}_{\max}$. Combining this with the previous step we receive

$$\begin{aligned}\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Exploitability}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))\right] &\geq \mathcal{C}_{\max}\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Exploitability}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))(s_3)\right] \\ &\geq \frac{\mathcal{C}_{\max}\varepsilon}{30}\exp\left(-N\text{KL}\left(\nu_{\text{Nash},\mathcal{G}_1}, \nu_{\text{Nash},\mathcal{G}_2}\right)\right) \\ &\geq \frac{\mathcal{C}_{\max}\varepsilon}{30}\exp\left(-N\chi^2\left(\frac{1}{2} - \frac{2\alpha\varepsilon}{8+4\varepsilon}, \frac{1}{2} - \frac{\alpha\varepsilon}{8+2\varepsilon}\right)\right) \\ &\geq \frac{\mathcal{C}_{\max}\varepsilon}{30}\exp\left(-\frac{N}{8}\varepsilon^2\right).\end{aligned}$$

Therefore, for any non-interactive Alg it requires an expert dataset of size $N = \Omega(\frac{\mathcal{C}_{\max}}{\varepsilon^2})$ to learn a $\mathcal{O}(\varepsilon)$ Nash equilibrium from data.

To complete the proof it remains to show that the considered policy class is enough. This follows directly from Lemma B.1. This completes the proof of Theorem 3.1. □

Next, we will provide the result, that all policies outside of the considered policy classes in the derived proof suffer from a higher worst case exploitability. In particular, we will show that the minimizer of the exploitability across the two games lies within the considered policy class.

**Lemma B.1.** *Let a parametrized Normal Form Game be given by*

$$\begin{pmatrix} 1+\Delta_{\mathcal{G}} & -1 \\ -1 & 1 \end{pmatrix} := R_{\mathcal{G}}$$

*for $\Delta_{\mathcal{G}} \in \{\varepsilon, 2\varepsilon\}$. Additionally, let the following strategies be given*

$$\begin{aligned} p_\alpha &= \frac{1}{2} - \left(\frac{2\alpha\varepsilon}{8+4\varepsilon} + \frac{(1-\alpha)\varepsilon}{8+2\varepsilon}\right) \\ q_\beta &= \frac{1}{2} - \left(\frac{2\beta\varepsilon}{8+4\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon}\right) \end{aligned}$$

*for $\alpha, \beta \in [0,1]$. Then, all other strategies $p \notin p_\alpha$ and $q \notin q_\beta$ suffer from a higher worst case exploitability across the two Normal Form Games.*

*Proof.* The idea is that we define a general policy and show that the minimizer of the maximal exploitability across the two Games lies within the considered policy class. We only complete the proof for $q$, it follows analogously for $p$.
We introduce a general policy $\mu = (q, 1-q) \in \Delta_2$. For a general policy $\mu$, a simple calculation provides the following potential exploitabilities from the perspective of player 1 stated as a function of $q$ across both Games:

$$\begin{aligned} f_1(q) &:= (2+2\varepsilon)q - 1 - \frac{\varepsilon}{2+\varepsilon}, \\ f_2(q) &:= (2+\varepsilon)q - 1 - \frac{\varepsilon/2}{2+\varepsilon/2}, \\ f_3(q) &:= 1 - 2q - \frac{\varepsilon}{2+\varepsilon}, \\ f_4(q) &:= 1 - 2q - \frac{\varepsilon/2}{2+\varepsilon/2}. \end{aligned}$$

Comparing $f_3(q)$ and $f_4(q)$, we obtain that they are the same functions except for the $\varepsilon$-term. It holds that $\frac{\varepsilon}{2+\varepsilon} > \frac{\varepsilon/2}{2+\varepsilon/2}$ and therefore $f_3(q) < f_4(q)$. Then, let us define the following convex function,

$$F(q) = \max\{f_1(q), f_2(q), f_4(q)\}.$$

Note that this function is indeed convex as a maximum of affine functions.
Next, we show that we can further simplify $F(q)$. Observe that if $f_2(q) > f_1(q)$, then

$$f_4(q) - f_2(q) = \left(1 - 2q - \tfrac{\varepsilon/2}{2+\varepsilon/2}\right) - \left((2+\varepsilon)q - 1 - \tfrac{\varepsilon/2}{2+\varepsilon/2}\right) = 2 - (4+\varepsilon)q.$$

Since $f_2 > f_1$ implies $(2+\varepsilon)q > (2+2\varepsilon)q - \frac{\varepsilon}{2+\varepsilon}$, one can check that this forces $q < \frac{2}{4+\varepsilon}$, in which case $f_4(q) > f_2(q)$. Hence whenever $f_2$ dominates $f_1$, we automatically have $f_4 > f_2$. Thus the maximum is always realized by either $f_1$ or $f_4$, and we can rewrite

$$F(q) = \max\{f_1(q), f_4(q)\}.$$

As stated $F(q)$ is the maximum of affine functions, therefore $F$ is convex and has a minimizer (see e.g. Boyd and Vandenberghe [2004, Section 3.2.3]). Knowing that $F$ is convex, we can use tools from convex optimization to find $q^* \in \operatorname{argmin} F(q)$. We know that $q^*$ is a minimizer of $F$ if and only if $F$ is subdifferentiable at $q^*$ and

$$0 \in \partial F(q^*).$$

As a first step, we note that we can write $F(q) = \max\{f_i\}_{i=1}^2$, where $f_i$ is an affine function $\forall i \in \{1,2\}$. This implies that the subdifferential at $q^*$ exists and is given by the convex hull of gradients of all active functions at $q^*$. Next, let us define the index set of all active functions at $q^*$, we get

$$A := \{i : f_i(q^*) = F(q^*)\}.$$

With this definition we have $0 \in \partial F(q^*) = \operatorname{conv}\{\nabla f_i(q^*) : i \in A\}$ and this can equivalently be expressed as $\lambda_i \geq 0$ for $i \in A$ with $\sum \lambda_i = 1$ s.t.

$$\sum_{i \in A} \lambda_i \nabla f_i(q^*) = 0.$$

Next, let us compute the gradients with respect to $q$. We get

$$\nabla f_1(q) = 2 + 2\varepsilon \quad \nabla f_4(q) = -2$$

Now, let us check that at $q^*$ both functions must be active. Observe that

$$\nabla f_1(q) = 2 + 2\varepsilon > 0, \qquad \nabla f_4(q) = -2 < 0.$$

Since neither gradient is zero, no single affine function has zero slope. This implies that an interior minimizer of the convex function $F(q)$ cannot occur at a point where only one $f_i$ is active. Hence, any interior minimizer must be attained at an intersection where at least two affine pieces are active. In particular, the necessary condition for an interior minimizer here is

$$f_1(q) = f_4(q).$$

Let us now check that indeed both functions can be active:

$$\lambda_1(2 + 2\varepsilon) - 2\lambda_4 = 0 \Leftrightarrow \lambda_1(2 + 2\varepsilon) - 2(1 - \lambda_1) = 0 \Leftrightarrow \lambda_1 = \frac{2}{4 + 2\varepsilon} \in (0, 1),$$

and therefore also $\lambda_4 = 1 - \frac{2}{4+2\varepsilon} \in (0, 1)$. This indicates that both functions are active at $q^*$, meaning that $A = \{1, 4\}$. From this we know that $f_1(q^*) = f_4(q^*)$ and we get

$$\begin{aligned} (2 + 2\varepsilon)q^* - 1 - \frac{\varepsilon}{2 + \varepsilon} &= 1 - 2q^* - \frac{\varepsilon/2}{2 + \varepsilon/2} \\ \Leftrightarrow q^* = \frac{2 + \frac{\varepsilon}{2+\varepsilon} - \frac{\varepsilon/2}{2+\varepsilon/2}}{(4 + 2\varepsilon)} &= \frac{2 + \frac{\varepsilon}{2+\varepsilon} - \frac{\varepsilon}{4+\varepsilon}}{2(2 + \varepsilon)}. \end{aligned}$$

It follows that it holds true that $q^* \in (0, 1)$. Next, we remind ourselves that in the lower bound proof we considered strategies of the form, with $\beta \in [0, 1]$:

$$q = \frac{1}{2} - \left( \frac{2\beta\varepsilon}{8 + 4\varepsilon} + \frac{(1 - \beta)\varepsilon}{8 + 2\varepsilon} \right).$$

Next, we explicitly calculate $\beta$ to see that indeed the policy lies within $q_\beta$. We remind ourselves, that $q_\beta := \frac{1}{2} - \left( \frac{2\beta\varepsilon}{8+4\varepsilon} + \frac{(1-\beta)\varepsilon}{8+2\varepsilon} \right)$. Let

$$F(\beta) = \frac{1}{2} - \left( \frac{2\beta\varepsilon}{8 + 4\varepsilon} + \frac{(1 - \beta)\varepsilon}{8 + 2\varepsilon} \right).$$

This can be rearranged as

$$F(\beta) = \frac{1}{2} - \frac{\varepsilon}{8 + 2\varepsilon} + \beta \left( \frac{\varepsilon}{8 + 2\varepsilon} - \frac{2\varepsilon}{8 + 4\varepsilon} \right).$$

Now solve $F(\beta) = q^*$. This yields

$$\beta = \frac{\varepsilon + 1}{\varepsilon + 2}.$$

Since for $\varepsilon > 0$ we have $\frac{1}{2} < \frac{\varepsilon+1}{\varepsilon+2} < 1$, this $\beta$ indeed lies in $[0, 1]$.
Therefore, for $\beta = \frac{\varepsilon+1}{\varepsilon+2}$ we recover $q^*$. This, together with the convexity of $F(q)$, shows that the maximal worst case exploitability is always higher for all strategies outside the considered policy class. This completes the proof. □

Finally, we briefly describe how the proof of Corollary 3.1 is extracted from the proof of Theorem 3.1.

*Proof of Corollary 3.1.* In the proof, for Theorem 3.1 we obtained

$$\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Exploitability}^{\nu}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))\right] \geq \frac{\varepsilon}{30}\exp\left(-N\text{KL}\left(\nu_{\text{Nash},\mathcal{G}_1}, \nu_{\text{Nash},\mathcal{G}_2}\right)\right),$$

where $N$ is the number of visits in $s_3$. However, if $\mathcal{C}_{\max} = \infty$, then $N = 0$ because $s_3$ is never visited. This implies

$$\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Exploitability}^{\nu}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))\right] \geq \frac{\varepsilon}{30}.$$

Repeating the same steps for the other player, and setting $\varepsilon = 1/4$ which is the largest possible value that ensures that the payoffs are bounded in $[0, 1]$ yields

$$\max_{\mathcal{G}\in\mathcal{H}} \mathbb{E}\left[\text{Nash-Gap}_{\mathcal{G}}(\text{Alg}(\mathcal{D}_{\mathcal{G}}))\right] \geq \frac{1}{60}.$$

At this point, let us consider that after playing one action in $s_3$, the agents move in another state which has exactly the same reward matrices of $s_3$, i.e. $R_{\mathcal{G}_1}$ and $R_{\mathcal{G}_2}$. The same transition is repeated for $H - 1$ times to ensure that the game has horizon $H$. Let us denote these $H$ steps games $\mathcal{G}_1'$, $\mathcal{G}_2'$ and the class $\mathcal{H}' := \{\mathcal{G}_1', \mathcal{G}_2'\}$. In this game, we then have

$$\max_{\mathcal{G}'\in\mathcal{H}'} \mathbb{E}\left[\text{Nash-Gap}_{\mathcal{G}'}(\text{Alg}(\mathcal{D}_{\mathcal{G}'}))\right] \geq \frac{H-1}{60}.$$

□

# C Omitted Proofs for MAIL-WARM

In this section, we provide a summary of the main steps used for our main result (Theorem 3.2), before we give all the missing details for the analysis of MAIL-WARM. After the summary, we give the pseudo-code of EULER, that we use in the reward-free warm-up phase. Then, we give the missing proof of Lemma 6.1 and last, we provide the concentration result used for the BC part of the algorithm.

We remind ourselves, that the analysis of the algorithm can be divided into two main steps. The first step concerns the reward-free warm-up phase. Here, we consider the expert-induced MDPs (Definition 6.1), which are constructed using access to the queriable experts. Informally, within these induced MDPs, we build datasets $(\mathcal{D}^{\nu^{\text{E}}}, \mathcal{D}^{\mu^{\text{E}}})$ that provide sufficient coverage of the relevant states. This result is formalized in Theorem 6.1.

In the second step, these datasets are used to recover the Nash equilibrium policies. Specifically, we apply Behavior Cloning on $\mathcal{D}^{\nu^{\text{E}}}$ to approximate the expert policy $\nu^{\text{E}}$ and on $\mathcal{D}^{\mu^{\text{E}}}$ to approximate $\mu^{\text{E}}$. To establish that this procedure effectively minimizes the Nash−Gap, we rely on Lemma 6.1, which decomposes the Nash−Gap in a way that leverages the dataset distributions. Finally, we invoke Lemma C.2 to bound the concentration error of Behavior Cloning on both datasets, thereby completing the proof.

## C.1 EULER algorithm

First, for completeness reason we state the EULER pseudo-code. EULER was introduced by Zanette and Brunskill [2019]. In our context it is used for the reward-free warm-up phase. In particular, it is used to solve the $SH$ RL problems that maximize the probability to reach a certain state. The full pseudo-code is given in Algorithm 2.

**Algorithm 2** EULER$(r, N_0, P)$

```
Require: Reward function $r$, episodes $N_0$, environment dynamics $P = \{P_h\}_{h=1}^H$ .
Initialize: $\delta' = \frac{1}{7}\delta,$ , $B_p = H\sqrt{2\ln\frac{(4SAN_0)}{\delta'}}$, $B_v = \sqrt{2\ln\frac{(4SAN_0)}{\delta'}}$, $J = H\ln\frac{(4SAN_0)/\delta'}{3}$
Initialize: $\pi_h^1 = \text{Uniform}(\mathcal{A})$ for all $h \in [H]$, for all $s \in \mathcal{S}$.
for $k = 1, 2, \dots, N_0$ do
    Sample a trajectory $(s_1^k, a_1^k, \dots, s_H^k, a_H^k)$ with policy $\pi^k$ in the environment with dynamics $P$.
    Set $V_{H+1}^k = 0$.
    for $h = H, H-1, \dots, 1$ do
        $N_h^k(s', s, a) = \sum_{\tau=1}^k \mathbb{1}\left\{s_{h+1}^\tau, a_h^\tau, s_h^\tau = s', a, s\right\}$
        $N_h^k(s, a) = \sum_{s' \in \mathcal{S}} N_h^k(s', s, a)$.
        $\hat{P}_h^k(s'|s, a) = \frac{N_h^k(s',s,a)}{N_h^k(s,a)}$.
        $b_h^k(s, a) = \sqrt{\frac{2\widehat{Var}_{\hat{P}_h^k}(\bar{V}_{h+1}^k)(s,a)\ln\frac{4SAT}{\delta'}}{N_h^k(s,a)}} + \frac{H\ln\frac{4SAT}{\delta'}}{3(N_h^k(s,a)-1)}$ where $\widehat{Var}_{\hat{P}_h^k}(V)(s, a) = \hat{P}_h^k(V - \hat{P}_h^k V(s, a))^2(s, a)$
        $B_k^h(s, a) = b_k^h(s, a) + \frac{1}{\sqrt{N_h^k(s,a)}}\left(\frac{4J+B_p}{\sqrt{N_h^k(s,a)}} + B_v\|\bar{V}_{h+1}^k - \underline{V}_{h+1}^k\|_{2,\hat{P}_h^k}\right)$
        $Q_h^k(s, a) = \min\left\{H - h, r_h(s, a) + \hat{P}_h^k\bar{V}_{h+1}^k(s, a) + B_h^k(s, a)\right\}$
        $\pi_h^k(s) = \arg\max_{a\in\mathcal{A}} Q_h^k(s, a)$
        $\bar{V}_h^k(s) = \max_{a\in\mathcal{A}} Q_h^k(s, a)$
        $\underline{V}_h^k(s) = \max\left\{0, r_h(s, a) + \hat{P}_h^k\underline{V}_{h+1}^k(s, a) - B_h^k(s, a)\right\}$
    end for
end for
```

## C.2 Proof of Lemma 6.1

For proving this result, the main idea is to decompose the exploitability into the Total variation between the estimated expert and the true one. Next, we will restate the Lemma.

**Lemma C.1 (Exploitability decomposition).** *For any policy pair $\nu, \nu'$, we define their total variation at state $s$ as $\text{TV}(\nu, \nu')(s) = \sum_{b\in\mathcal{B}} |\nu(b|s) - \nu'(b|s)|$ . It holds that*

$$\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\widehat{\mu},\nu^\star}\right\rangle \leq 2H\sum_{h=1}^H \sum_{\pi\in\{\hat{\mu},\hat{\nu}\}} \text{Err}_h(\pi).$$

*where we have defined* $\text{Err}_h(\hat{\nu}) := \max_{\mu_h\in\text{br}(\widehat{\nu}_h)} \mathbb{E}_{s\sim d_h^{\mu,\nu^{\text{E}}}}\left[\text{TV}\left(\nu_h^{\text{E}}, \widehat{\nu}_h\right)(s)\right]$ *for player 1 and additionally* $\text{Err}_h(\hat{\mu}) := \max_{\nu_h\in\text{br}(\widehat{\mu}_h)} \mathbb{E}_{s\sim d_h^{\mu^{\text{E}},\nu}}\left[\text{TV}\left(\mu_h^{\text{E}}, \widehat{\nu}_h\right)(s)\right]$ *for player 2.*

*Proof.* We start by upper bounding the decomposition as follows

$$
\begin{aligned}
\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\widehat{\mu},\nu^\star}\right\rangle &= \left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}\right\rangle + \left\langle d_0, V^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}} - V^{\widehat{\mu},\nu^\star}\right\rangle \\
&\le \left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\mu^\star,\nu^{\mathrm{E}}}\right\rangle + \left\langle d_0, V^{\mu^{\mathrm{E}},\nu^\star} - V^{\widehat{\mu},\nu^\star}\right\rangle,
\end{aligned}
$$

where $\mu^\star$ and $\nu^\star$ are arbitrary policies in the sets $\mathrm{br}(\hat{\nu})$ and $\mathrm{br}(\hat{\mu})$ respectively. At this point, we identified two pairs of value function differences where one policy is fixed, respectively $\mu^\star$ and $\nu^\star$. Therefore, applying the performance difference lemma (see e.g. Kakade and Langford [2002]) in the MDP induced by $\mu^\star$ we obtain

$$
\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\mu^\star,\nu^{\mathrm{E}}}\right\rangle \le \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\mu^\star,\nu^{\mathrm{E}}}} \left[\left\langle Q^{\mu^\star,\hat{\nu}}(s,\cdot), \hat{\nu}(\cdot|s) - \nu^{\mathrm{E}}(\cdot|s)\right\rangle\right].
$$

Then, by Hölder's inequality with $\|\cdot\|_1$ and $\|\cdot\|_\infty$ and additionally bounding the value function with its maximal value $H$, it holds that

$$
\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\mu^\star,\nu^{\mathrm{E}}}\right\rangle \le H \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\mu^\star,\nu^{\mathrm{E}}}} \left[\mathrm{TV}\left(\hat{\nu}(\cdot|s), \nu^{\mathrm{E}}(\cdot|s)\right)\right].
$$

Then, since we aim for a bound on the left hand side that holds for any $\mu^\star \in \mathrm{br}(\hat{\nu})$ we need to pick the maximizer over the right hand side.

$$
\left\langle d_0, V^{\mu^\star,\widehat{\nu}} - V^{\mu^\star,\nu^{\mathrm{E}}}\right\rangle \le H \max_{\mu\in\mathrm{br}(\hat{\nu})} \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\mu,\nu^{\mathrm{E}}}} \left[\mathrm{TV}\left(\hat{\nu}(\cdot|s), \nu^{\mathrm{E}}(\cdot|s)\right)\right]
$$

Equivalent steps for the second player to upper bound $\left\langle d_0, V^{\mu^{\mathrm{E}},\nu^\star} - V^{\widehat{\mu},\nu^\star}\right\rangle$ concludes the proof. □

## C.3 Behavior Cloning concentration

Last, we adapt the analysis of Lemma D.1 by Freihaut et al. [2025] to our setting. The main changes are that the expectation used in our setting is with respect to the dataset distributions $p^{\nu^{\mathrm{E}}}$ and $p^{\mu^{\mathrm{E}}}$ while their is with respect to the expert occupancy measure $d^{\mu^{\mathrm{E}},\nu^{\mathrm{E}}}$.

**Lemma C.2.** *Let $p_h^{\nu^{\mathrm{E}}}$ and $p_h^{\mu^{\mathrm{E}}}$ be two distributions received from running the reward-free warm-up of Algorithm MAIL-WARM and $N$ the size of the received datasets. Then, for all $s \in \mathcal{S}_{\delta,h}^{\nu^{\mathrm{E}}}$ it holds with probability of at least $1-\delta/2$ that*

$$
\mathbb{E}_{s\sim p_h^{\nu^{\mathrm{E}}}} \left[\mathrm{TV}\left(\nu_h^{\mathrm{E}}, \widehat{\nu}_h\right)(s)\right] \le \sqrt{\frac{SB\log(4S/\delta)}{N}}.
$$

*Similarly, for all $s \in \mathcal{S}_{\delta,h}^{\mu^{\mathrm{E}}}$ it holds with probability of at least $1-\delta/2$ that*

$$
\mathbb{E}_{s\sim p_h^{\mu^{\mathrm{E}}}} \left[\mathrm{TV}\left(\mu_h^{\mathrm{E}}, \widehat{\mu}_h\right)(s)\right] \le \sqrt{\frac{SA\log(4S/\delta)}{N}}.
$$

*Proof.* We only provide the proof for the distribution $p_h^{\nu^{\mathrm{E}}}$, it follows analogously for $p_h^{\mu^{\mathrm{E}}}$. We get

$$\begin{aligned}
\mathbb{E}_{s\sim p_h^{\nu^{\mathrm{E}}}}\left[\mathrm{TV}\left(\nu_h^{\mathrm{E}},\widehat{\nu}_h\right)(s)\right] &\overset{\text{(i)}}{\leq} \sum_{s\in\mathcal{S}} p_h^{\nu^{\mathrm{E}}}(s)\sqrt{\frac{2B\log(4S/\delta)}{\max\{N(s),1\}}} \\
&= \sum_{s\in\mathcal{S}} \sqrt{p_h^{\nu^{\mathrm{E}}}(s)}\sqrt{\frac{2B\, p_h^{\nu^{\mathrm{E}}}(s)\log(4S/\delta)}{\max\{N(s),1\}}} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\sum_{s\in\mathcal{S}} \frac{2B\, p_h^{\nu^{\mathrm{E}}}(s)\log(4S/\delta)}{\max\{N(s),1\}}} \\
&\overset{\text{(iii)}}{\leq} \sqrt{\sum_{s\in\mathcal{S}} \frac{16B\log^2(4S/\delta)}{N}} \\
&= 4\sqrt{\frac{SB\log^2(4S/\delta)}{N}},
\end{aligned}$$

where in (i) we applied Lemma E.1 and a union bound over the state space $\mathcal{S}$, in (ii) we applied Cauchy Schwarz and in (iii) we applied Lemma E.2, reminding ourselves that $N$ is the size of the dataset. □

## C.4 Extension to n-player general-sum Markov games

In this paragraph, we show that our approach is easily extendable to $n$-player general-sum games. Freihaut et al. [2025] also provide an extension for this setting, however their proof is hard to parse while our approach translates to the $n$-player setting smoothly. First, we introduce all the necessary notation.

**Notation General-Sum Markov games** A general-sum Markov game is defined by the tuple $\mathcal{G} = (n, H, \mathcal{S}, \mathcal{A}, P, r, d_0)$, where compared to zero-sum games, $n$ is now the number of players, $\mathcal{A} := \mathcal{A}_1 \times .. \times \mathcal{A}_n$ the joint action space composed of the individual action spaces $\mathcal{A}_i$; the reward function for the $i^{th}$ player at stage $h \in [H]$ is $r_{h,i} : \mathcal{S} \times \mathcal{A} \to [0,1]$ and $P$ is now the transition function that takes a joint action $a \in \mathcal{A}$ as an input. We denote a joint policy as $\pi := (\pi_1, \ldots, \pi_n)$, where $\pi_i : \mathcal{S} \to \Delta_{\mathcal{A}_i}$, where $\Delta_{\mathcal{A}_i}$ is the probability simplex over the individual action space $\mathcal{A}_i$. We denote the set of Markov policies for the agent $i$ as $\Pi_i$. Further let us denote the cardinality of an individual action space as $A_i := |\mathcal{A}_i|$ and the maximal cardinality of the individual state spaces as $A_{\max} := \max_i |\mathcal{A}_i|$ for $i \in [n]$. Additionally, we will make use of the convention that $\pi_{-i}$ denotes the policy of all agents but policy of agent $i$. The same notation is also used for an action $a_{-i} = (a_1, \ldots, a_{i-1}, a_{i-1}, \ldots, a_n)$ and the according action space is denoted by $\mathcal{A}_{-i}$. At this point, we can define the value function of the policy profile $\pi$ at stage $h \in [H]$ for a certain agent $i \in [n]$ as $V_{h,i}^{\pi}(s) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H-1} r_{i,t}(S_t, A_1, \ldots, A_n) \,\middle|\, S_h = s\right]$. Often times, we will shorten $V_{0,i}^{\pi}$ by $V_i^{\pi}$. Having defined the value function, we can introduce the Nash gap for general sum games as follows

$$\text{Nash-Gap}(\pi) = \max_{i\in[n]} \max_{\pi_i' : \mathcal{S}\to\Delta_{\mathcal{A}_i}} \langle d_0, V_i^{\pi_i', \pi_{-i}} - V_i^{\pi}\rangle.$$

Note that in general-sum games, the value of the NE does not need to be unique and the set of NE is not convex in general.

**Results for the $n$-player setting** With the given notation, let us redefine the expert induced MDP for this setting.

**Definition C.1** (Experts Induced MDP). *Let $\mathcal{G}$ be a Markov game and $\pi^E$ the expert policies, then $\mathcal{M}^{\pi^E_{-i}} := (\mathcal{S}, \mathcal{A}, P^{\pi^E_{-i}}, r^{\pi^E_{-i}}, H)$ is the MDP induced by the experts $\pi^E_{-i}$ with the transition model $P_h^{\pi^E_{-i}}(s' \mid s, a) := \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi^E_{-i,h}(a_{-i} \mid s) P_h(s' \mid s, a, a_{-i})$ and an arbitrary reward function $r_{i,h}^{\pi^E_{-i}}(s, a_i) \in \{0, 1\} \quad \forall (s, a_i) \in \mathcal{S} \times \mathcal{A}_i$.*

Now, we can restate Algorithm 3 for the n-player general-sum setting.
The reward free warm-up phase again produces a dataset that covers all $\delta$- significant states well.

**Corollary C.1.** *Let $\mathcal{M}^{\pi^E_{-i}}$ be the induced MDP defined in C.1 and the policy set $\Psi^{\pi_{-i}}$ is generated according to Algorithm 3. Then there exists an absolute constant $c > 0$ such that for any $\varepsilon > 0$ and $p \in (0, 1)$, if we set $N_0 \geq n^2 c S^2 A H^4 \iota_0^3/\delta$, where $\iota_0 := \log(SAH/p\delta)$, then with probability $1 - p$, the reward free exploration returns a sampling distribution $p^{\pi^E_{-i}}$ such that for all $\pi_i \in \Pi$*

$$\forall \delta - \text{significant}(s, h), \quad \max_{a,h} \frac{d_h^{\pi_i, \pi^E_{-i}}(s, a)}{p_h^{\pi^E_{-i}}(s, a)} \leq 2 S A_i H,$$

*where $\delta - \text{significant}$, means that the probability to reach a state $s$ in the induced MDP $\mathcal{M}^{\pi^E_{-i}}$ is lower bounded by $\delta$:*

$$\max_{\pi_i} d_h^{\pi_i, \pi^E_{-i}} \geq \delta.$$

Next, we can state the guarantees of MAIL-WARM for the $n$-player general-sum setting.

**Corollary C.2.** *For any $\varepsilon > 0$ and $\delta_{\text{fail}} \in (0, 1)$ if we execute Algorithm 3 and choose the parameters according to $N = \mathcal{O}(\frac{n^4 H^6 S^3 A^3_{\max} \log(S/\delta_{\text{fail}})}{\varepsilon^2})$ and $N_0 \geq \mathcal{O}\left(n^3 S^3 A^2_{\max} H^6 \iota_0^3/\varepsilon\right)$, we get with probability $1 - \delta_{\text{fail}}$ for the policies $\widehat{\pi}$ that*

$$\text{Nash}-\text{Gap}(\widehat{\pi}) \leq \mathcal{O}(\varepsilon).$$

*Proof.* Let us first remind ourselves that the Nash Gap for the $n$-player general-sum setting is slightly different. In particular, we have

$$\begin{aligned}
\text{Nash-Gap}(\hat{\pi}) &= \sum_{i=1}^{n} \max_{\pi'_i} \left\langle d_0, V_i^{\pi'_i, \hat{\pi}_{-i}} - V_i^{\hat{\pi}_i, \hat{\pi}_{-i}} \right\rangle \\
&= \sum_{i=1}^{n} \left\langle d_0, V_i^{\pi_i^\star, \hat{\pi}_{-i}} - V_i^{\pi_i^E, \pi_{-i}^E} \right\rangle + \left\langle d_0, V_i^{\pi_i^E, \pi_{-i}^E} - V_i^{\hat{\pi}_i, \hat{\pi}_{-i}} \right\rangle \\
&\leq \underbrace{\sum_{i=1}^{n} \left\langle d_0, \, V_i^{\pi_i^\star, \hat{\pi}_{-i}} - V_i^{\pi_i^\star, \pi_{-i}^E} \right\rangle}_{:=\text{Exploit-Gap}} + \underbrace{\sum_{i=1}^{n} \left\langle d_0, \, V_i^{\pi_i^E, \pi_{-i}^E} - V_i^{\hat{\pi}_i, \hat{\pi}_{-i}} \right\rangle}_{:=\text{Value-Gap}}.
\end{aligned}$$

**Algorithm 3** Multi-Agent Imitation Learning with reward-free warm-up (MAIL-WARM) for n-player Games

```
Input: iteration number $N_0$, $N$, queriable experts $\pi^E$.
Reward-free warm-up phase:
for all $i \in [n]$ do
    set policy class $\Psi^{\pi_{-i}^E} \leftarrow \emptyset$, and dataset $\mathcal{D} \leftarrow \emptyset$.
    for all $(s, h) \in \mathcal{S} \times [H]$ do
        $r_{i,h}^{\pi_{-i}^E}(s', a_i') \leftarrow \mathbf{1}[s' = s \text{ and } h' = h]$ for all $(s', a_i', h') \in \mathcal{S} \times \mathcal{A} \times [H]$.
        $\left\{\pi_i^{(s,h)}\right\}_{i=1}^{N_0} \leftarrow \text{EULER}(r^{\pi_{-i}^E}, N_0, P^{\pi_{-i}^E})$.
        Let $\Phi^{(s,h)} \leftarrow \left\{\pi_i^{(s,h)}\right\}_{i=1}^{N_0}$
        $\pi_{i,h'}(\cdot|s) \leftarrow \text{Unif}(\mathcal{A}_i), \ \ \forall \mu \in \Phi^{(s,h)}, \forall h' \geq h$.
        $\Psi^{\pi_{-i}^E} \leftarrow \Psi^{\pi_{-i}^E} \cup \Phi^{(s,h)}$.
    end for
    for $n = 1 \dots N$ do
        sample policy $\pi_i \sim \text{Unif}(\Psi^{\pi_{-i}^E})$.
        Collect $z_n = (s_1, a_1, a_{-i}, \dots, s_{H+1}) \sim \pi_i, \pi_{-i}^E$.
        $\mathcal{D}^{\pi_{-i}^E} \leftarrow \mathcal{D}^{\pi_{-i}^E} \cup \{z_n\}$
    end for
end for
Receive: datasets $\mathcal{D}^{\pi_{-i}^E}$ for all $i \in [n]$.
Imitation Learning
for $i \in [n]$ do
    Define the dataset $\mathcal{D}^{\pi_i^E} = \cup_{j \neq i} \mathcal{D}^{\pi_{-j}^E}$ to compute

        $$\widehat{\pi}_i = \underset{\pi_i \in \Pi_i}{\operatorname{argmin}} \sum_{s, a_i \in \mathcal{D}^{\pi_i^E}} - \log \pi_i(a_i|s)$$

    where $a_i$'s are sampled from $\pi_i^E(\cdot|s)$.
end for
Return Nash estimate $\widehat{\pi} = (\widehat{\pi}_1, \dots, \widehat{\pi}_n)$.
```

In the first step of the proof, we consider Exploit-Gap, which is the part where we require the reward-free warm-up phase.

In particular, we get

$$
\begin{aligned}
\text{Exploit-Gap} &\overset{(i)}{\leq} H \sum_{h=1}^{H} \sum_{i=1}^{n} \max_{\pi_i \in \mathrm{br}(\widehat{\pi}_{-i})} \mathbb{E}_{s \sim d_h^{\pi_i, \pi^E_{-i}}} \left[ \mathrm{TV}\big(\pi^E_{-i,h}(\cdot \mid s), \widehat{\pi}_{-i,h}(\cdot \mid s)\big) \right] \\
&= H \sum_{h=1}^{H} \sum_{i=1}^{n} \max_{\pi_i \in \mathrm{br}(\widehat{\pi}_{-i})} \Bigg( \sum_{s \in \mathcal{S}^i_{\delta,h}} \sum_{a \in \mathcal{A}} \frac{d_h^{\pi_i, \pi^E_{-i}}(s,a)}{p_h^{\pi^E_{-i}}(s,a)} \mathrm{TV}\left(\pi^E_{-i,h}, \widehat{\pi}_{-i,h}\right)(s) \\
&\qquad\qquad + \sum_{s \notin \mathcal{S}^i_{\delta,h}} \sum_{a \in \mathcal{A}} d_h^{\pi_i, \pi^E_{-i}}(s,a) \mathrm{TV}\left(\pi^E_{-i,h}, \widehat{\pi}_{-i,h}\right)(s) \Bigg) \\
&\overset{(ii)}{\leq} 2H^2 S A_{\max} \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbb{E}_{s \sim p_h^{\pi^E_{-i}}} \left[ \sum_{j \neq i} \mathrm{TV}\big(\pi^E_{j,h}(\cdot \mid s), \widehat{\pi}_{j,h}(\cdot \mid s)\big) \right] + nH^2 S A_{\max}\, \delta \\
&= 2H^2 S A_{\max} \sum_{h=1}^{H} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{E}_{s \sim p_h^{\pi^E_{-j}}} \left[ \mathrm{TV}\big(\pi^E_{i,h}(\cdot \mid s), \widehat{\pi}_{i,h}(\cdot \mid s)\big) \right] + nH^2 S A_{\max}\, \delta \\
&\leq 2n^2 S A_{\max} H^3 \sqrt{\frac{S A_{\max} \log(4S/\delta_{\mathrm{fail}})}{nN}} + n S A_{\max} H^2\, \delta \\
&= 2n^{\frac{3}{2}} S A_{\max} H^3 \sqrt{\frac{S A_{\max} \log(4S/\delta_{\mathrm{fail}})}{N}} + n S A_{\max} H^2\, \delta,
\end{aligned}
$$

where in $(i)$ we used the same argument as in Lemma 6.1 for the $n$ player setting. In $(ii)$ we used the fact that the policies are conditionally independent on $s$. This means that we now bound the TV separately for each player and get in total $(n-1)$ independent bounds, where we again can apply for example [Berend and Kontorovich, 2012, Theorem 2.1]. Moreover, we defined as $\mathcal{S}^i_{\delta,h}$ the set of $\delta$-reachable states at stage $h$ in the MDP $\mathcal{M}^{\pi^{\mathrm{E}}_{-i}}$.

For the value gap we can proceed similarly,

$$
\begin{aligned}
\text{Value-Gap} &\overset{(i)}{\leq} nH \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^E}} \left[\mathrm{TV}\big(\pi_h^E(\cdot \mid s), \widehat{\pi}_h(\cdot \mid s)\big)\right] \\
&= nH \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbb{E}_{s\sim d_h^{\pi^E}} \left[\mathrm{TV}\big(\pi_{i,h}^E(\cdot \mid s), \widehat{\pi}_{i,h}(\cdot \mid s)\big)\right] \\
&= nH \sum_{h=1}^{H}\sum_{i=1}^{n} \Bigg( \sum_{s\in\mathcal{S}_{\delta,h}^i} \sum_{a\in\mathcal{A}_i} \frac{d_h^{\pi_i^E,\pi_{-i}^E}(s,a)}{p_h^{\pi_{-i}^E}(s,a)} \mathrm{TV}\left(\pi_{i,h}^E, \widehat{\pi}_{i,h}\right)(s) \\
&\qquad\qquad + \sum_{s\notin\mathcal{S}_{\delta,h}^i} \sum_{a\in\mathcal{A}_i} d_h^{\pi_{i,h}^E,\pi_{-i,h}^E}(s,a) \mathrm{TV}\left(\pi_{i,h}^E, \widehat{\pi}_{i,h}\right)(s) \Bigg) \\
&\overset{(ii)}{\leq} 2nH^2 SA_{\max} \sum_{h=1}^{H}\sum_{i=1}^{n} \mathbb{E}_{s\sim p^{\pi_{-i}}} \left[\mathrm{TV}\big(\pi_{i,h}^E(\cdot \mid s), \widehat{\pi}_{i,h}(\cdot \mid s)\big)\right] + nH^2 SA_{\max}\,\delta \\
&= 2nH^2 SA_{\max} \sum_{h=1}^{H}\sum_{i=1}^{n} \mathbb{E}_{s\sim p^{\pi_{-i}^E}} \left[\mathrm{TV}\big(\pi_{i,h}^E(\cdot \mid s), \widehat{\pi}_{i,h}(\cdot \mid s)\big)\right] + nH^2 SA_{\max}\,\delta \\
&\leq 2n^2 SA_{\max} H^3 \sqrt{\frac{SA_{\max}\log(4S/\delta_{\text{fail}})}{N}} + nSA_{\max}H^2\,\delta \\
&= 2n^2 SA_{\max} H^3 \sqrt{\frac{SA_{\max}\log(4S/\delta_{\text{fail}})}{N}} + nSA_{\max}H^2\,\delta
\end{aligned}
$$

where steps $(i)$ and $(ii)$ holds exactly for the same reasons used in the upper bound of Exploit-Gap. Combining both parts completes the proof, giving that the total number of expert queries is given by $\mathcal{O}(\frac{n^4S^3A_{\max}^3H^6\log(S/\delta_{\text{fail}})}{(1-\gamma)^6\varepsilon^2})$. □

Some remarks are in order. Again this result needs a reward free warm-up phase. In particular, it requires a dataset for each expert, meaning $n$ datasets where each dataset depends on the other $n-1$ agents. Most importantly, we can see that the number of samples needed does not scale exponentially with the number of agents, instead it only scales quadratically. This is in contrast with learning Nash equilibria in the first places, where it is known that the number of samples scales with $A_{\max}^n$, known as the *curse of multi-agents* [Rubinstein, 2016]. However, this does not contradict with the lower bound as we already have access to data stemming from Nash equilibrium policies which provides additional information. That the lower bound does not hold in these settings has e.g. also been shown in offline general-sum settings [Cui and Du, 2022].

# D Experimental details

In this section, we provide details on the experimental setup used for our provided numerical verifications illustrated in Figure 2.

**Lower bound environment** The first experimental environment corresponds to the lower bound construction described earlier (see Figure 1), with a simplification of the game in state $s_3$. Instead of constructing an $\varepsilon$-perturbed Matching Pennies game, we use a normal-form game with a pure Nash equilibrium and unique value of 1.
Formally, the state space is $\mathcal{S} = \{s_1, s_2, s_3\}$. The action space is $\mathcal{A} = \{a_1, a_2\}$ for player 1 and $\mathcal{B} = \{b_1, b_2\}$ for player 2. The reward function is state-dependent in $s_1$ and $s_2$, with $r(s_1) = r(s_2) = 0$. At $s_3$ the reward structure is given by

$$r(s_3, a, b) := \begin{bmatrix} 1 & 1 \\ 0 & -12 \end{bmatrix},$$

where the row indicates the action of player 1 and the column the action of player 2.
The Nash equilibrium strategy for player 1 in this normal-form game is $\mu_{\text{Nash}}(\cdot \mid s_3) = (1, 0)$, while player 2 can play any strategy, since her expected reward is always $-1$. The unique Nash value from player 1's perspective is therefore 1.
As in the lower bound construction, in state $s_1$ player 2 strictly prefers $b_1$, which deterministically transitions to $s_2$. Consequently, under the Nash equilibrium profile, only states $\{s_1, s_2\}$ are visited, and the Nash value of the Markov game is 0.
If $s_3$ is never visited in the dataset, player 1's recovered policy by BC will be uniform, which can be exploited by player 2 through the best response

$$\nu_{\text{br}}(\cdot \mid s_3) = (0, 1),$$

leading to a reward of $-5.5$ for player 1 and $+5.5$ for player 2. In this case, player 1 is exploitable. Conversely, if $s_3$ is covered in the data, then player 1 requires only a single sample to recover the correct Nash strategy, and the exploitability becomes 0.
The probability of visiting $s_3$ in any given trajectory is $\rho(s_3)$. Thus, the number of trajectories required until $s_3$ is observed follows a geometric distribution with parameter $\rho(s_3)$, yielding an expected sample complexity of $1/\rho(s_3)$. Since $\mathcal{C}_{\max} = 1/\rho(s_3)$, varying $\rho(s_3) \in \{1, 0.5, 0.25, 0\}$ corresponds to $\mathcal{C}_{\max} \in \{1, 2, 4, \infty\}$, which is exactly reflected in the experimental results shown in Figure 2 (a). On the contrary, notice that for all values of $\rho(s_3)$, the value of $\mathcal{C}(\mu^{\text{E}}, \nu^{\text{E}})$ remains constant equal to 2 which is its smallest possible value. By simulating the geometric distribution over 100 runs across varying seeds and with the different parameters described above and tracking its standard deviation, we exactly recover the plots in Figure 2 (a).

**Gridworld** We next describe the setup of the considered zero-sum Gridworld environment, illustrated in Figure 4. The state space is given by the joint positions of the two agents on a 3 × 3 grid, subject to the restriction that both agents cannot occupy the same cell simultaneously. Formally,

$$\mathcal{S} = \{((i, j), (k, l)) \mid (i, j) \neq (k, l),\ i, j, k, l \in \{0, 1, 2\}\},$$

which yields 72 states in total. The action space is identical for both agents and defined as $\mathcal{A} = \{\text{left}, \text{right}, \text{up}, \text{down}\}$. The transition dynamics are deterministic: whenever an action would cause an agent to collide with a wall or with the other agent, the agent remains in its current position. If the initial distribution is chosen such that both agents are equidistant to the goal, the Nash value of the game is 0. In particular, we fix the deterministic starting state $((1, 0), (1, 2))$, from which both players require exactly three steps to reach the goal. Hence, the Nash equilibrium value is 0. Multiple

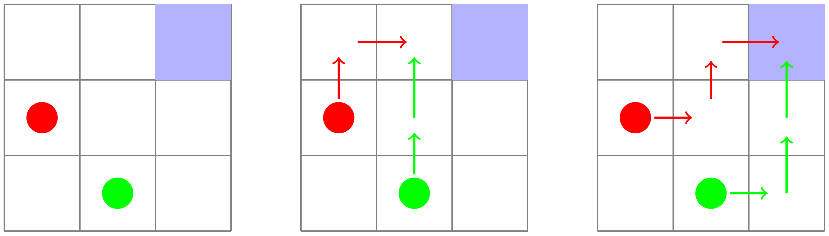


Figure 4: Zero-sum Gridworld environment and different Nash equilibrium paths.

Nash equilibria exist: any pair of paths in which both players ensure that the opponent cannot reach the goal earlier constitutes a Nash equilibrium. This is reflected in Figure 4, which illustrates different Nash paths obtained using zero-sum value iteration. Importantly, these policies also ensure that in all other states, the opponent cannot force an earlier goal arrival. For the second Gridworld experiment, we take convex combinations of different Nash paths to improve state coverage. Since the set of Nash equilibria in zero-sum games is convex, all such convex combinations remain valid Nash equilibria.

Both Gridworld experiments use the same environment specification. The only difference lies in the coverage provided by the expert demonstrations. In both cases, the expert policy is obtained by running zero-sum value iteration, which returns a Nash equilibrium policy pair.

**Algorithm setup** We next detail the implementation of MURMAIL and MAIL-WARM. For MURMAIL, we closely follow the description of Freihaut et al. [2025], and restate their pseudocode for completeness (see Algorithm 4). Note that all experiments have been run on a standard MacBook Pro with Chip M3 and 16GB of RAM. Since the Gridworld environment is more challenging than the lower bound construction considered in their work, we apply two modifications to improve convergence speed: (i) we set the learning rate to $\eta = 50$, and (ii) instead of sampling a single policy per RL inner loop, we average over 100 updates, effectively yielding a batched variant of MURMAIL better suited for larger environments. We set the inner RL loop horizon to $T = 10$.

For MAIL-WARM, we follow Algorithm 1, with the only practical adjustment being that the expert policies are non-stationary. Consequently, in Line 12 we return stationary approximations of the learned policies. Since EULER is not well suited for practice, and as Jin et al. [2020] show that any RL algorithm can be used to solve the $SH$ many RL problems in the reward-free phase, we instead employ Q-learning. Importantly, Q-learning does not require knowledge of the transition dynamics. We run Q-learning for 100 iterations for each RL problem in our experiments.

An interesting empirical observation is that solving the $SH$ many RL problems reveals that many states in the induced expert MDP are not reachable. For example, consider the Nash equilibrium policy illustrated in the middle of Figure 4. Fixing the green agent for the expert induced MDP, all states such that the position of the green agent is the bottom right corner are not reachable. Therefore, the effective visited states can significantly reduce compared to the whole state space.

## E Useful results

For completeness reasons, we provide Theorem 2.1 by Berend and Kontorovich [2012] which we used frequently throughout this work as well as a standard binomial concentration result.

**Algorithm 4** Maximum Uncertainty Response Multi-Agent Imitation Learning (MURMAIL)

```
**Input:**number of iterations $K$, learning rates $\eta$, inner iteration budget $T$, initial $(\mu_1, \nu_1)$
**Receive:**$\varepsilon$-Nash equilibrium $(\hat{\mu}, \hat{\nu})$
**for** $k = 1$ **to** $K$ **do Inner Single-Agent RL Updates:**
% Maximum uncertainty response to $\mu$-player update
Define single agent transition $P_{\mu_k}(s' \mid s, b) = \sum_{a \in \mathcal{A}} \mu_k(a \mid s) P(s' \mid s, a, b)$;
Define single agent stochastic reward $R_{\mu_k}(s) \to \mathbb{1}_{\{A_E = A'_E\}} - 2\mu_k(A_E \mid s) + \|\mu_k(\cdot|s)\|^2$ where $A_E, A'_E \sim \mu^{\mathrm{E}}(\cdot \mid s)$;
$y_k = \texttt{UCBVI}(T, P_{\mu_k}, R_{\mu_k})$;
% Maximum uncertainty response to $\nu$-player update
$P_{\nu_k}(s'|s, a) = \sum_{b \in \mathcal{B}} \nu_k(b|s) P(s' \mid s, a, b)$;
$R_{\nu_k}(s) \to \mathbb{1}_{\{A_E = A'_E\}} - 2\nu_k(A_E \mid s) + \|\nu_k(\cdot \mid s)\|^2$ where $A_E, A'_E \sim \nu^{\mathrm{E}}(\cdot \mid s)$;
$z_k = \texttt{UCBVI}(T, P_{\nu_k}, R_{\nu_k})$
**Update policies:**
Sample $S^\mu_k \sim d^{\mu_k, y_k}$, $A^\mu_k \sim \mu^{\mathrm{E}}(\cdot \mid S^\mu_k)$, $S^\nu_k \sim d^{z_k, \nu_k}$, $A^\nu_k \sim \nu^{\mathrm{E}}(\cdot \mid S^\nu_k)$.
$g^\mu_k(s, a) = \mu_k(a \mid S^\mu_k) \mathbb{1}_{S^\mu_k = s} - \mathbb{1}_{A^\mu_k = a}$
$g^\nu_k(s, a) = \nu_k(a \mid S^\nu_k) \mathbb{1}_{S^\nu_k = s} - \mathbb{1}_{A^\nu_k = a}$
$\mu_{k+1}(a \mid s) \propto \mu_k(a \mid s) \exp\left(-\eta g^\mu_k(s, a)\right) \quad \nu_{k+1}(b \mid s) \propto \nu_k(b \mid s) \exp\left(-\eta g^\nu_k(s, a)\right)$
**end for**
**Return**:$\mu_{\widehat{k}}$, $\nu_{\widehat{k}}$ for $\widehat{k} \sim \mathrm{Unif}([K])$
```

**Lemma E.1** (Concentration Inequality for Total Variation Distance, see e.g. Thm 2.1 by Berend and Kontorovich [2012]). *Let $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\}$ be a finite set. Let $P$ be a distribution on $\mathcal{X}$. Furthermore, let $\widehat{P}$ be the empirical distribution given $m$ i.i.d. samples $x_1, x_2, \cdots, x_n$ from $P$, i.e.,*

$$\widehat{P}(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i = j\}.$$

*Then, with probability at least $1 - \delta$, we have that*

$$\left\|P - \widehat{P}\right\|_1 := \sum_{x \in \mathcal{X}} \left|P(x) - \widehat{P}(x)\right| \leq \sqrt{\frac{2|\mathcal{X}| \log(1/\delta)}{n}}.$$

*Proof.* Define the function $f(x_1, \ldots, x_n) = \sum_{x \in \mathcal{X}} |\widehat{P}(x) - P(x)|$, where $\widehat{P}$ is the empirical distribution. Replacing one sample $x_i$ can change $f$ by at most $2/n$, since the empirical frequencies change by at most $1/n$ per coordinate and total variation sums these differences.
By McDiarmid's inequality, we have for any $\epsilon > 0$,

$$\Pr\left(f - \mathbb{E}[f] \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2}\right).$$

Berend and Kontorovich (2013) show that $\mathbb{E}[f] \leq \sqrt{\frac{|\mathcal{X}|}{n}}$. Setting the failure probability to $\delta$, we solve

$$\exp\left(-\frac{n\epsilon^2}{2}\right) = \delta \quad \Longrightarrow \quad \epsilon = \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Therefore, with probability at least $1 - \delta$,

$$\left\| P - \widehat{P} \right\|_1 \leq \sqrt{\frac{|\mathcal{X}|}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \leq \sqrt{\frac{2|\mathcal{X}|\log(1/\delta)}{n}},$$

□

**Lemma E.2** (Binomial concentration, see e.g. Lemma A.1 by Xie et al. [2021])**.** *Suppose* $N \sim \mathrm{Bin}(n, p)$ *where* $n \geq 1$ *and* $p \in [0, 1]$*. Then with probability at least* $1 - \delta$*, we have*

$$\frac{p}{N \vee 1} \leq \frac{8\log(1/\delta)}{n},$$

*where* $N \vee 1 := \max\{1, N\}$*.*

*Proof.* We consider two cases. Case 1: $p \leq \frac{8\log(1/\delta)}{n}$. As $N \vee 1 \geq 1$, we have $\frac{p}{N\vee 1} \leq p \leq \frac{8\log(1/\delta)}{n}$ almost surely. Case 2: $p > \frac{8\log(1/\delta)}{n}$. Note, that then $\mathbb{E}[N] = np > 8\log(1/\delta)$ and by the multiplicative Chernoff bound, for any $0 < \epsilon < 1$ it holds true that

$$\mathbb{P}\left(N < (1-\epsilon)np\right) \leq \exp\left(-\frac{\epsilon^2}{2}np\right).$$

Now, with $\epsilon = \frac{1}{2}$ we have

$$\mathbb{P}\left(N < (1-\epsilon)np\right) \leq \exp\left(-\frac{np}{8}\right) \leq \delta.$$

Therefore, with probability of at least $1 - \delta$ it holds $N \geq \frac{np}{2}$ and therefore on this event also $\frac{p}{n\vee 1} \leq \frac{2}{n}$. In total we get $\frac{p}{N\vee 1} \leq \frac{8\log(1/\delta)}{n}$. Combining both cases completes the proof. □