On Imitation in Mean-field Games

Giorgia Ramponi
ETH AI Center

Pavel Kolev
MPI for Intelligent Systems

Olivier Pietquin
Google DeepMind

Niao He
ETH Zurich

Mathieu Laurière
NYU Shanghai & Google DeepMind

Matthieu Geist
Google DeepMind

Abstract

We explore the problem of imitation learning (IL) in the context of mean-field games (MFGs), where the goal is to imitate the behavior of a population of agents following a Nash equilibrium policy according to some unknown payoff function. IL in MFGs presents new challenges compared to single-agent IL, particularly when both the reward function and the transition kernel depend on the population distribution. In this paper, departing from the existing literature on IL for MFGs, we introduce a new solution concept called the Nash imitation gap. Then we show that when only the reward depends on the population distribution, IL in MFGs can be reduced to single-agent IL with similar guarantees. However, when the dynamics is population-dependent, we provide a novel upper-bound that suggests IL is harder in this setting. To address this issue, we propose a new adversarial formulation where the reinforcement learning problem is replaced by a mean-field control (MFC) problem, suggesting progress in IL within MFGs may have to build upon MFC.

1 Introduction

Imitation learning (IL) is a popular framework involving an apprentice agent who learns to imitate the behavior of an expert agent by observing its actions and transitions. In the context of mean-field games (MFGs), IL is used to learn a policy that imitates the behavior of a population of infinitely-many expert agents that are following a Nash equilibrium policy, according to some unknown payoff function. Mean-field games are an approximation introduced to simplify the analysis of games with a large (but finite) number of identical players, where we can look at the interaction between a representative infinitesimal player and a term capturing the population's behavior. The MFG framework enables to scale to an infinite number of agents, where both the reward and the transition are population-dependent. The aim is to learn effective policies that can effectively learn and imitate the behavior of a large population of agents, which is a crucial problem in many real-world applications, such as traffic management [12, 30, 31], crowd control [11, 1], and financial markets [6, 5].

IL in MFGs presents new challenges compared to single-agent IL, as both the (unknown) reward function and the transition kernel can depend on the population distribution. Furthermore, algorithms will depend on whether we can only observe the trajectories drawn from the Nash Equilibrium (NE) or if we can access the MFG itself, either driven by the expert population or the imitating one.

The main question we address is whether IL in MFGs is actually harder than IL in single-agent settings and if we can use single-agent techniques to solve IL in the MFGs framework.

Although there exist IL algorithms for MFGs in the literature, none comes with a characterization of the quality of the learnt imitation policy. So as to compare algorithms on a rational basis, we provide an extension of the concept of imitation gap to this setting and study it. Our contributions are:

16th European Workshop on Reinforcement Learning (EWRL 2023).

• We provide a commented review of the existing literature on IL for MFGs. Notably, we will explain that they essentially amount to a reduction to classic IL, and explain the underlying possible issues.

• We introduce a new solution concept for IL in MFGs called *Nash imitation gap*, which is a strict generalization of the classic imitation gap and that we think may be more widely applicable to Multi-agent Reinforcement Learning (MARL).

• In light of this new criterion, we first study the setting where only the reward depends on the population's distribution, while the dynamics does not. This setting was largely studied in the past few years [16, 4, 1, 14, 22], and we show that in this case IL in MFGs reduces to single-agent IL with similar guarantees for Behavioral Cloning (BC) and Adversarial Imitation (ADV) type of algorithms.

• Then, we provide a similar analysis in the more general setting where the dynamics depends on the population's distribution. In this case, we provide for BC and ADV approaches upper-bounds that are exponential in the horizon, suggesting that IL is harder in this setting. On an abstract way, all previous works of the existing literature correspond to this setting.

• Due to these negative results, we introduce a new proxy to the Nash imitation gap, for which we can derive a quadratic upper bound on the horizon. Then, we discuss how a practical algorithm could be designed with an adversarial learning viewpoint. The idea behind it is to use an approach similar to adversarial IL, where the underlying RL problem is replaced by a Mean-Field Control (MFC) problem. We leave the design and experimentation of practical algorithms for future works, but this suggests that making progress on IL in MFGs may have to build upon MFC. We also provide a numerical illustration empirically supporting our claims in the appendix.

2 Background

2.1 Mean-field Games

Intuitively, an MFG corresponds to the limit of an N-player game when N tends towards infinity. We focus on the finite-horizon setting, in discrete time, and with finite state and action spaces. Mathematically, the MFG is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho_0)$ where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ is a transition kernel, $r : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \mathbb{R}$ is a reward function, H is a finite horizon and $\rho_0 \in \Delta_{\mathcal{S}}$ is a distribution over initial states. The first and second inputs of P and r represent respectively the individual player's state and action, while the third input represents the distribution of the population. This allows the model to capture interactions or mean field type. We denote $[H-1] = \{0, \dots, H-1\}$. Since the problem is set in finite time horizon, we consider non-stationary stochastic policies of the form $\pi = (\pi_0, \dots, \pi_{H-1})$ where for each $n \in [H-1]$, $\pi_n : \mathcal{S} \to \Delta_{\mathcal{A}}$. This means that, at time n, a representative player whose state is $s_n \in \mathcal{S}$ picks an action according to $\pi_n(s_n)$. We will also view π_n as a function from $\mathcal{S} \times \mathcal{A}$ to $[0, 1]$ and use the notation $\pi_n(a|s_n)$ to denote the probability to pick action a according to $\pi_n(s_n)$. We also introduce the set of population-independent non-stationary rewards uniformly bound by 1, that will be useful later, $\mathcal{R} = \{r : \mathcal{S} \times \mathcal{A} \times [H-1] \to [-1, 1]\}$.

When a player is at state s and uses action a while the population distribution is ρ, it gets the reward $r(s, a, \rho)$. In a finite-player game, ρ would be replaced by the empirical distribution of the other players' states. These players would be influenced by the player under consideration, leading to complex interactions. However, the influence of a single player on the population distribution becomes smaller as the number of players increases. In the limit, we can expect that each player has no influence on the distribution. We will use interchangeably the terms "agent" and "player".

The MFG framework allows us to formalize this idea. So the problem faced by a single representative player is a Markov Decision Process (MDP), in which the distribution is fixed: Given a mean field sequence $\rho = (\rho_n)_{n \in [H-1]}$, the player wants to find a policy π maximizing the value function V defined by:

$$V(\pi, \rho) = \mathbb{E}\left[\sum_{n=0}^{H-1} r(S_n, A_n, \rho_n)\right]$$

where S_0 is distributed according to the initial state distribution ρ_0, $A_n \sim \pi_n(S_n)$, and $S_{n+1} \sim P(S_n, A_n, \rho_n)$. The resulting policy is called a best response to the mean field ρ. We will sometime write $V_r(\pi, \rho)$ to make explicit to reward of the value function.

Definition 1 (Best response)**.** *A policy π is a best response to ρ if: $V(\pi, \rho) = \max_{\pi'} V(\pi', \rho)$.*

Intuitively, ρ is a Nash equilibrium if it is the distribution sequence obtained when all the players use a best response policy against ρ. To formalize this idea, we introduce the notion of population distribution sequence induced by a policy, that does influence the transition kernel.

Definition 2 (Population distribution sequence)**.** *We denote by $\rho^{(\pi)}$ the state-distribution sequence induced by the population following policy π, which is defined as:*

$$\begin{cases} \rho_0^{(\pi)}(s) = \rho_0(s), \qquad s \in \mathcal{S} \\ \rho_{n+1}^{(\pi)}(s') = \sum_s \rho_n^{(\pi)}(s) \sum_a \pi_n(a|s) P(s'|s, a, \rho_n^{(\pi)}), \qquad (n, s') \in [H-1] \times \mathcal{S}. \end{cases}$$

The state-action distribution sequence, denoted by $\mu^{(\pi)}$, is defined as:

$$\mu_n^{(\pi)}(s, a) = \pi_n(a|s) \rho_n^{(\pi)}(s), \qquad (n, s, a) \in [H-1] \times \mathcal{S} \times \mathcal{A}.$$

We can now give a formal definition of Nash equilibrium.[1]

Definition 3 (Nash equilibrium)**.** *A policy π is called a Nash equilibrium policy if it is a best response against ρ^π, i.e., π is a maximizer of $\pi' \mapsto V(\pi', \rho^\pi)$. The distribution sequence ρ^π induced by a Nash equilibrium policy π is called a Nash equilibrium mean field sequence.*

The value function can be written without expectation by introducing the state and state-action distributions for a single agent, which does not influence the transition kernel (only the population does, it is a core reason for considering the mean-field limit).

Definition 4 (Single-agent distribution sequence)**.** *Consider an agent using policy π' who evolves among a population using policy π. We denote by $\rho^{(\pi)\pi'}$ the state-distribution sequence of this single agent, which is defined by:*

$$\begin{cases} \rho_0^{(\pi)\pi'}(s) = \rho_0(s), \qquad s \in \mathcal{S} \\ \rho_{n+1}^{(\pi)\pi'}(s') = \sum_s \rho_n^{(\pi)\pi'}(s) \sum_a \pi'_n(a|s) P(s'|s, a, \rho_n^{(\pi)}), \qquad (n, s') \in [H-1] \times \mathcal{S}. \end{cases}$$

The state-action distribution sequence is defined as:

$$\mu_n^{(\pi)\pi'}(s, a) = \pi'_n(a|s) \rho_n^{(\pi)\pi'}(s), \qquad (n, s, a) \in [H-1] \times \mathcal{S} \times \mathcal{A}.$$

From these definitions we directly have that $\rho^{(\pi)\pi} = \rho^{(\pi)}$ and $\mu^{(\pi)\pi} = \mu^{(\pi)}$. Furthermore, for two policies π and π', $V(\pi', \rho^{(\pi)}) = \sum_{n=0}^{H-1} \sum_{s,a} \mu_n^{(\pi)\pi'}(s, a) r(s, a, \rho_n^{(\pi)})$.

Definition 5 (Exploitability)**.** *The exploitability of a policy π quantifies the gain for a representative player to replace its policy by a best response:*

$$\mathcal{E}(\pi) = \max_{\pi'} V(\pi', \rho^{(\pi)}) - V(\pi, \rho^{(\pi)}).$$

A Nash equilibrium policy can be defined equivalently as a policy such that its exploitability is 0.

Finding a Nash Equilibrium policy is different from trying to maximize the value function, which can be interpreted as a social optimum. This problem is sometimes referred to as mean field control problem (MFC) because it can be interpreted as an optimal control for an MDP where the state is augmented with the distribution.

Definition 6 (Social optimum)**.** *A policy π is socially optimal if $V(\pi, \rho^\pi) = \max_{\pi'} V(\pi', \rho^{\pi'})$.*

In general the Nash equilibrium and the social optimum do not coincide, i.e., the MFG policy and the MFC solution can be different. In other words, if π is an MFG solution, then we might have $V(\pi, \rho^\pi) < \max_{\pi'} V(\pi', \rho^{\pi'})$, and if π is an MFC policy, then we might have $\mathcal{E}(\pi) \neq 0$.

In this work, we will frequently make two common assumptions (e.g., [2, Asm. 1], or [34, Asm. 1]).

[1] We will sometimes omit the term "Nash" and simply write "equilibrium" when the context is clear.

Assumption 1. *The reward function r and the transition kernel P are Lipschitz continuous w.r.t. the population distribution and have corresponding Lipschitz constants L_r and L_P. In particular, for any state-action pair (s, a) it holds for any state distributions $\rho, \rho' \in \Delta_{\mathcal{S}}$ that*

$$|r(s, a, \rho) - r(s, a, \rho')| \leq L_r \|\rho - \rho'\|_1.$$
$$\|P(\cdot|s, a, \rho) - P(\cdot|s, a, \rho')\|_1 \leq L_P \|\rho - \rho'\|_1.$$

These assumptions mean that the reward and the transitions depend on the mean field in a smooth way. For example, $r(s, a, \rho) = -c_r \rho(s) + \tilde{r}(s, a)$ satisfies the assumption with $L_r = c_r$. A reward of this form penalizes the agent for being in a crowded state. In particular, these assumptions imply that r and P are continuous with respect to the distribution and since $\mathcal{S}$ and $\mathcal{A}$ are finite sets and we consider a discrete-time, finite horizon MFG, these assumptions are sufficient to ensure existence of a Nash equilibrium, see e.g. [10, Proposition 1].

Note that these assumptions do not assume that L_r or L_P need to be small. However, we will consider separately the case $L_P = 0$, which has received attention in the MFG literature and which corresponds to situations where the mean-field interactions occur through the reward only. Note that when $L_P = L_r = 0$, then there are no interactions so the Nash equilibrium condition becomes trivial and the problem reduces to a single agent MDP. We will write the value function as $V(\pi)$.

2.2 Classic imitation learning

The single-agent setting is a special case of the above, with $L_P = L_r = 0$. In this case, the Imitation Learning (IL) problem has been extensively studied [20]. In IL we assume that we observe state-action trajectories generated by an expert who is using an optimal policy π^E, i.e., $\pi^E \in \operatorname{argmax} V(\pi)$. We do not know π^E, V, r nor P. The goal is to learn a policy, denoted by π^A, which performs as well as the expert policy π^E according to the unknown reward: $V(\pi^A) = V(\pi^E) = \max_{\pi'} V(\pi')$.

Given the imitation policy π^A, the *imitation gap* is a non-negative quantity defined by $V(\pi^E) - V(\pi^A)$ [29]. The goal is to learn a policy π^A whose imitation gap is as close to 0 as possible. However, when we do not know the model, this quantity cannot be minimized directly. For this reason, different methods were introduced in literature. Here we focus on the following two prominent methods: Behavioral Cloning (BC) [24] and adversarial IL such as Generative Adversarial IL (GAIL) [18].

For simplicity, we will denote $\rho^E = \rho^{(\pi^E)\pi^E}$, $\mu^E = \mu^{(\pi^E)\pi^E}$, $\rho^A = \rho^{(\pi^A)\pi^A}$, and $\mu^A = \mu^{(\pi^A)\pi^A}$. When the dynamics does not depend on the population ($L_P = 0$), we will write $\rho^\pi = \rho^{(\pi^{\text{any}})\pi}$, $\mu^\pi = \mu^{(\pi^{\text{any}})\pi}$, as the population driven by π^{any} has no effect on the transition kernel.

Behavioral Cloning. In this work, we frame BC as minimizing the expected ℓ_1-distance between the action probability distributions of the expert and the imitation policy, where the expectation is over the expert state occupancy. Although in practice it consists in a reverse KL-divergence, or equivalently as the maximum likelihood estimation in supervised learning, here we consider the ℓ_1-norm distance (or scaled total variation), which is more convenient for the analysis. However, the analysis could be adapted to this maximum likelihood estimation with minor changes thanks to Pinsker's inequality. In the finite-horizon case, one obtains a bound on quantities of the form:

$$\epsilon_n^{BC} := \mathbb{E}_{s \sim \rho_n^E}[\|\pi_n^A(s) - \pi_n^E(s)\|_1], \qquad n \in \{0, \ldots, H-1\}.$$

That is, we solve one BC problem per time-step, from data generated by an expert. The single-agent imitation-learning gap is bounded by $\mathcal{O}\left(\epsilon^{BC} H^2\right)$, where $\epsilon^{BC} = \max_{n \in \{0,\ldots,H-1\}} \epsilon_n^{BC}$ [26]. We will retrieve this result as a special case of our analysis.

Adversarial Imitation Learning. Using adversarial-like approaches [18, 13] we control a distance or divergence between the state-action occupancy measures of the expert and the imitation policy. In the single-agent IL setting this divergence can be expressed as:

$$\epsilon_n^{ADV} := \|\mu_n^E - \mu_n^A\|_1, \qquad n \in \{0, \ldots, H-1\}. \tag{1}$$

Similar IL quantities are extensively studied in the single-agent case, leading to algorithms minimizing a divergence or distance such as Generative Adversarial Imitation Learning [18] or IQ-learn [13]. If we control the state-action occupancy measure errors, then we get a single-agent imitation-learning gap of order $\mathcal{O}\left(\epsilon^{ADV} H\right)$, where $\epsilon^{ADV} = \max_{n \in [0,\ldots,H-1]} \epsilon^{ADV}$, gaining an H factor compared to the BC bound [32].

To make the link to adversarial approaches, controlling the terms of Eq. (1) can be achieved through the following equivalent formulation, based on the Integral Probability Metric (IPM) formulation of the total variation (see Appx. A for a detailed derivation and additional discussions):

$$\min_{\pi} \sum_{n=0}^{H-1} \|\mu_n^E - \mu_n^\pi\|_1 = \max_{f \in \mathcal{R}} \min_{\pi} (V_f(\pi^E) - V_f(\pi)), \tag{2}$$

with $\mathcal{R}$ defined in Sec. 2.1 (non-stationary population-independent rewards functions uniformly bounded by 1). This form is similar to classic adversarial imitation approach, learning both a reward function and a policy, the inner problem being a reinforcement learning (RL) problem. Previous work usually consider an f-divergence [15, 19] between state-action occupation measures, which leads to different min-max problems. However, in this paper, we will focus on the ℓ_1-distance, that we think is a meaningful and practical abstraction of adversarial approaches. Our results could be extended to other settings by using tools like the Pinsker inequality.

Remark 1. *Remember that in the single-agent case (same as when we have $L_P = L_r = 0$) the transition dynamics and the reward function do not depend on the population distribution. In the mean-field game, on the other hand, the transition dynamics and the reward function can depend on the population distribution. For this reason, we need to consider different quantities to be controlled. We will see more about it in the next section.*

3 Related works

As a first contribution, we provide a commented review of the literature focusing on the fundamentals of the different approaches, their pros and their cons.

To the best of our knowledge, the first work addressing the IL problem in MFGs is [33]. The authors consider a discrete-time MFG over a complete graph, and they propose a reduction from MFGs to finite-horizon MDP (with a population-augmented state) and then use single-agent IL algorithms on the new MDP. In the reduction, the new reward function is computed in the following way: $\overline{r}(\rho, \pi) = \sum_s \rho(s) \sum_a \pi(a|s) r(s, a, \rho)$, considering that the state is the population distribution ρ and the actions are the possible policies π. Using this reduction, the authors implicitly assume that the observed expert is solving an MFC problem, i.e. she is looking for a socially optimal policy (see Def. 6). However, in general, this does not actually coincide with a Nash Equilibrium policy (see Def. 3). Then, the reduction works only for cooperative MFGs, but it is prone to biased reward inference in non-cooperative environments.

The first work enlightening this issue is [8]. The authors consider a discounted finite-horizon MFG and propose a novel method called Mean Field IRL (MFIRL). They reframe the problem as finding a reward function that makes the expert policy π^E the best response with respect to the expert population distribution ρ^E, i.e., find a population-dependent r such that $V(\pi^E, \rho^E) \in \operatorname{argmax}_\pi V(\pi, \rho^E)$. To solve this problem they use a max-margin approach similar to [25] in the single-agent case. In fact, fixing the population distribution ρ^E, the MFG is reduced to an MDP, and the imitation problem is reduced to single-agent IL. However, since they do not have access to the actual population distribution ρ^E, they need to estimate it from samples. In this way, the authors do not have any actual theoretical guarantee on the performances of the recovered policy, since it depends on the estimation of ρ^E.

Recently, [9] proposed a novel approach reusing ideas from adversarial learning [18] and maximum entropy [36]. The authors assume they have access to an expert optimizing a mean-field-regularized Nash-equilibrium, i.e., the value-function has an additional regularization term: $V(\pi', \rho^\pi) = \sum_{n=0}^{H-1} \sum_s \left\{ \mathcal{H}(\pi'(\cdot|s)) + \sum_a \mu_n^{(\pi)\pi'}(s, a) r(s, a, \rho_n^{(\pi)}) \right\}$, with $\mathcal{H}$ the Shannon entropy. This term guarantees that the MFG has a unique equilibrium solution, and only one best response policy. Then, they assume to have access to the population distribution ρ^E, or to be able to estimate this population distribution from samples. Fixing the population distribution, the MFG problem reduces to the single-agent setting, which allows applying GAIL [18] or maximum entropy IRL to learn an approximate policy π^A. This leads to two main problems. (i) We are actually estimating ρ^E from samples, trying to find a policy which gives us the true ρ^E; this circulating reasoning leads to many estimation errors. (ii) From an abstract viewpoint, the final theoretical guarantee we can hope to have for the recovered policy π^A would approximately be $\|\mu^{(\pi^E)\pi^A} - \mu^{(\pi^E)\pi^E}\|_1$ (in practice for a different distance or divergence, but involving the same occupancy measures), which may not

be sufficient as we will show in Sec. 4.2. We will see that the upper-bound for this case has an exponential dependency on the horizon.

In [35], the authors assume observing an MFG where the agents are acting according to a correlated equilibrium. The authors justify the study of a correlated equilibrium by some applications where we can have access to some correlation device (e.g., traffic network equilibrium induced from the public routing recommendations). Similar to [9], the authors of this work reuse ideas from adversarial learning [18] to recover a policy observing a correlated equilibrium policy. As in [9] they fix the population distribution to be the expert one and then apply single-agent GAIL to the problem. Then, although the solution concept is different, this approach faces similar problems as in [9], where instead the authors considered to observe an expert following a Nash Equilibrium policy. We will focus on experts achieving a Nash equilibrium. Extending our results (see Sec. 4) to more general equilibria such as coarse correlated ones is an interesting future research direction.

Notice also that all GAIL-like approaches discussed above learn an intermediate population-dependent reward function. This is complicated (as the population is a distribution over the state space, generally difficult to represent compactly), and in fact superfluous in their setting. Indeed, as all these approaches assume that the interaction is done with the MFG driven by the expert population (or a given approximation of it), it is sufficient to consider an intermediate population-independent but non-stationary reward. This will appears clearly in the IPM formulations we propose in Sec. 4.

4 Nash imitation gap and imitation in MFGs

We consider the imitation learning problem in MFGs. Similar to single-agent IL, we observe the interactions between an expert and a fixed MFG environment, for which we do not know the reward function r nor the transition kernel P. We only observe samples coming from an expert policy denoted by π^E, which we assume to be a Nash equilibrium policy, i.e., $\mathcal{E}(\pi^E) = 0$. In single-agent IL the goal is clear: Find a policy π^A to minimize the imitation gap $\max_\pi V(\pi) - V(\pi^A)$. However, in MFG, the goal is less clear. Previous works focused on solving IL for a single agent in a population that stays at equilibrium, but this is not relevant for many applications, in which the population might use the learnt policy. The learnt policy should thus not only be good at the individual level, but it should also be an equilibrium policy. As a first contribution, we propose a natural formulation for studying the performance of imitation learning in MFG, called Nash imitation gap:

Definition 7. *The Nash imitation gap (NIG) of a policy π^A is defined as:*

$$\mathcal{E}(\pi^A) = \max_{\pi' \in \Pi} V(\pi', \rho^{(\pi^A)}) - V(\pi^A, \rho^{(\pi^A)}).$$

Therefore, the NIG is simply defined as being the exploitability of the considered policy. The NIG has many interesting and useful properties. (i) If it is zero, the recovered policy π^A is a Nash equilibrium policy. (ii) It is a generalization of the single-agent imitation gap (see above and Section 2.2), which is recovered as a special case when $L_P = L_r = 0$.

As for the single-agent imitation gap, we cannot optimize it directly, since we do not know the reward function, but instead we can envision proxys, such as reducing the distance between the recovered policy π^A and the expert policy π^E (BC-like) or their occupancy measures (GAIL-like) as in the classic IL setting (see Sec. 2.2). In this section, we first discuss the imitation learning problem when the dynamics does not depend on the population, i.e. $L_P = 0$, a common setting largely explored in the last years [22, 23]. Then, we present our results for the general setting when the dynamics depends on the population, i.e. when $L_P > 0$.

For what follows, in addition to the Lipschitz assumption (Asm. 1), we will also assume that the unknown reward function r for which the expert π^E is a Nash equilibrium is uniformly bounded when the population is the expert one.

Assumption 2. *The unknown reward r satisfies*

$$\max_{s,a} |r(s, a, \rho^{(\pi^E)})| \leq r_{max}.$$

4.1 Population-independent dynamics: a reduction to classic imitation

We start by analyzing a simpler but commonly used setting (e.g., see [23] in the context of reinforcement learning methods, or [16, 4, 1] in the context of the analysis of discrete or continuous space

MFGs), where the dynamics does not depend on the population. Here the MFG interaction is without reward and it is only the (unknown) reward function that depends on the population. This setting is equivalent to observing an interaction between an expert and an MDP, since the state-distribution does not depend on the population, i.e., $\rho^{(\pi)\pi'} = \rho^{\pi'}$.

Behavioral Cloning. The behavioral cloning setting is the same as the single agent setting (see Sec. 2.2), i.e, we control:

$$\epsilon_n^{\text{BC}} := \mathbb{E}_{s\sim\rho_n^E}[\|\pi_n^A(s) - \pi_n^E(s)\|_1], \qquad n \in \{0, \dots, H-1\}, \tag{3}$$

where π^E is the expert policy and π^A the imitation policy. Under the assumption of the BC-type error, we give the following bound on the Nash imitation gap (proof in Appx. B).

Theorem 1. *Let $\epsilon^{BC} = \max_{n\in\{0,\dots,H-1\}} \epsilon_n^{BC}$. If $L_P = 0$, the Nash imitation gap satisfies*

$$\mathcal{E}(\pi^A) \leq H^2(r_{max} + 2L_r)\epsilon^{BC}.$$

Theorem 1 shows that when $L_P = 0$ we recover the single-agent imitation learning bound. Perhaps surprisingly, the simple BC approach was not previously considered for MFG, to the best of our knowledge.

Adversarial learning. In the adversarial setting, similar to the single-agent case (see Sec. 2.2) we control a distance or divergence between occupancy measures. More precisely, we consider the following distance between occupancy measures:

$$\epsilon_n^{\text{ADV}} := \|\mu_n^{\pi^A} - \mu_n^{\pi^E}\|_1, \qquad n \in \{0, \dots, H-1\}.$$

It is important to recall that in this setting, the dynamics do not depend on the population, thus $\mu^{(\pi)\pi''} = \mu^{(\pi')\pi''}$ for every triplet of policies (π, π', π'').

Since in this case the dynamics do not depend on the population distribution, the same IPM approach (see Eq. (2)) of single-agent imitation-learning also works in this context. Then we can practically solve the MFG IL problem using GAIL-like approaches [18, 15, 19, 13].

We now provide a novel bound on the Nash imitation gap (proof in Appx. B).

Theorem 2. *Let $\epsilon^{ADV} = \max_{n\in\{0,\dots,H-1\}} \epsilon_n^{ADV}$. If $L_P = 0$, the Nash imitation gap satisfies*

$$\mathcal{E}(\pi^A) \leq (2L_r + r_{\max})H\epsilon^{ADV}.$$

In contrast to the quadratic horizon dependence in Theorem 1, we derive here a linear horizon dependence. Furthermore, the bound in Theorem 2 is almost the same (similarly to the BC case) as in the single-agent IL problem. In fact, it recovers the bound of [32] by setting $L_r = 0$.

Discussion. When $L_P = 0$ interacting with the MFG without reward amounts to interact with and MDP without reward, so from a practical aspect any classic IL approach could be used, including Dagger-like approaches [27] that we do not analyse here. Our upper bounds show that when the dynamics is independent of the population, IL for MFG has similar guarantees as in single-agent imitation learning. In fact, our results suggest that a population-dependent unkown reward function affects the IL policy performance very moderately (through the L_r constant). However, new challenges may arise if we also want to recover the reward function, as in the case of Inverse Reinforcement Learning. We leave this interesting research direction for future work.

4.2 Population-dependent dynamics

When the dynamics of the MFG depend on the population, i.e., $L_P > 0$, then the previous results do not apply anymore. In this section, we present results on MFGs with population-dependent dynamics, for the same proxys introduced above (BC and adversarial).

Behavioral Cloning. The BC proxy to the NIG is the same as in Equation (3). In this case, however, the bound we get is no longer comparable to the classic IL setting (proof in Appx. B).

Theorem 3. *Let $\epsilon^{BC} = \max_{n\in\{0,\dots,H-1\}} \epsilon_n^{BC}$. If $L_P > 0$, the Nash imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq \left(H^2 r_{max} + 2(1+L_P)^H \frac{L_r + r_{\max}}{L_P^2}\right)\epsilon^{BC}.$$

We observe that when the dynamics depend on the population, the dependence on the horizon is no longer quadratic but exponential. From a technical viewpoint, this comes from the dependency of the transition kernel to the population. A worse dependency makes sense intuitively. In the classic IL setting, an error on policies will amplify at the occupancy measure level (drift phenomenon in imitation learning). Here, this problem is even more amplified by the fact that the transition kernel that defines the occupancy measure itself depends on the related population, amplifying even more the imitation error. Appx. C provides some empirical evidence that IL in MFGs is harder for BC as L_P increases.

Although we cannot claim our result is tight, this suggests that in MFGs we cannot hope to use BC to obtain a good imitation policy π^A, but we need to control the divergence between the state-action occupancy induced by π^A and the one induced by π^E.

Adversarial learning. In contrast to the population-independent case, the quantity to control in this setting is not trivial. In fact, it depends on the MFG interaction assumption. We can assume to have the possibility to interact with only the MFG or with the MFG driven by ρ^E. These two kind of interactions lead us to consider two different errors. We start by considering the error we would like to minimize if we can interact with the MFG driven by the expert population ρ^E. Assuming access to this MFG, the IL problem reduces to doing classic IL in an MDP without reward and with transition distribution $P(\cdot|s, a, \rho^{\pi^E})$. The error is defined as follows:

$$\epsilon_n^{\text{vanilla-ADV}} := \|\mu_n^{(\pi^E)\pi^A} - \mu_n^{(\pi^E)\pi^E}\|_1, \quad n \in \{0, \dots, H-1\}.$$

This assumption was implicitly made in all previous works (see Sec. 3), where the authors assume to fix the expert distribution ρ^E and then solve the IL problem. However, assuming to have access to the expert distribution may not be reasonable in practice, and due to this in [9, 8] the authors replace the expert population distribution ρ^E with its approximation from sampling $\hat{\rho}^E$, in order to learn $\mu^{(\pi^E)\pi^A}$. In reality, however, we seek to learn $\mu^{(\pi^E)\pi^A}$ and then ρ^E, which is the ultimate goal of the problem. This circular reasoning leads to not easily having theoretical guarantees for this setting.

In this case, for obtaining an adversarial formulation, we can use a similar approach as the one for the non-population dependent dynamics, Eq. (2), providing (details in Appx. A):

$$\min_{\pi} \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi}\|_1 = \max_{f \in \mathcal{R}} \min_{\pi} (V_f(\pi^E, \rho^{(\pi^E)}) - V_f(\pi, \rho^{(\pi^E)})).$$

We can observe that the inner problem is again an RL problem (for the MDP induced by $\rho^{(\pi^E)}$), and in practice any single-agent adversarial approach could be applied to solve for a similar proxy (related to a different min-max problem).

We provide a bound for this approach (proof in Appx. B).

Theorem 4. *Let $\epsilon^{\text{vanilla-ADV}} = \max_{n=[0,\dots,H-1]} \epsilon_n^{\text{vanilla-ADV}}$. If $L_P > 0$ the NE imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq r_{\max} H \epsilon^{\text{vanilla-ADV}} + 2(1 + L_P)^H \frac{r_{\max} + L_r}{L_P} \epsilon^{\text{vanilla-ADV}}.$$

Therefore, in this case, if in practice we can apply GAIL-like approaches to the MDP with transition distribution $P(\cdot|s, a, \rho^E)$, we have an exponential dependency in the horizon. Interestingly this is close to the quantity that previous works [9, 8] tried to control. Although we do not know if this bound is tight, reasonably controlling this quantity can lead to weak theoretical guarantees. Appx. C provides empirical evidences that, when H and L_P are large enough, vanilla-ADV behaves much like BC.

4.3 Population-dependent dynamics: a new efficient proxy

Although the vanilla-ADV error is a reasonable quantity to control, another interesting proxy consists in considering the occupancy measures induced when the population is the considered policy (and not always the expert one):

$$\epsilon_n^{\text{MFC-ADV}} := \|\mu_n^{(\pi^A)\pi^A} - \mu_n^{(\pi^E)\pi^E}\|_1, \quad n \in \{0, \dots, H-1\}.$$

To control this quantity, we do not need to have access to the expert population (except through data), we can directly interact with the MFG, driven by what we are learning. As we explained in Sec. 4.1, if $L_P = 0$ then this quantity is the same as the vanilla-ADV error. Before presenting and discussing an adversarial formulation (that will explain the naming choice for the error), we provide a novel bound for this quantity (proof in Appx. B):

Theorem 5. *Let* $\epsilon^{\text{MFC-ADV}} = \max_{n \in \{0,\dots,H-1\}} \epsilon_n^{\text{MFC-ADV}}$*. If* $L_P > 0$*, the Nash imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq (2L_r + r_{\max})H\epsilon^{\text{MFC-ADV}} + 3L_P r_{\max} H^2 \epsilon^{\text{MFC-ADV}}.$$

This gives us a significant improvement compared to the BC case and the vanilla-ADV error. Indeed, the dependency to the horizon is now quadratic and not anymore exponential. Comparing this result with classic imitation learning we have worse dependency on H, since in the classic setting is only linear (we can recover this dependency when $L_P = 0$). However, as explained before, the fact that the transition kernel does depend on the population indeed intuitively implies a larger amplification of errors. Appx. C provides empirical evidences that this approach is better than the two previous ones, with much less influence of H and L_P.

Now, we provide an adversarial formulation, to get a sense of what a practical algorithm could look like. Using again an IPM argument, we have that (details in Appx. A)

$$\min_{\pi} \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi)\pi}\|_1 = \min_{\pi} \max_{f \in \mathcal{R}} (V_f(\pi^E, \rho^{(\pi^E)}) - V_f(\pi, \rho^{(\pi)})).$$

Notice that it is not obvious if we can switch the min and the max here, due to the underlying set of policy-induced occupancy measures being not necessarily convex (due to the dependency of the dynamics on the population). Assuming we can, we still learn an intermediate population-independent non-stationary reward, but now the underlying control problem is no longer RL, it is an MFC problem, as it implies studying $\max_\pi V_f(\pi, \rho^{(\pi)})$, where the population does depend on the optimized policy. This suggests that in practice one could start from a classic adversarial IL approach, and replace the underlying RL algorithm by an inner MFC algorithm (e.g. [28, 7, 21, 17, 3]). We leave the design and implementation of such a practical algorithm for future work, which may be more complex than a straight replacement of the control part (notably, many MFC approaches learn a population-dependent policy). However, this overall suggests that making progress on IL for MFGs may have to build upon MFC.

5 Conclusion

In this paper, we have studied the recent question of imitation learning in mean-field games. We have reviewed the few previous works tackling this problem and provided a critical discussion on the proposed approach. We then have introduced the new solution concept of Nash imitation gap to quantify the quality of imitation. In the simpler case of a population-independent dynamics, we have shown that the problem basically reduces to single-agent imitation learning, and that abstractions of the canonical BC and adversarial approaches come with a similar performance guarantee. In the harder population-dependent case, we have provided upper-bounds that are *exponential* in the horizon, for both BC and adversarial IL, the latter being the approach adopted in previous work. We also introduce a new proxy that amounts to control different occupancy measures, not solely driven by the expert population but by the policy-induced population. From a practical viewpoint, it only implies interacting with the MFG (without access to the expert population for driving the dynamics), and from a theoretical viewpoint it enjoys a much better quadratic dependency to the horizon. The associated adversarial formulation suggests that classic adversarial IL approaches could be adapted by replacing the inner RL loop by an MFC one.

This works open a number of interesting research questions. Our last result suggests a simple modification of existing adversarial approaches, but in practice it may be more difficult than just replacing the control part, and could call for additional research in MFC and MFGs. Maybe also that controlling other kinds of occupancy measures, or even different quantities related to the policy and dynamics, may lead to even better guarantees, ideally recovering the linear dependency to the horizon of the single-agent case. We provided upper-bounds, but there is no known associated lower-bound in this setting, which is another interesting research direction. We focused on experts achieving a Nash equilibrium, and extending our work to different kind of equilibria is another possible direction. We focused on imitating the expert, not on recovering the underlying optimized reward (inverse RL). Doing so may require a different approach, as our adversarial formulations all rely on population-independent rewards (which is convenient from a practical viewpoint, whenever one just wants to recover the policy). We also plan to work on the extension of Nash imitation gap to other kinds of games.

References

[1] Yves Achdou and Jean-Michel Lasry. Mean field games for modeling crowd motion. *Contributions to partial differential equations and applications*, pages 17–42, 2019.

[2] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

[3] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

[4] Pierre Cardaliaguet and P Jameson Graber. Mean field games systems of first order. *ESAIM: Control, Optimisation and Calculus of Variations*, 21(3):690–722, 2015.

[5] Rene Carmona. Applications of mean field games in financial engineering and economic theory. *arXiv preprint arXiv:2012.05237*, 2020.

[6] René Carmona and Mathieu Laurière. Deep learning for mean field games and mean field control with applications to finance. *arXiv preprint arXiv:2107.04568*, 2021.

[7] René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[8] Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. Individual-level inverse reinforcement learning for mean field games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 253–262, 2022.

[9] Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654*, 2021.

[10] Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

[11] Christian Dogbé. Modeling crowd dynamics by the mean-field limit approach. *Mathematical and Computer Modelling*, 52(9-10):1506–1520, 2010.

[12] Mohammed Elhenawy, Ahmed A Elbery, Abdallah A Hassan, and Hesham A Rakha. An intersection game-theory-based traffic control algorithm in a connected vehicle environment. In *2015 IEEE 18th international conference on intelligent transportation systems*, pages 343–347. IEEE, 2015.

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

[14] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 489–497, 2022.

[15] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR, 2020.

[16] Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, 93(3):308–328, 2010.

[17] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021.

[18] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems (NeurIPS)*, 2016.

[19] Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pages 313–329. Springer, 2021.

[20] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[21] Barna Pasztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*, 2021.

[22] Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1028–1037, 2022.

[23] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

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

[25] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006.

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

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

[28] Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 251–259, 2019.

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

[30] Takashi Tanaka, Ehsan Nekouei, and Karl Henrik Johansson. Linearly solvable mean-field road traffic games. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 283–289. IEEE, 2018.

[31] Huisheng Wang, Yuejiang Li, and H Vicky Zhao. Research on intelligent traffic control methods at intersections based on game theory. *arXiv preprint arXiv:2009.05216*, 2020.

[32] Tian Xu, Ziniu Li, and Yang Yu. Error bounds of imitating policies and environments. *Advances in Neural Information Processing Systems*, 33:15737–15749, 2020.

[33] Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. Learning deep mean field games for modeling large population behavior. *arXiv preprint arXiv:1711.03156*, 2017.

[34] Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. *arXiv preprint arXiv:2212.14449*, 2022.

[35] Zhiyu Zhao, Renyuan Xu, Haifeng Zhang, Jun Wang, Mingyuan Zhang, and Yaodong Yang. Imitation learning for mean field games with correlated equilibria. *openreview preprint: id=VUdMeSbExWg*, 2023.

[36] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 3*, pages 1433–1438, 2008.

Appendix

Table of Contents

A More details on IPMs **12**
- A.1 Classic imitation setting . . . 12
- A.2 MFG adversarial imitation when $L_P = 0$ . . . 13
- A.3 MFG adversarial imitation when $L_P > 0$ . . . 13

B Proofs of stated theoretical results **14**
- B.1 Decomposition of the exploitability . . . 14
- B.2 Proofs for the case $L_P = 0$ . . . 16
- B.3 Proofs for the case $L_P > 0$ . . . 17
- B.4 Auxiliary lemmas . . . 20

C Numerical illustration **24**

A More details on IPMs

In this section, we provide a detailed derivation and additional discussions of the adversarial viewpoint of the minimization of the ℓ_1-distance between occupancy measures. Generally speaking, for a finite set X and Δ_X the associated simplex, we can express the ℓ_1-norm between probability distributions (their total variation) as an integral probability metric (IPM). For any $p, q \in \Delta_X$, we have

$$\|p - q\|_1 = \sup_{f: X \to [-1,1]} (\mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{x\sim q}[f(x)]).$$

This will be the building block for framing an adversarial formulation of imitation learning, and making links to classic approaches such as GAIL, even if they consider usually a different framework (single agent with γ-discounted infinite horizon).

A.1 Classic imitation setting

First, we consider the classic IL setting depicted in Sec. 2.2. Let $f : \mathcal{S} \times \mathcal{A} \times [H-1] \to \mathbb{R}$ be a (non-stationary) reward, we recall that the value function can be written as

$$V_f(\pi) = \sum_{s,a} \mu_n^\pi(s,a) f_n^\pi(s,a).$$

We claimed that adversarial IL can be framed as minimizing for all time step n the distance $\|\mu_n^E - \mu_n^A\|$. To see this, for a policy π with associated sequence of occupancy measures $\mu_0^\pi, \dots, \mu_{H-1}^\pi$, define

$$\tilde{\mu}^\pi = \frac{1}{H} \begin{pmatrix} \mu_0^\pi & \dots & \mu_{H-1}^\pi \end{pmatrix} \in \Delta_{\mathcal{S}\times\mathcal{A}\times[H-1]}.$$

Recall the set $\mathcal{R}$ defined in Sec. 2.1, $\mathcal{R} = \{\mathcal{S} \times \mathcal{A} \times [H-1] \to [-1,1]\}$. Using the IPM viewpoint, we can write

$$\begin{aligned}
\|\tilde{\mu}^E - \tilde{\mu}^\pi\|_1 &= \max_{f\in\mathcal{R}} (\mathbb{E}_{s,a\sim\tilde{\mu}_n^E}[f_n(s,a)] - \mathbb{E}_{s,a\sim\tilde{\mu}_n^\pi}[f_n(s,a)]) \\
&= \max_{f\in\mathcal{R}} \frac{1}{H} (\sum_{s,a} \mu_n^E(s,a) f_n(s,a) - \sum_{s,a} \mu_n^\pi(s,a) f_n(s,a)) \\
&= \max_{f\in\mathcal{R}} \frac{1}{H} (V_f(\pi^E) - V_f(\pi)).
\end{aligned}$$

Therefore, we can frame the imitation learning problem as finding a non-stationary policy $\pi^A \in \text{argmin}_\pi \|\tilde{\mu}^E - \tilde{\mu}^\pi\|_1$, which amounts to solve

$$\begin{aligned}\min_\pi \sum_{n=0}^{H-1} \|\mu_n^E - \mu_n^\pi\|_1 &= \min_\pi H\|\tilde{\mu}^E - \tilde{\mu}^\pi\|_1 \\ &= \min_\pi \max_{f\in\mathcal{R}} (V_f(\pi^E) - V_f(\pi)) \\ &= \max_{f\in\mathcal{R}} \min_\pi (V_f(\pi^E) - V_f(\pi)),\end{aligned}$$

where the last equation holds because the saddle-point objective is linear in both f and $\tilde{\mu}^\pi$. This is the result claimed in Sec. 2.2.

This is reminiscent of the classic adversarial approaches of the literature, with the inner problem consisting in solving a (non-stationary here) RL problem. There are important differences: the usual framework is γ-discounted infinite horizon, and as far as we know no practical approach is based on the IPM of the total variation. Rather, many of these adversarial approaches can be framed as minimizing an f-divergence between occupancy measures [15, 19], for example GAIL minimize an entropy-regularized Jensen-Shannon divergence [18]. We think that considering the total variation for our analysis and exposition is a practical and meaningful abstraction: it allows providing an analysis, and it could be an inspiration for deriving practical algorithms by applying a similar recipe.

A.2 MFG adversarial imitation when $L_P = 0$

When $L_P = 0$, the transition kernel of the MFG does not depend on the population. The reward does, but from the imitating agent viewpoint, there is an expert policy to imitate and an MDP without reward (the MFG without reward) to interact with. In other words, the minimization of the distance can be framed exactly as in the previous section:

$$\min_\pi \sum_{n=0}^{H-1} \|\mu_n^E - \mu_n^\pi\|_1 = \max_{f\in\mathcal{R}} \min_\pi (V_f(\pi^E) - V_f(\pi)).$$

So, no matter whether the expert policy is at a Nash equilibrium in an MFG or not, from the imitating agent this can be framed as a reduction to classical single agent imitation learning, with the exactly same guarantee.

A.3 MFG adversarial imitation when $L_P > 0$

First, consider the adversarial imitation approach studied in Sec. 4.2, that is we assume to control $\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1$. In essence, this means that we fix the population to be the expert one and ask a representative agent to imitate the expert policy. However, if we fix the population, the MFG without reward reduces to an MDP without reward, and we are again in the same case. We have

$$\begin{aligned}\|\tilde{\mu}^{(\pi^E)\pi^E} - \tilde{\mu}^{(\pi^E)\pi}\|_1 &= \max_{f\in\mathcal{R}} (\mathbb{E}_{s,a\sim\tilde{\mu}_n^{(\pi^E)\pi^E}}[f_n(s,a)] - \mathbb{E}_{s,a\sim\tilde{\mu}_n^{(\pi^E)\pi}}[f_n(s,a)]) \\ &= \max_{f\in\mathcal{R}} \frac{1}{H}(\sum_{s,a} \mu_n^{(\pi^E)\pi^E}(s,a) f_n(s,a) - \sum_{s,a} \mu_n^{(\pi^E)\pi}(s,a) f_n(s,a)) \\ &= \max_{f\in\mathcal{R}} \frac{1}{H}(V_f(\pi^E, \rho^{(\pi^E)}) - V_f(\pi, \rho^{(\pi^E)})).\end{aligned}$$

From this, as before we can deduce that

$$\min_\pi \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi}\|_1 = \max_{f\in\mathcal{R}} \min_\pi (V_f(\pi^E, \rho^{(\pi^E)}) - V_f(\pi, \rho^{(\pi^E)})).$$

The population being fixed to the expert one in both value functions, this is again a reduction to classic imitation, and the dual variable is a non-stationary reward that does not need to depend on the population. In other words, the inner problem is again a (non-stationary) RL problem. However, as discussed in Sec. 4.2, this does not come with encouraging theoretical guarantees, due to the exponential dependency on the horizon.

Eventually, let us consider the case of Sec. 4.3, that is, we assume to control $\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^A}\|_1$. Here, we no longer have a reduction to classic adversarial IL, because the two occupancy measures depends on different populations, but we can still obtain an adversarial formulation using the same IPM viewpoint. We have

$$\begin{aligned}\|\tilde{\mu}^{(\pi^E)\pi^E} - \tilde{\mu}^{(\pi)\pi}\|_1 &= \max_{f\in\mathcal{R}}(\mathbb{E}_{s,a\sim\tilde{\mu}_n^{(\pi^E)\pi^E}}[f_n(s,a)] - \mathbb{E}_{s,a\sim\tilde{\mu}_n^{(\pi)\pi}}[f_n(s,a)]) \\ &= \max_{f\in\mathcal{R}} \frac{1}{H}(\sum_{s,a}\mu_n^{(\pi^E)\pi^E}(s,a)f_n(s,a) - \sum_{s,a}\mu_n^{(\pi)\pi}(s,a)f_n(s,a)) \\ &= \max_{f\in\mathcal{R}} \frac{1}{H}(V_f(\pi^E,\rho^{(\pi^E)}) - V_f(\pi,\rho^{(\pi)})).\end{aligned}$$

From this, as before we can deduce that

$$\min_\pi \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi)\pi}\|_1 = \min_\pi \max_{f\in\mathcal{R}}(V_f(\pi^E,\rho^{(\pi^E)}) - V_f(\pi,\rho^{(\pi)})).$$

Notice that here, it is not obvious to know if we can switch the min and the max. Indeed, for this to hold, we need the set of policy-induced occupancy measures to be a convex set (in addition to the linearity of the value in both the reward and the occupancy measure). Whenever the dynamics does not depend on the population, this is true, this set is even a polytope. When the dynamics depends on the population, it is less clear, and ensuring the convexity of the underlying set may require additional assumptions on the transition kernel. We leave this interesting question for future work. For now, we assume that we can switch the min and the max, even if heuristically.

In this case, the underlying control problem is no longer an RL problem, but an MFC problem, as it implies solving for $\max_\pi V_f(\pi,\rho^{(\pi)})$, with again the reward being non-stationary, but still population-independent. This suggests that for obtaining such an adversarial IL approach for MFGs, one could start from an existing adversarial approach for the classic setting (for example, GAIL), and replace the underlying RL optimization problem by an MFC optimization problem. We leave the development of a more practical algorithm for future work, and it would probably call for more than just plugging an MFC algorithm in GAIL, but this suggests that making progress in IL for MFGs may have to build upon MFC.

B Proofs of stated theoretical results

In this section we report the proof of the theorems written in the paper. The main idea of the proof is to decompose the exploitability error.

B.1 Decomposition of the exploitability

In this subsection we provide the decomposition of the Nash imitation gap, that is the exploitability. For now, we do not make any assumption on how the policy π^A is obtained; it can be any policy. Our goal is to decompose:

$$\mathcal{E}(\pi^A) = \max_{\pi'} V(\pi',\rho^A) - V(\pi^A,\rho^A),$$

where we recall that we write ρ^A as a shorthand for $\rho^{(\pi^A)\pi^A}$, and similarly for other quantities (see Sec. 2.2). Then, proving our main results will amount to bound each term of the decomposition, depending on the setting ($L_P = 0$ or $L_P > 0$) and the kind of considered error.

Instead of bounding the exploitability, we bound the value difference for any policy π' (the bound on the exploitability is a simple corollary by maximizing over π').

Lemma 1. *Under Asm. 1 and 2, for any policies π^A and π', we have*

$$\begin{aligned}|V(\pi',\rho^A) - V(\pi^A,\rho^A)| &\leq 2L_r \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \\ &+ r_{max}\left(\sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 + \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1 + \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1\right).\end{aligned}$$

Proof. We start by decomposing the value difference:

$$V(\pi', \rho^A) - V(\pi^A, \rho^A) = \underbrace{V(\pi', \rho^A) - V(\pi^E, \rho^E)}_{A} + \underbrace{V(\pi^E, \rho^E) - V(\pi^A, \rho^A)}_{B}.$$

The idea is to study the distance between the value function with the quantity of interest, that is the Nash equilibrium policy π^E.

Term A. We decompose the term A again, summing and subtracting $V(\pi', \rho^E)$:

$$\begin{aligned} A &= V(\pi', \rho^A) - V(\pi^E, \rho^E) \\ &= \underbrace{V(\pi', \rho^A) - V(\pi', \rho^E)}_{A_1} + \underbrace{V(\pi', \rho^E) - V(\pi^E, \rho^E)}_{A_2}. \end{aligned}$$

The term A_2 can be interpreted as the gain we have using a different policy fixing the Nash equilibrium distribution ρ^E. Since (π^E, ρ^E) is a Nash equilibrium by assumption, then $\mathcal{E}(\pi^E) = 0$ and so $A_2 \leq 0$. Then, we need to study only the term A_1. Using Lemma 2 (see Appx. B.4) we have:

$$|V(\pi', \rho^A) - V(\pi', \rho^E)| \leq L_r \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 + r_{\max} \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1$$

Therefore, we have a bound for the term A:

$$|A| \leq L_r \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 + r_{\max} \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1.$$

Term B. We decompose the term B in the following way:

$$\begin{aligned} B &= V(\pi^E, \rho^E) - V(\pi^A, \rho^A) \\ &\leq \underbrace{V(\pi^E, \rho^E) - V(\pi^A, \rho^E)}_{B_1} + \underbrace{V(\pi^A, \rho^E) - V(\pi^A, \rho^A)}_{B_2}. \end{aligned}$$

We start by bounding the term B_1 (the positiveness of B_1 comes from π_E being a Nash equilibrium):

$$\begin{aligned} 0 \leq B_1 &= V(\pi^E, \rho^E) - V(\pi^A, \rho^E) \\ &= \sum_{n=0}^{H-1} \sum_{s,a} \left(\mu_n^{(\pi^E)\pi^E}(s,a) r(s,a,\rho^E) - \mu_n^{(\pi^A)\pi^E}(s,a) r(s,a,\rho^E) \right) \\ &= \sum_{n=0}^{H-1} \sum_{s,a} \left(\left(\mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^A)\pi^E}(s,a) \right) r(s,a,\rho^E) \right) \\ &\leq r_{\max} \sum_{n=0}^{H-1} \sum_{s,a} \left| \mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^A)\pi^E}(s,a) \right| \text{ (using Asm. 2)} \\ &= r_{\max} \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^E}\|_1. \end{aligned}$$

For the term B_2 we can use again Lemma 2:

$$|V(\pi^A, \rho^E) - V(\pi^A, \rho^A)| \leq L_r \sum_{n=0}^{H-1} \|\rho_n^E - \rho_n^A\|_1 + r_{\max} \sum_{n=0}^{H-1} \|\rho^{(\pi^E)\pi^A} - \rho^{(\pi^A)\pi^A}\|_1.$$

Then, putting things together:

$$|B| \leq r_{\max} \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^E}\|_1 + L_r \sum_{n=0}^{H-1} \|\rho_n^E - \rho_n^A\|_1 + r_{\max} \sum_{n=0}^{H-1} \|\rho^{(\pi^E)\pi^A} - \rho^{(\pi^A)\pi^A}\|_1.$$

Final bound. Applying the triangle inequality on the absolute value of the initial decomposition and injecting the bounds of $|A|$ and $|B|$, we obtain the stated result:

$$|V(\pi', \rho^A) - V(\pi^A, \rho^A)| \leq 2L_r \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1$$
$$+ r_{\max} \left(\sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 + \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1 + \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \right).$$

□

Notice that whenever $L_P = 0$, two terms of the above bound cancel out, $\sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1 = 0$ and $\sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 = 0$, as the occupancy measure does not depend on the population. This will be useful in the next section.

B.2 Proofs for the case $L_P = 0$

In this section we analyze the case in which the transition model is independent from the population distribution. As explained above, with the terms canceling out, the value difference is bounded as:

$$|V(\pi', \rho^A) - V(\pi^A, \rho^A)| \leq 2L_r \underbrace{\sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1}_{T_1} + r_{\max} \underbrace{\left(\sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \right)}_{T_2}.$$

We analyze now the two errors considered: the one from Behavioral Cloning and the adversarial one.

Behavioral Cloning. Recall the definition of the term ϵ_n^{BC} and the stated result.

$$\epsilon_n^{BC} := \mathbb{E}_{s \sim \rho_n^E} \left[\|\pi_n^A(s) - \pi_n^E(s)\|_1 \right].$$

Theorem 1. *Let $\epsilon^{BC} = \max_{n \in \{0, \ldots, H-1\}} \epsilon_n^{BC}$. If $L_P = 0$, the Nash imitation gap satisfies*

$$\mathcal{E}(\pi^A) \leq H^2 (r_{max} + 2L_r) \epsilon^{BC}.$$

Proof. Both terms T_1 and T_2 can be bounded using Lemma 3 in Appx. B.4 which connects the ℓ_1 distance between the two policies with the state-action distribution sequence, giving the stated result as a corollary. For the term T_1, to apply the lemma, it may be worth emphasizing that when $L_P = 0$, we have that $\|\rho_n^A - \rho_n^E\|_1 = \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^E}\|_1 = \|\rho_n^{(\pi^E)\pi^A} - \rho_n^{(\pi^E)\pi^E}\|_1$. □

Adversarial learning. Recall the definition of the term ϵ_n^{ADV} and the stated result.

$$\epsilon_n^{\text{ADV}} := \|\mu_n^{\pi^A} - \mu_n^{\pi^E}\|_1, \qquad n \in \{0, \ldots, H-1\}.$$

Theorem 2. *Let $\epsilon^{ADV} = \max_{n \in \{0, \ldots, H-1\}} \epsilon_n^{ADV}$. If $L_P = 0$, the Nash imitation gap satisfies*

$$\mathcal{E}(\pi^A) \leq (2L_r + r_{\max}) H \epsilon^{ADV}.$$

Proof. We start by bounding the term T_1.

$$\begin{aligned}
\|\rho_n^A - \rho_n^E\|_1 &= \sum_s |\rho_n^A(s) - \rho_n^E(s)| \\
&= \sum_s |\sum_a \mu_n^A(s,a) - \mu_n^E(s,a)| \\
&\leq \sum_{s,a} |\mu_n^A(s,a) - \mu_n^E(s,a)| = \|\mu_n^E - \mu_n^A\|_1 \leq \epsilon^{\text{ADV}},
\end{aligned}$$

and thus:

$$\sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \leq H\epsilon^{\text{ADV}}.$$

The second term, T_2, is bounded by $H\epsilon^{\text{ADV}}$ by definition of ϵ^{ADV}. Then putting things together we recover the stated result. □

B.3 Proofs for the case $L_P > 0$

We report in this section the results for the more general case in which the transition dynamics depends on the population. We recall the bound on the value difference given by Lemma 1:

$$|V(\pi', \rho^E) - V(\pi^E, \rho^E)| \leq 2L_r \underbrace{\sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1}_{T_1} + r_{\max}\Bigg(\underbrace{\sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1}_{T_2}$$
$$+ \underbrace{\sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1}_{T_3} + \underbrace{\sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1}_{T_4} \Bigg).$$

Behavioral cloning. Recall the definition of the term ϵ_n^{BC} and the stated result.

$$\epsilon_n^{\text{BC}} := \mathbb{E}_{s\sim\rho_n^E}\left[\|\pi_n^A(s) - \pi_n^E(s)\|_1\right].$$

Theorem 3. *Let $\epsilon^{BC} = \max_{n\in\{0,\dots,H-1\}} \epsilon_n^{BC}$. If $L_P > 0$, the Nash imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq \left(H^2 r_{max} + 2(1+L_P)^H \frac{L_r + r_{\max}}{L_P^2} \right) \epsilon^{BC}.$$

Proof. We start by bounding the term T_1. Now, the two sequences of involved occupancy measures differ by their underlying policy, as for the case $L_P = 0$, but also by their underlying dynamics, driven by different populations. The bound of the term T_1 is given by Lemma 4 in Appx. B.4:

$$T_1 = \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{\text{BC}}.$$

Next, we consider the term T_2. The two sequences of involved occupancy measures differ by their underlying policies, but they share the same dynamics, driven by the expert population. Therefore, as for the case $L_P = 0$, we can apply Lemma 3 (see Appx. B.4), and obtain

$$T_2 = \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \leq H^2\epsilon^{\text{BC}}.$$

Eventually, we consider the terms T_3 and T_4. They have in common that the two sequences of involved occupancy measures differ by their underlying dynamics (driven by different populations), but have the same underlying policy. Both terms can be bounded as a direct corollary of Lemma 5 in Appx. B.4, by instantiating this common policy. The resulting bounds are:

$$T_3 = \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{\text{BC}},$$
$$T_4 = \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{\text{BC}}.$$

Putting everything together we obtain the stated bound (noticing that the bound does not depend on the policy π', so a bound on the value difference readily gives a bound on the exploitability of π^A, that is the Nash imitation gap). □

Vanilla-ADV. Recall the definition of the term $\epsilon_n^{\text{vanilla-ADV}}$ and the stated result.

$$\epsilon_n^{\text{vanilla-ADV}} := \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \qquad \forall n \in [0, \dots, H-1].$$

Theorem 4. *Let $\epsilon^{\textit{vanilla-ADV}} = \max_{n=[0,\dots,H-1]} \epsilon_n^{\textit{vanilla-ADV}}$. If $L_P > 0$ the NE imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq r_{\max} H \epsilon^{\textit{vanilla-ADV}} + 2(1+L_P)^H \frac{r_{\textit{max}} + L_r}{L_P} \epsilon^{\textit{vanilla-ADV}}.$$

Proof. We start by bounding the term T_1. We have that:

$$\begin{aligned}
\|\rho_n^A - \rho_n^E\|_1 &= \sum_s |\sum_a \mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^A)\pi^A}(s,a)| \\
&\leq \underbrace{\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1}_{\leq \epsilon^{\text{vanilla-ADV}} \text{ by def.}} + \underbrace{\|\mu_n^{(\pi^E)\pi^A} - \mu_n^{(\pi^A)\pi^A}\|_1}_{= \|\rho_n^{(\pi^E)\pi^A} - \rho_n^{(\pi^A)\pi^A}\|_1 \text{ (same policy)}} \\
&\leq \epsilon^{\text{vanilla-ADV}} + \|\rho_n^{(\pi^E)\pi^A} - \rho_n^{(\pi^A)\pi^A}\|. \qquad (4)
\end{aligned}$$

From Eq. (8), an intermediate result of the proof of Lemma 5, we have the following inequality, for any policy π:

$$\|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 \leq L_P \|\rho_n^A - \rho_n^E\| + \|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1.$$

Instantiating this inequality with $\pi = \pi^A$ and injecting Eq. (4), we obtain

$$\begin{aligned}
\|\rho_{n+1}^{(\pi^A)\pi^A} - \rho_{n+1}^{(\pi^E)\pi^A}\|_1 &\leq L_P \|\rho_n^A - \rho_n^E\| + \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
&\leq L_P(\epsilon^{\text{vanilla-ADV}} + \|\rho_n^{(\pi^E)\pi^A} - \rho_n^{(\pi^A)\pi^A}\|) + \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
&= L_P \epsilon^{\text{vanilla-ADV}} + (1+L_P)\|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
&\leq L_P \epsilon^{\text{vanilla-ADV}} \sum_{k=0}^{n} (1+L_P)^k \\
&= L_P \epsilon^{\text{vanilla-ADV}} \frac{(1+L_P)^{n+1} - 1}{L_P} \\
&= \epsilon^{\text{vanilla-ADV}}((1+L_P)^{n+1} - 1).
\end{aligned}$$

Next, we can inject back this last inequality in Eq. (4) to get a bound on $\|\rho_n^A - \rho_n^E\|_1$, and then sum to obtain the bound on T_1.

$$\|\rho_n^A - \rho_n^E\|_1 \leq \epsilon^{\text{vanilla-ADV}}(1+L_P)^n \qquad (5)$$

$$\text{and thus } T_1 = \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \leq \frac{(1+L_P)^H - 1}{L_P} \epsilon^{\text{vanilla-ADV}} \leq \frac{(1+L_P)^H}{L_P} \epsilon^{\text{vanilla-ADV}}.$$

The term T_2 is directly bounded by $H\epsilon^{\text{vanilla-ADV}}$, by definition of $\epsilon^{\text{vanilla-ADV}}$.

Eventually, we need to bound the terms T_3 and T_4, implying occupancy measures for different populations (thus dynamics) but the same policy. We start again from Eq. (8) from the proof of Lemma 5, for an arbitrary policy π, and inject the bound we just obtained on $\|\rho_n^A - \rho_n^E\|_1$ (see Eq. (5)):

$$\begin{aligned}
\|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 &\leq L_P \|\rho_n^A - \rho_n^E\| + \|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1 \\
&\leq L_P \epsilon^{\text{vanilla-ADV}}(1+L_P)^n + \|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1 \\
&\leq \epsilon^{\text{vanilla-ADV}} L_P \sum_{k=1}^{n} (1+L_P)^k \\
&\leq \epsilon^{\text{vanilla-ADV}}(1+L_P)^{n+1}.
\end{aligned}$$

Summing over time we obtain (as the bound does not depend on the common policy π)

$$T_3, T_4 \leq \frac{(1+L_P)^H}{L_P} \epsilon^{\text{vanilla-ADV}}.$$

Putting everything together we recover the final bound. □

MFC-ADV Recall the definition of the term $\epsilon_n^{\text{MFC-ADV}}$ and the stated result.

$$\epsilon_n^{\text{MFC-ADV}} = \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^A}\|_1 \qquad \forall n \in [0, \dots, H-1].$$

Theorem 5. *Let $\epsilon^{\textit{MFC-ADV}} = \max_{n\in\{0,\dots,H-1\}} \epsilon_n^{\textit{MFC-ADV}}$. If $L_P > 0$, the Nash imitation gap satisfies:*

$$\mathcal{E}(\pi^A) \leq (2L_r + r_{\max})H\epsilon^{\textit{MFC-ADV}} + 3L_P r_{\max} H^2 \epsilon^{\textit{MFC-ADV}}.$$

Proof. The term T_1 is easily bounded by $H\epsilon^{MFC-ADV}$. Indeed, we have

$$\begin{aligned}
\|\rho_n^A - \rho_n^E\|_1 &= \sum_s |\rho_n^A(s) - \rho_n^E(s)| \\
&= \sum_s |\sum_a (\mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^A)\pi^A}(s,a))| \\
&\leq \sum_{s,a} |\mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^A)\pi^A}(s,a)| = \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^A}\|_1 \\
&\leq \epsilon^{MFC-ADV},
\end{aligned}$$

then by summing over time steps:

$$T_1 = \sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \leq H\epsilon^{MFC-ADV}.$$

We now focus on the terms T_3 and T_4. Starting again from Eq. (8), the intermediate result of the proof of Lemma 5, for an arbitrary policy π, we have:

$$\|\rho_{n+1}^{(A)\pi} - \rho_{n+1}^{(E)\pi}\|_1 \leq L_P \|\rho_n^A - \rho_n^E\| + \|\rho_n^{(A)\pi} - \rho_n^{(E)\pi}\|_1.$$

Then, using the definition of $\epsilon^{\text{MFC-ADV}}$ and by induction

$$\|\rho_{n+1}^{(A)\pi} - \rho_{n+1}^{(E)\pi}\|_1 \leq L_P \epsilon^{\text{MFC-ADV}} + \|\rho_n^{(A)\pi} - \rho_n^{(E)\pi}\|_1 \leq (n+1)L_P\epsilon^{\text{MFC-ADV}}. \tag{6}$$

This being true for any policy π, we obtain the bounds on T_3 and T_4 by summing:

$$\begin{aligned}
T_3 &= \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi'} - \rho_n^{(\pi^E)\pi'}\|_1 \leq H^2 L_P \epsilon^{\text{MFC-ADV}}, \\
T_4 &= \sum_{n=0}^{H-1} \|\rho_n^{(\pi^A)\pi^A} - \rho_n^{(\pi^E)\pi^A}\|_1 \leq H^2 L_P \epsilon^{\text{MFC-ADV}}.
\end{aligned}$$

Eventually, we bound the remaining term T_2. We have that

$$\begin{aligned}
\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\| &= \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^A} + \mu_n^{(\pi^A)\pi^A} - \mu_n^{(\pi^E)\pi^A}\|_1 \\
&\leq \underbrace{\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^A)\pi^A}\|_1}_{\leq \epsilon^{\text{MFC-ADV}} \text{ by def.}} + \underbrace{\|\mu_n^{(\pi^A)\pi^A} - \mu_n^{(\pi^E)\pi^A}\|_1}_{\leq nL_P\epsilon^{\text{MFC-ADV}} \text{ by Eq. (6) and same policy}} \\
&\leq \epsilon^{\text{MFC-ADV}} + nL_P\epsilon^{\text{MFC-ADV}}.
\end{aligned}$$

Then, summing over time, we obtain

$$T_2 = \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \leq H\epsilon^{\text{MFC-ADV}} + H^2 L_P \epsilon^{\text{MFC-ADV}}.$$

Putting all the terms together we recover the stated bound. □

B.4 Auxiliary lemmas

In this section we report some auxiliary lemmas used in the proofs.

The first lemma bounds the value difference for a common policy but different populations.

Lemma 2. *For every three policies π^1, π^2, π^3 and the associated population distributions $\rho^{(\pi^1)}, \rho^{(\pi^2)}$, we have under Asm. 1 and writing here $r_{max} = \max_{s,a} |r(s,a,\rho^{(\pi^2)})|$:*

$$|V(\pi^3, \rho^{(\pi^1)}) - V(\pi^3, \rho^{(\pi^2)})| \leq L_r \sum_{n=0}^{H-1} \|\rho_n^{(\pi^1)} - \rho_n^{(\pi^2)}\|_1 + r_{\max} \sum_{n=0}^{H-1} \|\rho_n^{(\pi^1)\pi^3} - \rho_n^{(\pi^2)\pi^3}\|_1.$$

Remark 2. *Notice that π^1 and π^2 play symmetric roles, and that we'll only call this result with $\pi^2 = \pi^E$, hence Asm. 2.*

Proof. In the proof, to lighten notations, we write $\rho^1 = \rho^{(\pi^1)}$ and $\rho^2 = \rho^{(\pi^2)}$. We start by decomposing the value difference as follows, by starting from the definition of the value, adding and subtracting the term $\mu_n^{(\pi^1)\pi^3}(s,a) r(s,a,\rho_n^2)$, and using the triangle inequality:

$$\begin{aligned}
&|V(\pi^3, \rho^1) - V(\pi^3, \rho^2)| \\
&= \left| \sum_{n=0}^{H-1} \sum_{s,a} \left(\mu_n^{(\pi^1)\pi^3}(s,a) r(s,a,\rho_n^1) - \mu_n^{(\pi^2)\pi^3}(s,a) r(s,a,\rho_n^2) \right) \right| \\
&= \left| \sum_{n=0}^{H-1} \sum_{s,a} \left(\mu_n^{(\pi^1)\pi^3}(s,a)(r(s,a,\rho_n^1) - r(s,a,\rho_n^2)) + (\mu_n^{(\pi^1)\pi^3}(s,a) - \mu_n^{(\pi^2)\pi^3}(s,a)) r(s,a,\rho_n^2) \right) \right| \\
&\leq \sum_{n=0}^{H-1} \sum_{s,a} \left| \mu_n^{(\pi^1)\pi^3}(s,a)(r(s,a,\rho_n^1) - r(s,a,\rho_n^2)) \right| + \sum_{n=0}^{H-1} \sum_{s,a} \left| (\mu_n^{(\pi^1)\pi^3}(s,a) - \mu_n^{(\pi^2)\pi^3}(s,a)) r(s,a,\rho_n^2) \right|.
\end{aligned}$$

We have two terms in the previous bound, and we upper-bound each of them. For the first one:

$$\begin{aligned}
\sum_{n=0}^{H-1} \sum_{s,a} \left| \mu_n^{(\pi^1)\pi^3}(s,a)(r(s,a,\rho_n^1) - r(s,a,\rho_n^2)) \right| &= \sum_{n=0}^{H-1} \sum_{s,a} \mu_n^{(\pi^1)\pi^3}(s,a) |(r(s,a,\rho_n^1) - r(s,a,\rho_n^2))| \\
&\leq \sum_{n=0}^{H-1} \sum_{s,a} \mu_n^{(\pi^1)\pi^3}(s,a) L_r \|\rho_n^1 - \rho_n^2\|_1 \\
&= L_r \sum_{n=0}^{H-1} \|\rho_n^1 - \rho_n^2\|_1,
\end{aligned}$$

where the inequality is due to the Lipschitz assumption (Asm. 1).

For the second term to be bounded, we have:

$$\begin{aligned}
\sum_{n=0}^{H-1} \sum_{s,a} \left| (\mu_n^{(\pi^1)\pi^3}(s,a) - \mu_n^{(\pi^2)\pi^3}(s,a)) r(s,a,\rho_n^2) \right| &\leq r_{\max} \sum_{n=0}^{H-1} \sum_{s,a} |\mu_n^{(\pi^1)\pi^3}(s,a) - \mu_n^{(\pi^2)\pi^3}(s,a)| \\
&= r_{\max} \sum_{n=0}^{H-1} \sum_{s,a} \pi_n^3(a|s) |\rho_n^{(\pi^1)\pi^3}(s) - \rho_n^{(\pi^2)\pi^3}(s)| \\
&= r_{\max} \sum_{n=0}^{H-1} \|\rho_n^{(\pi^1)\pi^3} - \rho_n^{(\pi^2)\pi^3}\|_1.
\end{aligned}$$

The first line relies on the assumption that the reward is uniformly bounded the second line is by definition of the joint occupancy measure, and the last line due to the probabilities summing to 1.

Putting things together, we obtain the stated bound. □

The next lemma provides intermediate bounds for the BC error when the dynamics is solely driven by the expert population, which also applies when $L_P = 0$ (as the dependency of the dynamics to the population disappear).

Lemma 3. *Recall that* $\epsilon^{BC} = \max_{0\leq n\leq H-1} \mathbb{E}_{s\sim\rho^E}[\|\pi_n^E(\cdot|s) - \pi_n^A(\cdot|s)\|_1]$. *We have that:*

$$\sum_{n=0}^{H-1} \|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \leq H^2\epsilon^{BC} \qquad \textit{and} \qquad \sum_{n=0}^{H-1} \|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \quad \leq H^2\epsilon^{BC}.$$

Proof. We start by working on the sequence of state occupancy measures. We proceed by induction. When $n = 0$, the two distributions are identical. For $n \geq 0$, assume that $\|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \leq n\epsilon^{\text{BC}}$. Then,

$$\begin{aligned}
&\|\rho_{n+1}^{(\pi^E)\pi^E} - \rho_{n+1}^{(\pi^E)\pi^A}\|_1 \\
=&\sum_s |\rho_{n+1}^{(\pi^E)\pi^E}(s) - \rho_{n+1}^{(\pi^E)\pi^A}(s)| \\
=&\sum_s |\sum_{x,a} \rho_n^{(\pi^E)\pi^E}(x)\pi_n^E(a|x)P(s|x,a,\rho^E) - \rho_n^{(\pi^E)\pi^A}(x)\pi_n^A(a|x)P(s|x,a,\rho^E)| \text{ (by def.)} \\
\leq&\sum_{x,a} |\rho_n^{(\pi^E)\pi^E}(x)\pi_n^E(a|x) - \rho_n^{(\pi^E)\pi^A}(x)\pi_n^A(a|x)| \underbrace{\sum_s P(s|x,a,\rho^E)}_{=1} \\
=&\sum_{x,a} |\rho_n^{(\pi^E)\pi^E}(x)\pi_n^E(a|x) - \rho_n^{(\pi^E)\pi^E}(x)\pi_n^A(a|x) + \rho_n^{(\pi^E)\pi^E}(x)\pi_n^A(a|x) - \rho_n^{(\pi^E)\pi^A}(x)\pi_n^A(a|x)| \\
\leq&\sum_x \rho_n^{(\pi^E)\pi^E}(x) \sum_a |\pi_n^E(a|x) - \pi_n^A(a|x)| + \sum_x |\rho_n^{(\pi^E)\pi^E}(x) - \rho_n^{(\pi^E)\pi^A}(x)| \underbrace{\sum_a \pi_n^A(a|x)}_{=1} \\
=&\mathbb{E}_{x\sim\rho_n^E}\left[\|\pi_n^E(\cdot|x) - \pi_n^A(\cdot|x)\|_1\right] + \|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
\leq&\epsilon^{\text{BC}} + \|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
\leq&(n+1)\epsilon^{\text{BC}}.
\end{aligned}$$

Then, summing over time provides the stated result:

$$\sum_{n=0}^{H-1} \|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \leq \epsilon^{\text{BC}} \sum_{n=0}^{H-1} n \leq H^2\epsilon^{\text{BC}}.$$

Building upon the previous result, we now work on the sequence of state-action occupancy measures.

$$\begin{aligned}
&\|\mu_n^{(\pi^E)\pi^E} - \mu_n^{(\pi^E)\pi^A}\|_1 \\
&= \sum_{s,a} |\mu_n^{(\pi^E)\pi^E}(s,a) - \mu_n^{(\pi^E)\pi^A}(s,a)| \\
&= \sum_{s,a} |\rho_n^{(\pi^E)\pi^E}(s)\pi^E(a|s) - \rho_n^{(\pi^E)\pi^A}(s)\pi^A(a|s)| \\
&= \sum_{s,a} |\rho_n^{(\pi^E)\pi^E}(s)\pi^E(a|s) - \rho_n^{(\pi^E)\pi^E}(s)\pi^A(a|s) + \rho_n^{(\pi^E)\pi^E}(s)\pi^A(a|s) - \rho_n^{(\pi^E)\pi^A}(s)\pi^A(a|s)| \\
&\leq \mathbb{E}_{s\sim\rho_n^E}[\|\pi_n^E(\cdot|s) - \pi_n^A(\cdot|s)\|_1] + \|\rho_n^{(\pi^E)\pi^E} - \rho_n^{(\pi^E)\pi^A}\|_1 \\
&\leq \epsilon^{\text{BC}} + n\epsilon^{\text{BC}} = (n+1)\epsilon^{\text{BC}}.
\end{aligned}$$

From this, by summing over time, we obtain the same stated bound. □

The next technical lemma considers the propagation of errors when bounding a term involving a sequence of occupancy measures relying on both different policies and different populations (so different dynamics).

Lemma 4. *Recall that* $\epsilon^{BC} = \max_{0\leq n\leq H-1} \mathbb{E}_{s\sim\rho^E}[\|\pi_n^E(\cdot|s) - \pi_n^A(\cdot|s)\|_1]$ *and assume that* $L_P > 0$. *We have that:*

$$\sum_{n=0}^{H-1} \|\rho_n^A - \rho_n^E\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{BC}.$$

Proof. We will bound the term $\|\rho_{n+1}^A - \rho_{n+1}^E\|_1$. The idea is to make use of the definition of the occupancy measure to make appear both ρ_n^A and ρ_n^E, to add and subtract various terms (namely $\rho_n^E(x)\pi_n^A(a|x)P(s|x,a,\rho_n^A)$ and $\rho_n^E(x)\pi_n^E(a|x)P(s|x,a,\rho_n^A)$), to use the triangle inequality, and then to bound each of the resulting terms.

$$\begin{aligned}
\|\rho_{n+1}^A - \rho_{n+1}^E\|_1 &= \sum_s |\rho_{n+1}^A(s) - \rho_{n+1}^E(s)| \\
&= \sum_s |\sum_{x,a} \rho_n^A(x)\pi_n^A(a|x)P(s|x,a,\rho_n^A) - \rho_n^E(x)\pi_n^E(a|x)P(s|x,a,\rho_n^E)| \\
&\leq \sum_{s,a,x} |\rho_n^A(x)\pi_n^A(a|x)P(s|x,a,\rho_n^A) - \rho_n^E(x)\pi_n^A(a|x)P(s|x,a,\rho_n^A)| \\
&\quad + \sum_{s,a,x} |\rho_n^E(x)\pi_n^A(a|x)P(s|x,a,\rho_n^A) - \rho_n^E(x)\pi_n^E(a|x)P(s|x,a,\rho_n^A)| \\
&\quad + \sum_{s,a,x} |\rho_n^E(x)\pi_n^E(a|x)P(s|x,a,\rho_n^A) - \rho_n^E(x)\pi_n^E(a|x)P(s|x,a,\rho_n^E)| \\
&= \underbrace{\sum_x \sum_{s,a} \pi_n^A(a|x)P(s|x,a,\rho_n^A)|\rho_n^A(x) - \rho_n^E(x)|}_{=\|\rho_n^A - \rho_n^E\|_1} \\
&\quad + \underbrace{\sum_{x,a} \rho_n^E(x)|\pi_n^E(a|x) - \pi_n^A(a|x)|}_{=\mathbb{E}_{x\sim\rho^E}[\|\pi_n^E(\cdot|x) - \pi_n^A(\cdot|x)\|_1]\leq\epsilon^{\text{BC}}} \underbrace{\sum_s P(s|x,a,\rho_n^A)}_{=1} \\
&\quad + \sum_{x,a} \rho_n^E(x)\pi_n^E(a|x) \underbrace{\sum_s |P(s|x,a,\rho_n^A) - P(s|x,a,\rho_n^E)|}_{\leq L_P\|\rho_n^A - \rho_n^E\|_1 \text{ by Asm. 1}} \\
&\leq (1+L_P)\|\rho_n^A - \rho_n^E\|_1 + \epsilon^{\text{BC}}.
\end{aligned}$$

By direct induction, we obtain

$$\|\rho_{n+1}^A - \rho_{n+1}^E\|_1 \leq \epsilon^{\text{BC}} \sum_{k=0}^{n} (1+L_P)^k.$$

Notice that if $L_P = 0$, we retrieve the result in the proof of Lemma 3, that is $\|\rho_{n+1}^A - \rho_{n+1}^E\|_1 \leq (n+1)\epsilon^{\text{BC}}$. If $L_P > 0$, this simplifies to

$$\|\rho_{n+1}^A - \rho_{n+1}^E\|_1 \leq \epsilon^{\text{BC}} \frac{(1+L_P)^{n+1} - 1}{L_P}. \tag{7}$$

Summing over time, we obtain the stated result,

$$\sum_{n=0}^{H-1} \|\rho_{n+1}^A - \rho_{n+1}^E\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{\text{BC}}.$$

□

The last technical lemma we provide considers the case when the involved sequences of occupancy measures have the same (arbitrary) policy but different populations (thus different dynamics).

Lemma 5. *Recall that* $\epsilon^{BC} = \max_{0\leq n\leq H-1} \mathbb{E}_{s\sim\rho^E}[\|\pi_n^E(\cdot|s) - \pi_n^A(\cdot|s)\|_1]$ *and assume that* $L_P > 0$. *Let* π *be an arbitrary policy. We have:*

$$\sum_{n=0}^{H-1} \|\rho_n^{(A)\pi} - \rho_n^{(E)\pi}\|_1 \leq \frac{(1+L_P)^H}{L_P^2}\epsilon^{BC}.$$

Proof. The proof follows a similar idea as in the proof of Lemma 4. We make use of the definition of occupancy measure to make appear the measures at the previous time step, we add and subtract a term (namely $\rho_n^{(\pi^A)\pi}(x)P(s|x,a,\rho_n^E)$ here) and use the triangle inequality, to eventually bound each of the resulting terms.

$$\begin{aligned}
\|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 &= \sum_s |\rho_{n+1}^{(\pi^A)\pi}(s) - \rho_{n+1}^{(\pi^E)\pi}(s)| \\
&= \sum_s |\sum_{x,a} \rho_n^{(\pi^A)\pi}(x)\pi_n(a|x)P(s|x,a,\rho_n^A) - \rho_n^{(\pi^E)\pi}(x)\pi_n(a|x)P(s|x,a,\rho_n^E)| \\
&\leq \sum_{s,a,x} \pi_n(a|x)|\rho_n^{(\pi^A)\pi}(x)P(s|x,a,\rho_n^A) - \rho_n^{(\pi^E)\pi}(x)P(s|x,a,\rho_n^E)| \\
&\leq \sum_{s,a,x} \pi_n(a|x)|\rho_n^{(\pi^A)\pi}(x)P(s|x,a,\rho_n^A) - \rho_n^{(\pi^A)\pi}(x)P(s|x,a,\rho_n^E)| \\
&\quad + \sum_{s,a,x} \pi_n(a|x)|\rho_n^{(\pi^A)\pi}(x)P(s|x,a,\rho_n^E) - \rho_n^{(\pi^E)\pi}(x)P(s|x,a,\rho_n^E)| \\
&= \sum_{a,x} \rho_n^{(\pi^A)\pi}(x)\pi_n(a|x) \underbrace{\sum_s |P(s|x,a,\rho_n^A) - P(s|x,a,\rho_n^E)|}_{\leq L_P\|\rho_n^A - \rho_n^E\|_1 \text{ by Asm. 1}} \\
&\quad + \underbrace{\sum_x \sum_{s,a} \pi_n(a|x)P(s|x,a,\rho_n^E)|\rho_n^{(\pi^A)\pi}(x) - \rho_n^{(\pi^E)\pi}(x)|}_{=\|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1},
\end{aligned}$$

$$\text{so } \|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 \leq L_P\|\rho_n^A - \rho_n^E\|_1 + \|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1. \tag{8}$$

If $L_P = 0$, this readily implies that $\|\rho_n^{(A)\pi} - \rho_n^{(E)\pi}\|_1 \leq \|\rho_0^{(A)\pi} - \rho_0^{(E)\pi}\|_1 = 0$, which is obviously the correct result (if the dynamics does not depend on the population, then the occupancy measures are the same, the underlying policy being the same).

Now consider the case of interest, $L_P > 0$. From Eq. (7) of the proof of Lemma 4, we know that

$$\|\rho_n^A - \rho_n^E\|_1 \leq \epsilon^{\text{BC}}\frac{(1+L_P)^n}{L_P}.$$

Injecting this into Eq. (8), we have

$$\|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 \leq \epsilon^{\text{BC}}(1+L_P)^n + \|\rho_n^{(\pi^A)\pi} - \rho_n^{(\pi^E)\pi}\|_1.$$

By direct induction we obtain

$$\|\rho_{n+1}^{(\pi^A)\pi} - \rho_{n+1}^{(\pi^E)\pi}\|_1 \leq \epsilon^{\text{BC}}\sum_{k=0}^{n}(1+L_P)^k \leq \epsilon^{\text{BC}}\frac{(1+L_P)^{n+1}}{L_P},$$

and summing over time we obtain the stated result:

$$\sum_{n=0}^{H-1} \|\rho_n^{(A)\pi} - \rho_n^{(E)\pi}\|_1 \leq \epsilon^{\text{BC}}\frac{(1+L_P)^H}{L_P^2}.$$

□

C Numerical illustration

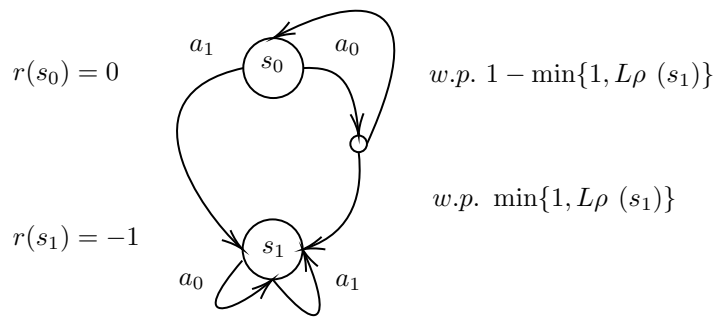

Figure 1: The "attractor" mean-field game.

In order to provide some empirical evidences of the insights given by our analysis (influence of the horizon and the dependency of the dynamics to the population on the various considered proxys to the Nash imitation gap), we introduce the "Attractor MFG" depicted in Fig. 1.

This is a 2-state and 2-action MFG with initial distribution satisfying $\rho_0(s_0) = 1$, with horizon H and with Lipschitz parameter L. The reward only depends on the state (not on the distribution nor the action) and satisfies for all $a \in \mathcal{A}$, $\rho \in \Delta_{\mathcal{S}}$,

$$r(s_0, a, \rho) = 0 \text{ and } r(s_1, a, \rho) = -1.$$

In the state s_1, any choice of actions leads deterministically to s_1, the transition kernel satisfies for all $a \in \mathcal{A}$, $\rho \in \Delta_{\mathcal{S}}$,

$$P(s_1|s_1, a, \rho) = 1.$$

In the state s_0, the action a_0 leads deterministically to s_1, while action a_0 leads stochastically to one of the two states: the higher the fraction of the population in s_1, the higher the chance to transit to s_1 after choosing a_0:

$$P(s_1|s_0, a_1, \rho) = 1 \text{ and } P(s_1|s_0, a_0, \rho) = \min\{1, L\rho(s_1)\}.$$

Therefore, the state s_1 is an attractor, hence the chosen name for the MFG.

Any policy choosing action a_0 in state s_0 for every timestep is a Nash equilibrium. Denoting by π^E such a policy, its value is $V(\pi^E, \rho^E) = 0$ (it is also a social equilibrium). The associated population obviously satisfies $\rho_n^E(s_0) = 1$ for all time steps $n = [0, \dots, H]$. We can also bound the Nash imitation gap, as any policy choosing action a_1 at timestep 0 in s_0 (and any action afterwards) will lead to the lowest possible value, that is for any policy π,

$$\mathcal{E}(\pi) \leq H - 1.$$

The Nash equilibrium being stationary, we consider the policy π_α being parameterized by the single scalar parameter $\alpha \in [0, 1]$, defined as (recall that the action selection on s_1 has no influence):

$$\pi^\alpha(a_1|s_0) = \alpha.$$

So, $\pi^{\alpha=0}$ is a Nash equilibrium, and $\pi^{\alpha=1}$ is a worst-case policy (of value $-(H-1)$). For such a policy, we directly get the BC error as

$$\epsilon_n^{\text{BC}}(\pi^\alpha) = \mathbb{E}_{s\sim\rho^E}[\|\pi_n^\alpha(\cdot|s) - \pi_n^E(\cdot|s)\|_1] = \|\pi_n^\alpha(\cdot|s_0) - \pi_n^E(\cdot|s_0)\|_1 = 2\alpha.$$

We can also easily compute the occupancy measures of interest by induction (it is sufficient to do so in the state s_1, as there are only two states):

$$\rho_0^{(E)\pi^\alpha}(s_1) = 0, \quad \rho_{n+1}^{(E)\pi^\alpha}(s_1) = \rho_n^{(E)\pi^\alpha}(s_1) + (1 - \rho_n^{(E)\pi^\alpha}(s_1))\alpha,$$
$$\rho_0^{(\pi^\alpha)}(s_1) = 0, \quad \rho_{n+1}^{(\pi^\alpha)}(s_1) = \rho_n^{(\pi^\alpha)}(s_1) + (1 - \rho_n^{(\pi^\alpha)}(s_1))(\alpha + (1-\alpha)\min\{1, L\rho_n^{(\pi^\alpha)}(s_1)\}).$$

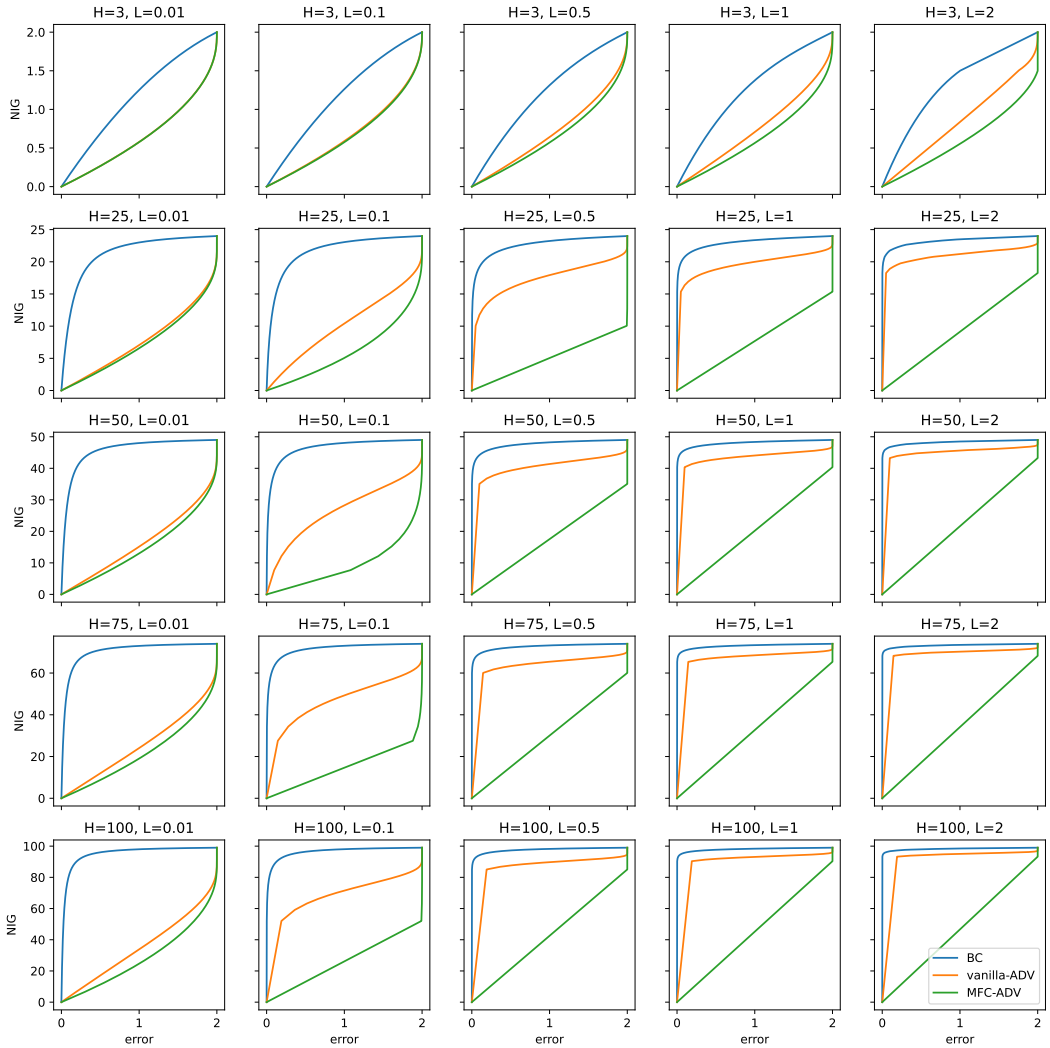

Figure 2: Nash imitation gap as a function of the considered maximum errors (ϵ^{BC}, $\epsilon^{\text{vanilla-ADV}}$ and $\epsilon^{\text{MFC-ADV}}$) in the attractor MFG, for various values of L (column-wise) and H (row-wise).

From this, we can easily get the related errors,

$$\epsilon_n^{\text{vanilla-ADV}}(\pi^\alpha) = \|\mu_n^{(E)} - \mu_n^{(E)\pi^\alpha}\|_1 = 2(\alpha + \rho_n^{(E)\pi^\alpha}(s_1)(1-\alpha)),$$
$$\epsilon_n^{\text{MFC-ADV}}(\pi^\alpha) = \|\mu_n^{(E)} - \mu_n^{(\pi^\alpha)}\|_1 = 2(\alpha + \rho_n^{(\pi^\alpha)}(s_1)(1-\alpha)).$$

With the above quantity, we also directly have the Nash imitation gap,

$$\mathcal{E}(\pi^\alpha) = \sum_{n=0}^{H-1} \rho_n^{(\pi^\alpha)}(s_1).$$

From this, we can also compute the maximum errors $\epsilon^{\text{BC}}(\pi_\alpha)$, $\epsilon^{\text{vanilla-ADV}}(\pi^\alpha)$ and $\epsilon^{\text{MFC-ADV}}(\pi^\alpha)$.

The numerical illustration we propose consists in computing these quantities for a grid of values of α, for various values of $L \in \{0.01, 0.1, 0.5, 1, 2\}$ and of $H \in \{3, 25, 50, 75, 100\}$, and showing the NIG as a function of respectively ϵ^{BC}, $\epsilon^{\text{vanilla-ADV}}$ and $\epsilon^{\text{MFC-ADV}}$, for the considered parameterized policies. The results are provided in Fig. 2.

We observe the following:

- NIG has a worse dependency to the BC errors than to the other ones. When either L or H increases, this dependency worsens, in the sense that smaller values of ϵ^{BC} are required for ensuring a given NIG.
- Vanilla-ADV and MFC-ADV behave similarly for small L and H (recall also that they are the same quantity for $L = 0$). However, whenever L or H increases, the dependency of the NIG to $\epsilon^{\text{vanilla-ADV}}$ worsens. Indeed, whenever L and/or H are large enough, one can observe than vanilla-ADV behaves more like BC than like MFC-ADV.
- MFC-ADV is also influenced by the values of L and H, but much less, and is always the best approach.

Overall, this supports the insights from our analysis. In particular MFC-ADV is better than vanilla-ADV, which itself is better than BC. When L and H are small enough, vanilla-ADV and MFC-ADV behave similarly, and when L and/or H are large enough, vanilla-ADV behaves more closely to BC. Overall, this suggests that a practical IL approach for MFGs should take into account the fact that the dynamics does depend on the population, and this may build upon MFC as suggested in the main paper.