A Variational Inequality Approach to Independent Learning in Static Mean-Field Games

BATUHAN YARDIM, Department of Computer Science, ETH Zürich, Zürich, Switzerland
SEMIH CAYCI, Department of Mathematics, RWTH Aachen University, Aachen, Germany
NIAO HE, Department of Computer Science, ETH Zürich, Zürich, Switzerland

Competitive games involving thousands or even millions of players are prevalent in real-world contexts, such as transportation, communications, and computer networks. However, learning in these large-scale multi-agent environments presents a grand challenge, often referred to as the "curse of many agents". In this article, we formalize and analyze the Static Mean-Field Game (SMFG) under both full and bandit feedback, offering a generic framework for modeling large population interactions while enabling independent learning.

We first establish close connections between SMFG and variational inequality (VI), showing that SMFG can be framed as a VI problem in the infinite agent limit. Building on the VI perspective, we propose independent learning and exploration algorithms that efficiently converge to approximate Nash equilibria, when dealing with a finite number of agents. Theoretically, we provide explicit finite sample complexity guarantees for independent learning across various feedback models in repeated play scenarios, assuming (strongly-)monotone payoffs. Numerically, we validate our results through both simulations and real-world applications in city traffic and network access management.

CCS Concepts: • **Computing methodologies** → **Multi-agent systems**; • **Theory of computation** → **Convergence and learning in games;**

Additional Key Words and Phrases: Mean-field games, independent learning, variational inequality, multi-agent systems, bandit feedback

ACM Reference Format:
Batuhan Yardim, Semih Cayci, and Niao He. 2025. A Variational Inequality Approach to Independent Learning in Static Mean-Field Games. *ACM/IMS J. Data Sci.* 2, 2, Article 7 (July 2025), 43 pages. https://doi.org/10.1145/3731553

1 Introduction

Multi-agent reinforcement learning (**MARL**) has been an active area of research, with a vast range of successful applications in games such as Chess, Shogi [63], Go [64], Stratego [51], as well as real-world applications for instance in robotics [41]. However, the applications of MARL to

This project was supported by Swiss National Science Foundation (SNSF) under the framework of NCCR Automation and SNSF Starting Grant. S. Cayci's research was funded under the Excellence Strategy of the Federal Government and the Länder.
Authors' Contact Information: Batuhan Yardim, Department of Computer Science, ETH Zürich, Zürich, Switzerland; e-mail: alibatuhan.yardim@inf.ethz.ch; Semih Cayci, Department of Mathematics, RWTH Aachen University, Aachen, Germany; e-mail: cayci@mathc.rwth-aachen.de; Niao He, Department of Computer Science, ETH Zürich, Zürich, Switzerland; e-mail: niao.he@inf.ethz.ch.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).
ACM 2831-3194/2025/07-ART7
https://doi.org/10.1145/3731553

games of a much larger scale involving thousands or millions of agents remain a theoretical and practical challenge [69].

Despite this limitation, competitive games with many players are ubiquitous and typically high-stakes. In many real-world games, extremely large-scale competitive multi-agency is the rule rather than the exception: for instance, when commuting every morning between cities using infrastructure shared with millions of other commuters, when periodically accessing an internet resource by querying servers used simultaneously by many users, or when sending information over common communication channels. Crucially, in such settings, learning can only feasibly occur in a *decentralized manner* due to the massive scale of the problem. Each agent can only utilize their local observations (often in the form of partial/bandit feedback) to maximize their selfish utilities. Another common feature in such games is *statelessness*. For instance, the capacity of an intercity highway or an internet server is only a function of its (and other highways'/servers') immediate load and not a function of time[1]. In our setting, we are interested in *independent learning (IL) using repeated play* in such games, that is, algorithms where each agent learns from their own (noisy) interactions without observing others or a centralized coordinator. IL, despite being theoretically challenging, is natural for such games as centralized control can be an unrealistic assumption for large populations.

1.1 Game Formulation

We formalize the **static mean-field game** (**SMFG**), the main mathematical object of interest in this work. The SMFG problem with repeated plays consists of the following.

- Finite set of players $\mathcal{N} = \{1, \dots, N\}$ with $|\mathcal{N}| = N \in \mathbb{N}_{>0}$,
- Set of finitely many actions $\mathcal{A}$, with $|\mathcal{A}| = K \in \mathbb{N}_{>0}$,
- A payoff function $\mathbf{F} : \Delta_{\mathcal{A}} \to [0,1]^K$, which maps the occupancy measure of the population over actions to a corresponding payoff in $[0,1]$ for each action.

Here $\Delta_{\mathcal{A}}$ denotes the probability simplex over a finite set $\mathcal{A}$. For $\mu \in \Delta_{\mathcal{A}}$, $a \in \mathcal{A}$, we denote the entry of $\mathbf{F}(\mu) \in \mathbb{R}^{\mathcal{A}}$ corresponding to action a as $\mathbf{F}(\mu, a) = \mathbf{F}(\mu)(a) \in \mathbb{R}$.

SMFG is played across multiple rounds where players are allowed to alter their strategies in between observations, where at each round $t \in \mathbb{N}_{\geq 0}$,

(1) Each player $i \in \mathcal{N}$ picks an action $a_t^i \in \mathcal{A}$,
(2) The empirical occupancy measure over actions is set to be $\widehat{\mu}_t := \frac{1}{N} \sum_{i=1}^N \mathbf{e}_{a_t^i}$
(3) Depending on the feedback model, each agent $i \in \mathcal{N}$ observes noisy rewards:
Full feedback: The (noisy) payoffs for *each* action $\mathbf{r}_t^i := \mathbf{F}(\widehat{\mu}_t) + \mathbf{n}_t^i \in \mathbb{R}^K$ or,
Bandit feedback: The (noisy) payoff for their chosen action $r_t^i := \mathbf{r}_t^i(a_t^i) \in \mathbb{R}$.
(4) Each agent receives the payoff r_t^i.

Here $\mathbf{e}_a \in \mathbb{R}^{\mathcal{A}}$ is the standard unit vector with coordinate a set to 1, and we assume for each action $a \in \mathcal{A}$ the noise $\mathbf{n}_t^i(a)$ are independent for all $i \in \mathcal{N}, t \geq 0$ and have zero mean and variance upper bounded by σ^2. Intuitively, SMFG models games where the payoff obtained from choosing an action depends on how the population is statistically distributed over actions, without distinguishing between particular players. Hence, each agent observes a noisy version of the payoff of their action (or the payoff of all actions with full feedback) under the current empirical occupancy over actions.

We motivate the relevance of the SMFG framework with three examples from real-world applications.

[1]Unless the timescale is years/decades, in which long-term degradation effects will become significant.

(1) **Network resource management.** Assume there are K resources (e.g., servers or computational nodes) available on a computer network shared by a large number of (say, $N \gg K$) users. These resources might have varying capabilities, hence the delay in accessing each will be a non-linear function of the overall load. At each time step, each user tries to access a resource and experiences a delay that increases with the number of users trying to access the same node. The Tor network application is a particularly good example of SMFGs: the network has 2–3 million users and functions by each user choosing an entry node to use in a decentralized fashion [68].

(2) **Repeated commuting with large populations.** Every morning, N commuters from city A to city B choose among K routes to drive to their target (typically $N \gg K$), observing only how long it takes them to commute. The distribution of choices among the population affects how much time each person spends traveling. The commute times in modern road infrastructure can be a very complex function of overall load [26] due to non-trivial feedback loops and adaptive systems such as load-dependent traffic lights. On the other hand, city traffic is a prime example of finite resources (with potentially complex payoff functions) shared by thousands or millions of inhabitants, especially during heavy commute hours [3]. The empirical study of macroscopic flow diagrams [3, 20] suggests that at least in aggregate, city traffic flow is explained by its momentary occupancy. This empirical observation motivates modeling periodic commutes as a stateless game.

(3) **Multi-player multi-armed bandits with soft collisions.** Multiplayer MAB have already been studied in the special case where collisions (i.e., multiple players choosing the same arm) result in zero returns. In many real-world applications, arms used by multiple players yield diminished (but non-zero) utilities when occupied by multiple players. For instance, in many radio communications applications (Bluetooth, Wi-Fi), common frequencies are automatically used via *time-sharing*, yielding delayed but successful communication when collisions occur.

1.2 Learning Objectives

We consider competitive games, that is, each agent aims to maximize their personal expected payoff without regard to social welfare. We allow agents to play randomized actions (mixed strategies), where each agent i randomly chooses their actions at time t with respect to their mixed strategy $\boldsymbol{\pi}_t^i \in \Delta_{\mathcal{A}}$. The primary solution concept for such a game is the **Nash equilibrium** (**NE**).

Definition 1.1 (Expected payoff, exploitability, Nash equilibrium). For an N-tuple of mixed strategies $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \in \Delta_{\mathcal{A}}^N$, we define the expected payoff V^i of an agent $i \in \mathcal{N}$ as

$$V^i(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) := \mathbb{E}\left[\mathbf{F}\left(\frac{1}{N}\sum_{j=1}^{N} \mathbf{e}_{a^j}\right)(a^i) \middle| a^j \sim \boldsymbol{\pi}^j, \forall j = 1, \ldots, N\right].$$

We define the *exploitability* of agent i for the tuple as $\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^j\}_{j=1}^N) := \max_{\boldsymbol{\pi}' \in \Delta_{\mathcal{A}}} V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) - V^i(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N)$. An N-tuple $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \in \Delta_{\mathcal{A}}^N$ is called a NE if for all i, $\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^j\}_{j=1}^N) = 0$, namely, $V^i(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) = \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-i})$, and it is called a δ-NE ($\delta > 0$) if for all i, $\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^j\}_{j=1}^N) \leq \delta$, namely, $V^i(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \geq \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-i}) - \delta$.

Intuitively, under a mixed strategy profile $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N)$ that is a NE, no agent has the incentive to deviate from their mixed strategy as in expectation each agent plays optimally with respect to the rest. A NE of the SMFG can be shown to always exist in this setting without any further assumptions [46], as the SMFG is also an N-player normal form game when $\mathcal{A}$ is finite.

Goal. With the SMFG problem definition and the solution concept introduced, we state our objective: in both feedback models, we would like to find *sample efficient* algorithms[2] that learn an *approximate NE* (in the sense of low exploitability) *independently* from repeated plays by N agents when N is *large*.

We will solve the SMFG in the regime when N is very large by leveraging its connection to the corresponding problem in the infinite-player limit, i.e., $N \to \infty$. For this purpose, we introduce the following MF-NE concept.

Definition 1.2 (MF-NE). A **mean-field Nash equilibrium (MF-NE)** $\boldsymbol{\pi}^* \in \Delta_{\mathcal{A}}$ associated with payoff operator $\mathbf{F}$ is a probability distribution over actions such that $\sum_a \boldsymbol{\pi}^*(a)\mathbf{F}(\boldsymbol{\pi}^*, a) = \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \sum_a \boldsymbol{\pi}(a)\mathbf{F}(\boldsymbol{\pi}^*, a)$. If it holds that $\sum_a \boldsymbol{\pi}^*(a)\mathbf{F}(\boldsymbol{\pi}^*, a) \geq \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \sum_a \boldsymbol{\pi}(a)\mathbf{F}(\boldsymbol{\pi}^*, a) - \delta$ for some $\delta > 0$, we call $\boldsymbol{\pi}^*$ a δ-MF-NE.

Intuitively, the mean-field limit simplifies the SMFG problem by assuming each agent follows the same policy and replacing the notion of N independent agents with a single agent playing against a continuum of infinitely many identical agents. Notably, finding an MF-NE is equivalent to solving the following **variational inequality (VI)** problem corresponding to the operator $\mathbf{F}$ and domain $\Delta_{\mathcal{A}}$:

$$\text{Find } \boldsymbol{\pi}^* \in \Delta_{\mathcal{A}} \text{ s.t. } \mathbf{F}(\boldsymbol{\pi}^*)^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}) \geq 0, \forall \boldsymbol{\pi} \in \Delta_{\mathcal{A}}. \tag{MF-VI}$$

The δ-MF-NE is related to the strong gap function of (MF-VI), as $\boldsymbol{\pi}^*$ is a δ-MF-NE if and only if $\min_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \mathbf{F}(\boldsymbol{\pi}^*)^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}) \geq -\delta$. While the MF-VI is not a concrete game and MF-NE is not strictly speaking a Nash equilibrium of any game, they will be crucial tools in constructing **independent learning** (IL) algorithms for SMFG.

1.3 Summary of Contributions

Based on the VI perspective, we develop algorithms that efficiently converge to approximate NE solutions for the SMFG problem, providing finite-sample, finite-agent, and IL guarantees under different feedback models. Our key technical contributions are as follows:

(1) **VI formulation and approximation.** We formulate the infinite-agent mean-field game limit as a VI. We make connections between MF-NE, VIs with regularization, and NE of the SMFG, providing explicit bounds on the bias introduced due to approximating the N-agent game with an infinite player game. In particular, we show that a solution of the (MF-VI) is a $\mathcal{O}(1/\sqrt{N})$-NE of the N-player game, thereby justifying the usefulness of the approximation (MF-VI) when N is large.

(2) **Independent learning with N agents and full feedback.** Leveraging techniques from optimization and VIs, we analyze independent learning in the N-agent setting and prove finite sample bounds with regularized learning. Our work highlights how VI and operator theory can provide insights into IL with full or partial feedback. In each feedback model, we propose a τ-Tikhonov regularization scheme that stabilizes the divergence between agent policies without communication, which introduces a $\mathcal{O}(\tau)$ bias. When agents have full (though noisy) feedback of reward payoffs, we show that after T rounds of play, our algorithm produces policies with expected exploitability upper bounded by $\mathcal{O}(\frac{\tau^{-2}}{\sqrt{T}} + \frac{\tau^{-1}}{\sqrt{N}} + \tau)$ with monotone payoffs. For λ-strongly monotone payoffs, we improve this to $\mathcal{O}(\frac{\tau^{-3/2}\lambda^{-1/2}}{\sqrt{T}} + \frac{\tau^{-1/2}\lambda^{-1/2}}{\sqrt{N}} + \tau)$.

(3) **Learning guarantees with bandit feedback.** We further analyze the learning algorithm under a bandit feedback model and prove finite sample guarantees. In this case, we propose

[2]We rigorously formalize the notion of an algorithm in the different feedback models in the appendix (Section B).

a probabilistic exploration scheme that enables agents to obtain low-variance estimates of the payoff operator $\mathbf{F}$ without centralized coordination. In this scenario, we show that after T rounds of play in expectation the exploitability in the *unregularized* SMFG is bounded by $\widetilde{O}(\frac{N^{3/4}}{\sqrt{T}} + \frac{1}{\sqrt[4]{N}})$ for monotone payoffs and by $\widetilde{O}(\frac{N^{3/4}\lambda^{-1/2}}{\sqrt{T}} + \frac{\lambda^{-1/2}}{\sqrt[4]{N}})$ for λ-strongly monotone payoffs.

(4) **Empirical validation.** We verify our theoretical results through synthetic and real-world experiments such as the beach bar problem, linear payoffs, city traffic management, and network access within the Tor network. These experiments demonstrate the effectiveness of the SMFG formulation and our IL algorithms for addressing games with large populations.

1.4 Related Work and Comparisons

This work is situated at the intersection of multiple areas of research, and we discuss each.

Mean-field games. Mean-field games (MFG), originally proposed independently by Lasry and Lions [35] and Huang et al. [27], have been an active research area in multi-agent RL literature. MFG is a useful theoretical tool for analyzing a specific class of MARL problems consisting of a large number of players with symmetric (but competitive) interests by formulating the infinite-agent limit. The MFG formalism has been used successfully to model large-scale problems such as auctions [54], electricity markets [21], epidemics [6, 28], and carbon markets [10].

In this work, we analyze a learning problem in a static (or stateless) finite-action MFG. Closer to our setting, discrete-time dynamic MFGs have been analyzed under various assumptions, reward models, and solution concepts, such as Lasry-Lions games [52, 53], stationary MFG [4, 71, 73, 76], linear quadratic MFG [18, 24], and MFGs on graphs [22, 72]. However, finite-sample learning guarantees for mean-field games only exist under specific assumptions. For Lasry-Lions games, convergence guarantees exist with an infinite agent oracle [53, 77], while IL guarantees with finite agents in this setting do not exist to the best of our knowledge. Recently, theoretical learning guarantees for centralized learning on monotone MFGs with graphon structure [77] and potential MFGs [19] have been shown. IL in stationary MFGs has been studied either under a subjective equilibrium solution concept [75] or with strong regularization bias and poor sample complexity [73].

Variational inequalities (VI). VIs and relevant algorithms have been an active area of research in classical and recent optimization literature; see e.g., Facchinei and Pang [17], Kotsalis et al. [33], Lin and Jordan [38], Nemirovski [47], Nesterov [48], just to list a few. VIs have been extensively analyzed in the case of (strongly) monotone operators. Classical algorithms in monotone VIs include the projected gradient method [17], the extragradient method [32], and proximal point algorithms [57]. Several works provide extensions to more general operator classes such as generalized monotone operators [33] or pseudo-monotone operators [31]. A recent survey for methods for solving VIs can be found in Beznosikov et al. [8]. In this work, we use VIs to formalize the infinite-player limit and use algorithmic tools from VI literature to analyze the more challenging IL setting. Our algorithms are adaptations of the projected gradient method to the decentralized learning under partial feedback. However, in the SMFG setting, it is not possible to have oracle access to the VI operator (or even unbiased samples of it), posing an algorithmic and theoretical design challenge. Strictly speaking, the VI given by (MF-VI) exists only as an abstract approximation in our work and not as an oracle model.

Decentralized VI algorithms. Another setting that is related but orthogonal to our problem is the distributed/decentralized VI problem studied for example by Kovalev et al. [34], Mukherjee and Chakraborty [45], Srivastava et al. [66]. In this setting, the VI operator is assumed to decompose into M components $\mathbf{F} := \sum_{m=1}^{M} \mathbf{F}_m$ each of which can be evaluated locally by M workers. Workers are permitted to occasionally transmit information along a communication graph: the communi-

cation complexity of algorithms as well as the computational complexity is of interest. Our setting fundamentally differs from the decentralized VI problem. Firstly, the VI operator in our case exists only implicitly as an approximation of the finite-player game dynamics: strictly speaking, the problem becomes a VI only at the infinite-player limit. Furthermore, the VI operator $\mathbf{F}$ in our setting does not admit components that can be independently evaluated by any player due to the game interactions: our stochastic oracle is much more restrictive. Finally, the independent players cannot communicate at all throughout the repeated plays, introducing the algorithmic challenge of fully decentralized learning.

Multi-player MAB (MMAB). Our model is related to the MMAB problem. The standard reward model for MMAB is collisions [5], where agents receive a reward of 0 if more than one pulls the same arm. Results have focused on the so-called no-communications setting, proving regret guarantees for the cooperative setting without a coordinator [7, 9, 58]. Our SMFG model can capture much more general reward models than binary collisions. Moreover, in our competitive setting, our primary metric is exploitability or proximity to NE rather than regret as typically employed in the collaborative setting of MMAB (similar to the NE for MMAB result established by Lugosi and Mehrabian [40]).

Other work. A setting that has a similar set of keywords is mean-field bandits [23, 70], as mentioned in the survey by Laurière et al. [36]. However, these models do not analyze a *competitive (Nash)* equilibrium, but rather a population steady-state reached by an infinite population under prescribed unknown agent behavior. In these works, optimality or exploitability is not a concern, hence a direct comparison with our work is not possible. On the other hand, our model is equivalent to decentralized learning on static MFGs [36] with finitely many agents and partial feedback. The problem of IL is fundamental in algorithmic game theory and has been separately investigated in zero-sum Markov games [12, 49, 62] and potential games [2, 14, 25]. Another related literature is population games and evolutionary dynamics [55, 61], where competitive populations are analyzed with differential equations. While solution concepts overlap, we are interested in IL with repeated play under partial feedback, not the continuous-time dynamic system.

1.5 Organization and Notations

The article is structured as follows. In Section 2, we introduce our assumptions on the SMFG operator and discuss their theoretical and practical relevance. In Section 3, we theoretically analyze the SMFG at the limit $N \to \infty$ and make connections to variational inequalities in optimization literature. In Sections 4 and 5, we formulate our algorithms and present the finite-time analysis of *independent learning* in full and bandit feedback cases. Finally, in Section 6, we experimentally verify our theory in numerical examples and two real-world use cases in city traffic and access to the Tor network.

Notation. For any $M \in \mathbb{N}_{>0}$, let $[M] := \{1, \dots, M\}$. $\Delta_{\mathcal{A}}$ denotes the probability simplex over a finite set $\mathcal{A}$, and $\Delta_{\mathcal{A},N} \subset \Delta_{\mathcal{A}}$ is the set of possible empirical measures with N samples, that is, $\Delta_{\mathcal{A},N} := \{\mathbf{u} = \{u_i\}_i \in \Delta_{\mathcal{A}} | N u_i \in \mathbb{N}_{\geq 0}\}$. For a set $\mathcal{A}$ and a map $\mathbf{F} : \mathcal{X} \to \mathbb{R}^{\mathcal{A}}$ defined on an arbitrary set $\mathcal{X}$, for $x \in \mathcal{X}$, $a \in \mathcal{A}$, we denote the entry of $\mathbf{F}(x) \in \mathbb{R}^{\mathcal{A}}$ corresponding to a as $\mathbf{F}(x, a) \in \mathbb{R}$. For a vector $\mathbf{u} \in \mathbb{R}^{\mathcal{A}}$, $\mathbf{u}(a) \in \mathbb{R}$ denotes its entry corresponding to $a \in \mathcal{A}$. $\mathbf{1}_K \in \mathbb{R}^K$ denotes a vector with all entries 1. For N-tuple $\mathbf{v} \in \mathcal{X}^N$ and $u, v \in \mathcal{X}$, $(v, \mathbf{v}^{-i})$ is the N-tuple with the ith entry of $\mathbf{v}$ replaced by v and $(u, v, \mathbf{v}^{-i,-j})$ is the N-tuple with the ith and jth entries of $\mathbf{v}$ replaced by u, v respectively. $\mathbf{e}_a \in \mathbb{R}^{\mathcal{A}}$ is the standard unit vector with coordinate a set to 1. $\mathbb{M}^{D_1, D_2}$ denotes the set of $D_1 \times D_2$ matrices, $\mathbb{S}^D_{++}$ denotes the set of positive definite $D \times D$ symmetric matrices. For a compact convex set $K \subset \mathbb{R}^H$, $\Pi_K : \mathbb{R}^H \to K$ denotes the projection operator. For an event (i.e. measurable set) E in a generic probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote its complement $\Omega \setminus E$ as $\overline{E}$.

2 Assumptions on Payoffs and Examples

In general, finding a NE in normal form games is known to be challenging and possibly computationally intractable [11, 50]. This difficulty persists even in the mean-field regime [74]. Therefore, we will analyze SMFG under certain structural assumptions on $\mathbf{F}$ that are closely aligned with many real-world applications.

2.1 Assumption on Lipschitz Continuity

Our first assumption is the Lipschitz continuity of the payoff. In practice, action payoffs should not change drastically with small variations in population behavior: Assumption 1 formalizes this intuition.

Assumption 1 (Lipschitz continuous payoffs). *The payoff map $\mathbf{F} : \Delta_{\mathcal{A}} \to [0,1]^K$ is called Lipschitz continuous with parameter $L > 0$ if for any $\mu, \mu' \in \Delta_{\mathcal{A}}$, $\|\mathbf{F}(\mu) - \mathbf{F}(\mu')\|_2 \leq L\|\mu - \mu'\|_2$.*

As a direct consequence of the Lipschitz continuous payoffs, we have the following technical lemmas, which will be used in the subsequent analysis.

Lemma 2.1. *For any policy profile $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \in \Delta_{\mathcal{A}}^N$, it holds that*

$$|V^i(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) - \boldsymbol{\pi}^{i,\top}\, \mathbb{E}\left[\mathbf{F}(\widehat{\mu})\Big|_{\widehat{\mu} := \frac{1}{N}\sum_{j=1}^N \mathbf{e}_{a^j}}^{a^j \sim \boldsymbol{\pi}^j, \forall j}\right]| \leq \frac{L\sqrt{2}}{N}.$$

In addition, we can also obtain the Lipschitz continuity of V^i and $\mathcal{E}^i_{\text{exp}}$, as stated below.

Lemma 2.2 ($V^i, \mathcal{E}^i_{\text{EXP}}$ are Lipschitz). *For any $i \in \mathcal{N}$, the value functions $V^i : \Delta_{\mathcal{A}}^N \to \mathbb{R}$ and exploitability functions $\mathcal{E}^i_{exp} : \Delta_{\mathcal{A}}^N \to \mathbb{R}$ are Lipschitz continuous, that is, for any $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \in \Delta_{\mathcal{A}}^N, \boldsymbol{\pi}, \bar{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}$,*

$$\begin{aligned}|V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-j}) - V^i(\bar{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-j})| &\leq L_{j,i}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2,\\ |\mathcal{E}^i_{exp}(\boldsymbol{\pi}, \boldsymbol{\pi}^{-j}) - \mathcal{E}^i_{exp}(\bar{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-j})| &\leq \bar{L}_{j,i}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2,\end{aligned}$$

with the Lipschitz moduli given by

$$L_{i,j} = \begin{cases}\sqrt{K}, \text{if } i = j,\\ \frac{2L\sqrt{2K}}{N}, \text{if } i \neq j,\end{cases} \qquad \bar{L}_{i,j} = \begin{cases}\sqrt{K}, \text{if } i = j,\\ \frac{4L\sqrt{2K}}{N}, \text{if } i \neq j.\end{cases}$$

We defer the proofs to Sections D.1 and D.2 in the appendix.

2.2 Assumption on Monotonicity

The next fundamental assumption, monotonicity, is standard in VI literature [17] and is similar in form to Lasry-Lions conditions in mean-field games literature (initially studied by [35], later analyzed with reinforcement learning in MFGs by [52, 53] and many others).

Assumption 2 (Monotone/Strongly monotone payoff). *The vector map $\mathbf{F} : \Delta_{\mathcal{A}} \to [0,1]^K$ is called monotone if for some $\lambda \geq 0$, for all $\forall \mu_1, \mu_2 \in \Delta_{\mathcal{A}}$, it holds that*

$$\begin{aligned}(\mathbf{F}(\mu_1) - \mathbf{F}(\mu_2))^\top (\mu_1 - \mu_2) &= \sum_{a \in \mathcal{A}} (\mu_1(a) - \mu_2(a))(\mathbf{F}(\mu_1, a) - \mathbf{F}(\mu_2, a))\\ &\leq -\lambda\|\mu_1 - \mu_2\|_2^2.\end{aligned}$$

Furthermore, if the above holds for some $\lambda > 0$, $\mathbf{F}$ is called λ-strongly monotone.

Monotone payoffs, as in the case of Lasry-Lions conditions, can be intuitively thought of as modeling problems where the payoff of an action on average decreases as the occupancy increases. In comparison to [35, 53], since the game is stateless in our case, the monotonicity condition degenerates to "decreasing payoffs with increasing *action occupancy*".

While Assumption 2 is abstract, one large class of concrete monotone payoff functions are given by games where actions have payoffs non-increasing with their occupancy and there is no interaction between actions, demonstrated below by Example 2.3.

Example 2.3 (Non-increasing payoffs). Let $F_a : [0,1] \to [0,1]$ for $a \in \mathcal{A}$ be Lipschitz continuous and non-increasing functions, define $\mathbf{F}(\mu) := \sum_a F_a(\mu(a))\mathbf{e}_a$. Then $\mathbf{F}$ is monotone and Lipschitz. If F_a is also *strictly* decreasing and there exists $\lambda > 0$ such that $|F_a(x) - F_a(x')| \geq \lambda |x - x'|$ for all $x, x' \in [0,1], a \in \mathcal{A}$, then $\mathbf{F}$ is λ-strongly monotone.

Applications where such an F_a exists are common and intuitive: non-increasing F_a models congestion effects on common resources. Even theoretically, the existence of a non-increasing F_a in wireless communications can be motivated by well-known results on the capacity of multiple access channels in information theory [16, 65]. Likewise, the conceptual models of traffic flow (e.g., see the macroscopic flow diagram [39]) imply traffic *speed* is monotonically decreasing with occupancy in aggregate. As a special case of Example 2.3, the classical multi-player multi-armed bandits setting with collisions yields a monotone Lipschitz continuous payoff. We state this explicitly in the following example.

Example 2.4 (Multi-player MAB with Collisions). For the N-player game, for each action $a \in \mathcal{A}$, take the functions $F^a_{col} : [0,1] \to [0,1]$ such that

$$F^a_{col}(x) = \begin{cases} \alpha^a, \text{if } x \leq \frac{1}{N}, \\ \alpha^a N(\frac{2}{N} - x), \text{if } \frac{1}{N} \leq x \leq \frac{2}{N}, \\ 0, \text{if } x \geq \frac{2}{N}. \end{cases}$$

where $\alpha^a \in [0,1]$ is the expected payoff of the action a when no collision occurs. Take the payoff map $\mathbf{F}(\mu) := \sum_{a \in \mathcal{A}} F^a_{col}(\mu_a)\mathbf{e}_a$. The payoff map $\mathbf{F}$ is Lipschitz continuous and monotone and corresponds to the classical multi-player MAB with collisions.

Due to this fact, one interpretation of our SMFG setting is that we allow soft collisions for action payoffs, as actions may yield non-zero but diminished payoffs when chosen by multiple players. In fact, a model with hard collisions would not be feasible (or realistic) in applications where $N \gg K$ as any policy supported by all actions would lead to collisions with probability approaching 1. On the other hand, a general non-increasing function F^a_{col} as in Example 2.3 can model more realistic use cases where collisions happen almost surely due to a large number of players/users.

Before concluding the discussion of monotonicity of payoffs, we also compare our payoff assumptions with two other common game-theoretical settings—potential games[3] [44] and their subset, congestion games [59], for which algorithms with IL guarantees already exist [37]. We show that SMFGs with monotone payoffs are not special cases of either of these settings. Specifically, we provide a counter-example of SMFG that does not admit a game potential.

Remark 2.5 (SMFG is not a potential game). The SMFG is a *potential game* if there exists a map $P : \mathcal{A}^N \to \mathbb{R}$ such that for all $i \in \mathcal{N}, \mathbf{a} = (a_1, \ldots, a_N) \in \mathcal{A}^N, a \in \mathcal{A}$, it holds that

$$V^i(a^i, \mathbf{a}^{-i}) - V^i(a, \mathbf{a}^{-i}) = P(a^i, \mathbf{a}^{-i}) - P(a, \mathbf{a}^{-i}), \tag{1}$$

[3]Not to be confused with $\mathbf{F}$ that admits a potential Φ such that $\mathbf{F} = \nabla\Phi$, also called a gradient field.

where $V^i(\mathbf{a})$ denotes the expected reward of player i when each player $j \in \mathcal{N}$ (deterministically) plays action a^j. Note that $V^i(\mathbf{a}) = \mathbf{F}(\widehat{\mu}(\mathbf{a}))(a^i)$ where $\widehat{\mu}(\mathbf{a}) \in \Delta_{\mathcal{A}}$ is the empirical distribution of actions over actions induced by action profile $\mathbf{a}$.

We provide an example of an SMFG where no such P exists. Let $\mathbf{S} \in \mathbb{S}^D_{++}$ be a symmetric positive definite matrix, $\mathbf{b} \in \mathbb{R}^D$ be an arbitrary vector, and $\mathbf{X} \in \mathbb{M}^{D,D}$ be a anti-symmetric matrix such that $\mathbf{X} = -\mathbf{X}^\top$. Take the payoff operator

$$\mathbf{F}(\mu) = (-\mathbf{S} + \mathbf{X})\mu + \mathbf{b},$$

which can be trivially shown to be monotone. In particular, take $N = 3, K = 3, \mathcal{A} = \{a_1, a_2, a_3\}$, and define the reward $\mathbf{F}$ as

$$\mathbf{F}\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} = \begin{pmatrix}-\mu_1 - \mu_2\\ \mu_1\\ -\mu_3\end{pmatrix},$$

which is monotone. Assume that a potential P exists and satisfies Equation (1). Then, it would follow that:

$$\begin{aligned}
V^1(a_2, a_1, a_1) - V^1(a_3, a_1, a_1) &= P(a_2, a_1, a_1) - P(a_3, a_1, a_1),\\
V^2(a_3, a_1, a_1) - V^2(a_3, a_2, a_1) &= P(a_3, a_1, a_1) - P(a_3, a_2, a_1),\\
V^1(a_3, a_2, a_1) - V^1(a_2, a_2, a_1) &= P(a_3, a_2, a_1) - P(a_2, a_2, a_1),\\
V^2(a_2, a_2, a_1) - V^2(a_2, a_1, a_1) &= P(a_2, a_2, a_1) - P(a_2, a_1, a_1).
\end{aligned}$$

Hence, adding the inequalities above, we have that

$$\begin{aligned}
0 &= V^1(a_2, a_1, a_1) - V^1(a_3, a_1, a_1) + V^2(a_3, a_1, a_1) - V^2(a_3, a_2, a_1)\\
&\quad + V^1(a_3, a_2, a_1) - V^1(a_2, a_2, a_1) + V^2(a_2, a_2, a_1) - V^2(a_2, a_1, a_1)\\
&= \mathbf{F}((2/3, 1/3, 0), a_2) - \mathbf{F}((2/3, 0, 1/3), a_3) + \mathbf{F}((2/3, 0, 1/3), a_1) - \mathbf{F}((1/3, 1/3, 1/3), a_2)\\
&\quad + \mathbf{F}((1/3, 1/3, 1/3), a_3) - \mathbf{F}((1/3, 2/3, 0), a_2) + \mathbf{F}((1/3, 2/3, 0), a_2) - \mathbf{F}((2/3, 1/3, 0), a_1)\\
&= 2/3 - (-1/3) + (-2/3) - 1/3 + (-1/3) - 1/3 + 1/3 - (-1) \neq 0,
\end{aligned}$$

leading to a contradiction. As a result, no such potential P could exist, and this (monotone) SMFG cannot be a potential game.

2.3 Assumption on Stochastic Noise

Finally, our results require the standard assumption of independent noise with bounded variance. We formalize this below in Assumption 3.

Assumption 3 (Independent, bounded variance noise). *We assume that the payoff noise is zero mean, has entrywise variance upper bounded by σ^2, and is independent. That is, we assume the following hold:*

(1) For all $t \geq 0, i \in \mathcal{N}$ and $a \in \mathcal{A}$, it holds that $\mathbb{E}[\mathbf{n}^i_t(a)] = 0$ and $\mathbb{E}[\mathbf{n}^i_t(a)^2] \leq \sigma^2$,
(2) $\{\mathbf{n}^i_t(a)\}_{i \in \mathcal{N}, t \geq 0, a \in \mathcal{A}}$ are independent random variables,
(3) For any $t \geq 0, i \in \mathcal{N}$, the random vector $\mathbf{n}^i_t$ is independent from $\{a^j_{t'}\}_{t' \leq t, j \in \mathcal{N}}$ (i.e., actions taken up to step t by all players).

The assumption of independent noise with bounded variance is common in bandits literature [5, 7, 9]. We emphasize that the noise vectors $\mathbf{n}^i_t$ have entry-wise bounded variance by σ^2 in Assumption 3. In optimization or VI settings, the assumption of independent sources of noise in stochastic oracles is also standard, although typically a bound on $\mathbb{E}[\|\mathbf{n}^i_t\|_2^2]$ is assumed [30].

3 The VI Approximation as $N \to \infty$

Our first set of theoretical results presented in this chapter makes the connection between the NE and a solution of (MF-VI) explicit. We will show that solutions of (MF-VI) form good approximations of the true NE in the N-player game if N is large.

Remark 3.1 (Existence and Uniqueness of MF-NE). Let $\mathbf{F} : \Delta_{\mathcal{A}} \to [0,1]^K$ be a continuous function. Then $\mathbf{F}$ has at least one MF-NE $\boldsymbol{\pi}^*$, and the set of MF-NE is compact. Furthermore, if $\mathbf{F}$ is also λ-strongly monotone for some $\lambda > 0$, then the MF-NE is unique. This can be seen as follows. The MF-NE corresponds to solutions of the VI: $\forall \boldsymbol{\pi} \in \Delta_{\mathcal{A}}, \mathbf{F}(\boldsymbol{\pi}^*)^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}) \geq 0$. The domain set $\Delta_{\mathcal{A}}$ is compact and convex, and the assumption that $\mathbf{F}$ is continuous yields the existence of a solution using Corollary 2.2.5 of Ref. [17]. For uniqueness in the case of strong monotonicity, see Theorem 2.3.3 of Ref. [17].

The following theorem shows that the solution of (MF-VI), when deployed by all players, is a $O(1/\sqrt{N})$ solution of the N-player game. Therefore, the MF-NE solution will be an arbitrarily good approximation of the true NE when $N \to \infty$, and the bias introduced by studying the N-player game can be explicitly quantified.

Theorem 3.2. *Let $\mathbf{F}$ be L-Lipschitz, $\delta \geq 0$ arbitrary, and let $\boldsymbol{\pi}^*$ be a δ-MF-NE. Then, the strategy profile $(\boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \in \Delta_{\mathcal{A}}^N$ is a $O(\delta + \frac{L}{\sqrt{N}})$-NE of the N-player SMFG.*

Proof. Firstly, define the independent random variables $a^j \sim \boldsymbol{\pi}^*$ for all $j \in \mathcal{N}$ for a δ-MF-NE $\boldsymbol{\pi}^* \in \Delta_{\mathcal{A}}$. Define the random variable $\widehat{\mu} := {}^1/_N \sum_{j=1}^N \mathbf{e}_{a^j}$, which is the empirical distribution of players over actions in a single round of an SMFG. The proof will proceed by formally proving that if N is large enough, then $\mathbb{E}[\mathbf{F}(\widehat{\mu})] \approx \mathbf{F}(\boldsymbol{\pi}^*)$ and a^j is approximately independent from $\mathbf{F}(\widehat{\mu})$.

It is straightforward that $\mathbb{E}[\widehat{\mu}] = \boldsymbol{\pi}^*$. Furthermore, by independence of the random vectors $\mathbf{e}_{a^j}$, we have

$$\mathbb{E}\left[\|\widehat{\mu} - \boldsymbol{\pi}^*\|_2\right] \leq \sqrt{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^N \mathbf{e}_{a^j} - \boldsymbol{\pi}^*\right\|_2^2\right]} \leq \sqrt{\frac{1}{N^2}\sum_{j=1}^N \mathbb{E}\left[\|\mathbf{e}_{a^j} - \boldsymbol{\pi}^*\|_2^2\right]} \leq \frac{2}{\sqrt{N}}.$$

Hence, as $\mathbf{F}$ is L-Lipschitz, we have that

$$\|\,\mathbb{E}[\mathbf{F}(\widehat{\mu})|a_j \sim \boldsymbol{\pi}^*] - \mathbf{F}(\boldsymbol{\pi}^*)\|_2 \leq \mathbb{E}[\|\mathbf{F}(\widehat{\mu}) - \mathbf{F}(\boldsymbol{\pi}^*)\|_2] \leq \frac{2L}{\sqrt{N}}. \tag{2}$$

Now let $i \in \mathcal{N}$ be arbitrary, and let $\boldsymbol{\pi}' \in \Delta_{\mathcal{A}}$ be any distribution over actions that satisfies $V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) = \max_{\boldsymbol{\pi}} V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{*,-i})$. We also define the quantities

$$\overline{\mathbf{F}}_1 = \mathbb{E}\left[\mathbf{F}(\widehat{\mu})\middle|a^j \sim \boldsymbol{\pi}^*, \forall j \in \mathcal{N}\right], \qquad \overline{\mathbf{F}}_2 = \mathbb{E}\left[\mathbf{F}(\widehat{\mu})\middle|a^j \sim \boldsymbol{\pi}^* \text{ for } \forall i \neq j, \quad a^i \sim \boldsymbol{\pi}'\right].$$

We will bound $V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) - V^i(\boldsymbol{\pi}^*, \boldsymbol{\pi}^{*,-i})$. Combining Lemma 2.1 and the inequality (2), we observe

$$\begin{aligned}
|V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}'^\top \mathbf{F}(\boldsymbol{\pi}^*)| \leq & |V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}'^\top \overline{\mathbf{F}}_2| + \left|\boldsymbol{\pi}'^\top \overline{\mathbf{F}}_2 - \boldsymbol{\pi}'^\top \mathbf{F}\left(\frac{N-1}{N}\boldsymbol{\pi}^* + \frac{1}{N}\boldsymbol{\pi}'\right)\right| \\
& + \left|\boldsymbol{\pi}'^\top \mathbf{F}\left(\frac{N-1}{N}\boldsymbol{\pi}^* + \frac{1}{N}\boldsymbol{\pi}'\right) - \boldsymbol{\pi}'^\top \mathbf{F}(\boldsymbol{\pi}^*)\right| \\
\leq & \frac{L\sqrt{2}}{N} + \frac{2L}{\sqrt{N}} + \frac{2L}{N},
\end{aligned}$$

since $\mathbf{F}$ is L-Lipschitz. Likewise, using Lemma 2.1 once again, we have

$$\begin{aligned}|V^i(\boldsymbol{\pi}^*, \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}^{*,\top}\mathbf{F}(\boldsymbol{\pi}^*)| &\leq |V^i(\boldsymbol{\pi}^*, \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}^{*,\top}\overline{\mathbf{F}}_1| + |\boldsymbol{\pi}^{*,\top}\overline{\mathbf{F}}_1 - \boldsymbol{\pi}^{*,\top}\mathbf{F}(\boldsymbol{\pi}^*)| \\ &\leq \frac{L\sqrt{2}}{N} + \frac{2L}{\sqrt{N}}.\end{aligned}$$

Finally, using the definition of a δ-MF-NE, it holds that

$$\begin{aligned}V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) - V^i(\boldsymbol{\pi}^*, \boldsymbol{\pi}^{*,-i}) \leq& \mathbf{F}(\boldsymbol{\pi}^*)^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) + |V^i(\boldsymbol{\pi}^*, \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}^{*,\top}\mathbf{F}(\boldsymbol{\pi}^*)| \\ &+ |V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{*,-i}) - \boldsymbol{\pi}'^{\top}\mathbf{F}(\boldsymbol{\pi}^*)| \\ \leq& \delta + \frac{L(2\sqrt{2}+4)}{N} + \frac{4L}{\sqrt{N}}.\end{aligned}$$

□

Recall our goal in the context of the N-player SMFG is to find policies $\{\boldsymbol{\pi}^j\}_{j=1}^N$ with low exploitability $\mathcal{E}^i_{\text{exp}}$ for all i. Theorem 3.2 considers that agents adopt the same policy $\boldsymbol{\pi}^*$ from solving the VI corresponding to operator $\mathbf{F}$ to obtain a low-exploitability approximation. We will generalize this result to explicitly bound $\mathcal{E}^i_{\text{exp}}$ when agent policies can also deviate and when agents can employ regularization. In our algorithms, regularizing the MF-VI problem will play a crucial role in the IL setting, as it will prevent the policies of agents from diverging when there is no centralized controller synchronizing the policies of agents. For this reason, our algorithms in the later sections will introduce extraneous regularization to (MF-VI) and instead solve the following τ-Tikhonov regularized VI problem:

$$\text{Find } \boldsymbol{\pi}^* \in \Delta_{\mathcal{A}} \text{ s.t. } (\mathbf{F} - \tau\mathbf{I})(\boldsymbol{\pi}^*)^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}) \geq 0, \forall \boldsymbol{\pi} \in \Delta_{\mathcal{A}}. \tag{MF-RVI}$$

The following theorem quantifies the additional exploitability incurred in the N-player game due to (1) extraneous regularization, which is useful for algorithm design, and (2) deviations in agent policies from the MF-NE, potentially due to stochasticity in learning. Theorem 3.3 will be a more useful result later in a learning setting since the learned policies $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N)$ will only approximate the solution of (MF-RVI).

Theorem 3.3. *Let $\mathbf{F}$ be monotone, L-Lipschitz. Let $\boldsymbol{\pi}^*_\tau \in \Delta_{\mathcal{A}}$ be the (unique) MF-NE of the regularized map $\mathbf{F} - \tau\mathbf{I}$. Let $\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N \in \Delta_{\mathcal{A}}$ be such that $\|\boldsymbol{\pi}^i - \boldsymbol{\pi}^*_\tau\|_2 \leq \delta$ for all i, then it holds that $\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N) = O(\tau + \delta + 1/\sqrt{N})$ for all $i \in \mathcal{N}$.*

Proof. By the Lipschitz continuity of exploitability (Lemma 2.2), we have

$$\begin{aligned}\mathcal{E}^i_{\text{exp}}\left(\{\boldsymbol{\pi}^j\}_{j=1}^N\right) \leq& \mathcal{E}^i_{\text{exp}}\left(\{\boldsymbol{\pi}^*_\tau\}_{j=1}^N\right) + \sqrt{K}\|\boldsymbol{\pi}^i - \boldsymbol{\pi}^*_\tau\|_2 + \sum_{j\neq i}\frac{4L\sqrt{2K}}{N}\|\boldsymbol{\pi}^j - \boldsymbol{\pi}^*_\tau\|_2 \\ \leq& \mathcal{E}^i_{\text{exp}}\left(\{\boldsymbol{\pi}^*_\tau\}_{j=1}^N\right) + \delta\sqrt{K} + 4L\sqrt{2K}\delta. \end{aligned} \tag{3}$$

Since $\boldsymbol{\pi}^*_\tau$ is the unique MF-NE of the operator $\mathbf{F} - \tau\mathbf{I}$, it holds by definition that

$$(\mathbf{F} - \tau\mathbf{I})(\boldsymbol{\pi}_\tau)^\top\boldsymbol{\pi}_\tau \geq (\mathbf{F} - \tau\mathbf{I})(\boldsymbol{\pi}_\tau)^\top\boldsymbol{\pi}.$$

Organizing both sides, we have

$$\mathbf{F}(\boldsymbol{\pi}_\tau)^\top\boldsymbol{\pi}_\tau \geq \mathbf{F}(\boldsymbol{\pi}_\tau)^\top\boldsymbol{\pi} + \tau\boldsymbol{\pi}_\tau^\top(\boldsymbol{\pi}_\tau - \boldsymbol{\pi}) \geq \mathbf{F}(\boldsymbol{\pi}_\tau)^\top\boldsymbol{\pi} - 2\tau,$$

as $|\boldsymbol{\pi}_\tau^\top(\boldsymbol{\pi}_\tau - \boldsymbol{\pi})| \leq \|\boldsymbol{\pi}_\tau\|_2\|\boldsymbol{\pi}_\tau - \boldsymbol{\pi}\|_2 \leq 2$. Then, $\boldsymbol{\pi}^*_\tau$ is a 2τ-MF-NE for the operator $\mathbf{F}$, and by Theorem 3.2, $\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^*_\tau\}_{j=1}^N) \leq O(\tau + 1/\sqrt{N})$. Placing this in (3) proves the theorem. □

ALGORITHM 1: TRPA-Full: IL with Full Feedback Algorithm for Each Agent $i \in \mathcal{N}$.

Require: Number of actions K, regularization $\tau > 0$, learning rate $\{\eta_t\}_{t=0}^{T}$, rounds $T > 0$.
$\boldsymbol{\pi}_0^i \leftarrow \frac{1}{K}\mathbf{1}$
for $t = 0, \dots, T-1$ **do**
  Play action with current policy $a_t^i \sim \boldsymbol{\pi}_t^i$.
  Observe payoff $\mathbf{r}_t^i$
  $\boldsymbol{\pi}_{t+1}^i = \Pi_{\Delta_{\mathcal{A}}}((1-\tau\eta_t)\boldsymbol{\pi}_t^i + \eta_t \mathbf{r}_t^i)$
end for
Return $\boldsymbol{\pi}_T^i$

To summarize, this section presented key approximation results linking the solutions of (MF-VI) and (MF-RVI) to the N-player exploitability in the SMFG. The next sections will be devoted to designing sample-efficient IL algorithms.

4 Convergence in the Full Feedback Case

We first present an IL algorithm for the full feedback setting, as a first step towards analyzing the more interesting bandit feedback setting. In this setting, while there is no centralized controller, independent noisy reports of all action payoffs are available to each agent after each round.

Is it possible to simply solve MF-RVI in our IL setting? Before we present our results, we note the following: Past works in MFG have already proved approximation results of N-agent games by MFG albeit in different settings [60, 74], but these results do not consider when *learning itself* is carried out with N agents. If N agents cannot communicate, it is theoretically challenging to approximate the MF-RVI and to tackle bandit feedback. Most importantly, the IL algorithms formalized in Section B cannot query an operator oracle or maintain a common iterate throughout repeated plays. Therefore, the approximation properties of (MF-VI) do not immediately imply the MF-NE can be learned using VI algorithms. In this section and the next, we prove the more challenging result of convergence with IL, first under full feedback and later under partial (bandit) feedback.

Our analysis builds up on **Tikhonov regularized projected ascent** (**TRPA**). The TRPA operator is defined as

$$\Gamma^{\eta,\tau}(\boldsymbol{\pi}) := \Pi_{\Delta_{\mathcal{A}}}(\boldsymbol{\pi} + \eta(\mathbf{F} - \tau\mathbf{I})(\boldsymbol{\pi})) = \Pi_{\Delta_{\mathcal{A}}}((1-\eta\tau)\boldsymbol{\pi} + \eta\mathbf{F}(\boldsymbol{\pi})), \tag{TRPA}$$

for a learning rate $\eta > 0$ and regularization $\tau > 0$. Intuitively, $\Gamma^{\eta,\tau}$ uses $\mathbf{F}$ evaluated at $\boldsymbol{\pi}$ to modify action probabilities in the direction of the greatest payoff, incorporating an ℓ_2 regularizer of τ. Furthermore, the unique MF-NE $\boldsymbol{\pi}^*$ of (MF-RVI) is also a fixed point of $\Gamma^{\eta,\tau}$. The analysis of TRPA is standard and known to converge for monotone $\mathbf{F}$ [17, 47], when (stochastic) oracle access to $\mathbf{F}$ is assumed. Naturally, the main complication in applying the method above will be the fact that in the IL setting, agents cannot evaluate the operator $\mathbf{F}$ arbitrarily, but rather can only observe (a noisy) estimate of $\mathbf{F}$ as a function of the empirical population distribution and not of their policy $\boldsymbol{\pi}$. In the full feedback setting, we analyze the following dynamics:

$$\boldsymbol{\pi}_0^i := \mathrm{Unif}(\mathcal{A}) = \frac{1}{K}\mathbf{1}_K, \quad \boldsymbol{\pi}_{t+1}^i = \Pi_{\Delta_{\mathcal{A}}}\left((1-\tau\eta_t)\boldsymbol{\pi}_t^i + \eta_t \mathbf{r}_t^i\right), \tag{TRPA-Full}$$

for a time varying learning rate η_t, for each agent $i \in \mathcal{N}$. The extraneous ℓ_2-regularization incorporated in each agent running TRPA-Full is critical for the analysis and convergence in IL, as it allows explicit synchronization of policies of agents without communication. We state the TRPA-Full algorithm in Algorithm 1 for reference.

We state the following standard result regarding the TRPA operator without proof, as it will be used later.

Lemma 4.1 (cf. Theorem 12.1.2 of [17]). *Assume* $\mathbf{F}$ *is* $\lambda \geq 0$*-monotone and* L*-Lipschitz. Then* $\Gamma^{\eta,\tau}$ *is Lipschitz with constant* $\sqrt{1-2(\lambda+\tau)\eta+\eta^2(L+\tau)^2}$ *with respect to the* ℓ_2*-norm.*

For abuse of notation, let $\boldsymbol{\pi}^* \in \Delta_{\mathcal{A}}$ be the unique solution of (MF-RVI) for the regularization $\tau > 0$. Also define the sigma algebra $\mathcal{F}_t := \mathcal{F}(\{\boldsymbol{\pi}_{t'}^i\}_{t'=0,\dots,t}^{i=1,\dots,N})$. We maintain the definitions of the core random variables of the SMFG dynamics introduced in Section 1.1,

$$\widehat{\mu}_t := \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_{a_t^i}, \qquad \mathbf{r}_t^i := \mathbf{F}(\widehat{\mu}_t) + \mathbf{n}_t^i.$$

Under TRPA-Full dynamics, we also define the following random variables that assist our analysis.

$$\bar{\mu}_t := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}_t^i, \qquad e_t^i := \left\|\boldsymbol{\pi}_t^i - \bar{\mu}_t\right\|_2^2, \qquad u_t^i := \mathbb{E}\left[\left\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\right\|_2^2\right].$$

We call $\bar{\mu}_t$ the mean policy, e_t^i the mean policy deviation, and u_t^i the expected ℓ_2-deviation from the regularized MF-NE. Our goal is to bound the sequence or error terms u_t^i; however, the process is complicated by the fact that in general the policy deviations of agents e_t^i are nonzero. Our strategy is as follows: (1) derive a recursion for u_t^i incorporating the terms e_t^i, (2) bound the terms e_t^i, showing the deviation of the policies of the agents goes to zero in expectation, and (3) solve the recursion to obtain the convergence rate.

The following lemma captures the first step and provides a recurrence for the evolution of u_t^i under TRPA-Full.

Lemma 4.2 (Error recurrence under full feedback). *Under TRPA-Full with learning rates* η_t*, it holds for* L*-Lipschitz and* λ*-strongly monotone* $\mathbf{F}$ *that*

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}^*\right\|_2^2\right] \leq & 3\eta_t^2 K(1+\sigma^2) + 2\eta_t^2(L+\tau)^2 + \frac{4\eta_t L^2\lambda^{-1}}{N} \\ & + 2\eta_t L^2\lambda^{-1}\,\mathbb{E}\left[e_t^i\right] + (1-2\eta_t(\lambda/2+\tau))\,\mathbb{E}\left[\left\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\right\|_2^2\right],\end{aligned}$$

and for L*-Lipschitz and monotone* $\mathbf{F}$ *that*

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}^*\right\|_2^2\right] \leq & 3\eta_t^2 K(1+\sigma^2) + 2\eta_t^2(L+\tau)^2 + \frac{4\tau^{-1}\eta_t L^2\delta^{-1}}{N} \\ & + \tau^{-1}\eta_t L^2\delta^{-1}\,\mathbb{E}\left[e_t^i\right] + (1-2\tau\eta_t(1-\delta))\,\mathbb{E}\left[\left\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\right\|_2^2\right],\end{aligned}$$

for arbitrary $\delta \in (0,1)$.

Proof. We analyze for any $i \in [N]$ the error term $\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2$. Denote $\alpha_t := (1-\tau\eta_t)$. For the regularized solution $\boldsymbol{\pi}^*$, we have the fixed point result

$$\Pi_{\Delta_{\mathcal{A}}}((1-\tau\eta_t)\boldsymbol{\pi}^* + \eta_t\mathbf{F}(\boldsymbol{\pi}^*)) = \Pi_{\Delta_{\mathcal{A}}}(\boldsymbol{\pi}^* + \eta_t(\mathbf{F}-\tau\mathbf{I})(\boldsymbol{\pi}^*)) = \boldsymbol{\pi}^*.$$

The proof strategy is to decompose the ℓ_2 distance of player policies to $\boldsymbol{\pi}^*$ into 3 components using this property. We can bound the quantity $\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}^*\|_2^2$ by using the non-expansiveness of $\Pi_{\Delta_{\mathcal{A}}}$:

$$\begin{aligned}\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}^*\|_2^2 =& \|\Pi_{\Delta_{\mathcal{A}}}(\alpha_t\boldsymbol{\pi}_t^i + \eta_t\mathbf{r}_t^i) - \Pi_{\Delta_{\mathcal{A}}}(\alpha_t\boldsymbol{\pi}^* + \eta_t\mathbf{F}(\boldsymbol{\pi}^*))\|_2^2 \\ \leq& \|\alpha_t\boldsymbol{\pi}_t^i + \eta_t\mathbf{F}(\boldsymbol{\pi}_t^i) - \alpha_t\boldsymbol{\pi}^* - \eta_t\mathbf{F}(\boldsymbol{\pi}^*) + \eta_t(\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i))\|_2^2 \\ =& \eta_t^2\|\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2 + 2\eta_t(\alpha_t(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*) + \eta_t(\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*)))^\top(\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i))\end{aligned}$$

$$
\begin{aligned}
&+ \|\alpha_t(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*) + \eta_t(\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*))\|_2^2 \\
&\leq \underbrace{\eta_t^2 \|\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2 + 2\eta_t^2 (\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*))^\top (\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i))}_{(a)} \\
&+ \underbrace{2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i))}_{(b)} + \underbrace{\|\alpha_t(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*) + \eta_t(\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*))\|_2^2}_{(c)}. \qquad (4)
\end{aligned}
$$

We analyze the three marked terms separately. For term (a), using the independence assumption of the noise vectors and Young's inequality, in expectation we obtain

$$
\begin{aligned}
\mathbb{E}[(a)] \leq& \eta_t^2 \, \mathbb{E}[\|\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2] + \eta_t^2 \, \mathbb{E}[\|\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*)\|_2^2] + \eta_t^2 \, \mathbb{E}[\|\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2] \\
\leq& 2\eta_t^2 \, \mathbb{E}[\|\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2] + \eta_t^2 \, \mathbb{E}[\|\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*)\|_2^2] \\
\leq& 2\eta_t^2 \, \mathbb{E}[\|\mathbf{r}_t^i - \mathbf{F}(\widehat{\mu}_t)\|_2^2 + \|\mathbf{F}(\widehat{\mu}_t) - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2] + \eta_t^2 K \\
\leq& 2\eta_t^2 \sigma^2 K + 3\eta_t^2 K \leq 3\eta_t^2 K(\sigma^2 + 1)
\end{aligned}
$$

For the term (c), we obtain

$$
\begin{aligned}
(c) =& \|\alpha_t(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*) + \eta_t(\mathbf{F}(\boldsymbol{\pi}_t^i) - \mathbf{F}(\boldsymbol{\pi}^*))\|_2^2 \\
=& \|(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*) + \eta_t(\mathbf{F}(\boldsymbol{\pi}_t^i) - \tau\boldsymbol{\pi}_t^i - \mathbf{F}(\boldsymbol{\pi}^*) + \tau\boldsymbol{\pi}^*)\|_2^2 \\
\leq& \left(1 - 2(\lambda + \tau)\eta_t + (L + \tau)^2 \eta_t^2\right) \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2 \\
\leq& \left(1 - 2(\lambda + \tau)\eta_t\right) \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2 + 2(L + \tau)^2 \eta_t^2
\end{aligned}
$$

where the last inequality holds from the Lipschitz continuity result of Lemma 4.1.

For the term (b), first taking the strongly monotone problem $\lambda > 0$, we have that

$$
\begin{aligned}
(b) =& 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\boldsymbol{\pi}_t^i)) \\
=& 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\widehat{\mu}_t)) + 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{F}(\widehat{\mu}_t) - \mathbf{F}(\bar{\mu}_t)) \\
&+ 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{F}(\bar{\mu}_t) - \mathbf{F}(\boldsymbol{\pi}_t^i)) \\
\leq& 2\eta_t \alpha_t \left(\frac{\lambda}{4} \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{\lambda} \|\mathbf{F}(\widehat{\mu}_t) - \mathbf{F}(\bar{\mu}_t)\|_2^2\right) + 2\eta_t \alpha_t \left(\frac{\lambda}{4} \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{\lambda} \|\mathbf{F}(\bar{\mu}_t) - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2\right) \\
&+ 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\widehat{\mu}_t)) \\
\leq& \eta_t \lambda \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2^2 + 2\eta_t \lambda^{-1} \|\mathbf{F}(\widehat{\mu}_t) - \mathbf{F}(\bar{\mu}_t)\|_2^2 + 2\eta_t \lambda^{-1} \|\mathbf{F}(\bar{\mu}_t) - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2 \\
&+ 2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\widehat{\mu}_t)),
\end{aligned}
$$

which follows from applications of Young's inequality. For the last three terms we observe:

$$
\begin{aligned}
\mathbb{E}\left[2\eta_t \alpha_t (\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*)^\top (\mathbf{r}_t^i - \mathbf{F}(\widehat{\mu}_t)) | \mathcal{F}_t\right] =& 0, \\
\mathbb{E}[\|\mathbf{F}(\widehat{\mu}_t) - \mathbf{F}(\bar{\mu}_t)\|_2^2 | \mathcal{F}_t] \leq& L^2 \, \mathbb{E}\left[\|\widehat{\mu}_t - \bar{\mu}_t\|_2^2 | \mathcal{F}_t\right] \\
\leq& L^2 \, \mathbb{E}\left[\frac{1}{N^2} \| \sum_i \boldsymbol{\pi}_t^i - \sum_i \mathbf{e}_{a_t^i} \|_2^2 | \mathcal{F}_t\right] \\
=& \frac{L^2}{N^2} \sum_i \mathbb{E}[\|\boldsymbol{\pi}_t^i - \mathbf{e}_{a_t^i}\|_2^2 | \mathcal{F}_t] \leq \frac{2L^2}{N}, \\
\|\mathbf{F}(\bar{\mu}_t) - \mathbf{F}(\boldsymbol{\pi}_t^i)\|_2^2 \leq& L^2 \|\bar{\mu}_t - \boldsymbol{\pi}_t^i\|_2^2 = L^2 e_t^i.
\end{aligned}
$$

The second inequality above follows from the fact that $\widehat{\mu}_t$ is the sum of N independent random variables and has expectation $\bar{\mu}_t$. Hence, putting in the bounds for (a), (b), (c) and taking expectations, we obtain the inequality

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}^i_{t+1}-\boldsymbol{\pi}^*\|_2^2\right] \leq & 3\eta_t^2K(1+\sigma^2)+\frac{4\eta_tL^2}{\lambda N}+\frac{2\eta_tL^2}{\lambda}\mathbb{E}\left[e^i_t\right]\\ &+(1-2(\lambda/2+\tau)\eta_t)\,\mathbb{E}\left[\|\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*\|_2^2\right]+2(L+\tau)^2\eta_t^2.\end{aligned}$$

Turning back to the monotone case, if $\lambda = 0$, vary the upper bound on (b) as follows. Take any arbitrary $\delta \in (0,1)$. Then, once again applying Young's inequality, we obtain

$$\begin{aligned}(b) =& 2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{r}^i_t-\mathbf{F}(\boldsymbol{\pi}^i_t))\\ =& 2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{r}^i_t-\mathbf{F}(\widehat{\mu}_t))+2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{F}(\widehat{\mu}_t)-\mathbf{F}(\bar{\mu}_t))\\ &+2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{F}(\bar{\mu}_t)-\mathbf{F}(\boldsymbol{\pi}^i_t))\\ \leq& 2\eta_t\alpha_t\left(\frac{\tau\delta}{2}\|\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*\|_2^2+\frac{1}{2\tau\delta}\|\mathbf{F}(\widehat{\mu}_t)-\mathbf{F}(\bar{\mu}_t)\|_2^2\right)+2\eta_t\alpha_t\left(\frac{\tau\delta}{2}\|\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*\|_2^2+\frac{1}{2\tau\delta}\|\mathbf{F}(\bar{\mu}_t)-\mathbf{F}(\boldsymbol{\pi}^i_t)\|_2^2\right)\\ &+2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{r}^i_t-\mathbf{F}(\widehat{\mu}_t))\\ \leq& 2\eta_t\tau\delta\|\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*\|_2^2+\frac{\eta_t}{\tau\delta}\|\mathbf{F}(\widehat{\mu}_t)-\mathbf{F}(\bar{\mu}_t)\|_2^2+\frac{\eta_t}{\tau\delta}\|\mathbf{F}(\bar{\mu}_t)-\mathbf{F}(\boldsymbol{\pi}^i_t)\|_2^2\\ &+2\eta_t\alpha_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^*)^\top(\mathbf{r}^i_t-\mathbf{F}(\widehat{\mu}_t)).\end{aligned}$$

Applying the same bounds for the terms (a), (c) as before yields the lemma. □

This above lemma has two key features: a dependence on expected mean policy deviation $\mathbb{E}[e^i_t]$, and a term that scales as $O(1/N)$. While the $O(1/N)$ term can be anticipated (and asymptotically ignored when N is large) due to the finite-agent mean-field bias (as shown previously in Section 3), the term $\mathbb{E}[e^i_t]$ must be controlled separately in the independent learning setting, where policies cannot be synchronized through explicit communication between agents. The term $\mathbb{E}[e^i_t]$ reflects the core difference of the SMFG model from typical VI stochastic oracles. Unlike typical VI oracle models, in SMFG the operator $\mathbf{F}$ cannot be evaluated at the current iterate $\boldsymbol{\pi}^i_t$ of a player i but only approximately at the mean $\bar{\mu}_t$. This is due to decentralized learning: players can only evaluate the current payoffs at the "mean-iterate" given by $\mathbf{F}(\widehat{\mu}_t) \approx \mathbf{F}(\bar{\mu}_t)$ (up to some stochastic noise) that is almost surely different than their iterates $\{\boldsymbol{\pi}^i_t\}_i$ apart from the case with degenerate/zero noise. Furthermore, Lemma 4.2 suggests that the algorithmic scheme must guarantee that $\mathbb{E}[e^i_t]$ decays with the rate at least $O(1/t)$ to obtain a non-vacuous bound on exploitability. Taking inspiration from algorithmic stability literature [1, 78], we utilize a regularization scheme to ensure the iterates of players do not diverge. The following lemma shows that by introducing explicit regularization $\tau > 0$, the expected mean policy deviation can be controlled throughout training.

Lemma 4.3 (Policy variations bound). *Under TRPA-Full with learning rates* $\eta_t := \frac{\tau^{-1}}{t+2}$, *we have* $\mathbb{E}[e^i_t] \leq \frac{14\tau^{-2}K\sigma^2+14}{t+2}$.

Proof. Note that for any $i, j \in \mathcal{N}$ such that $i \neq j$, using the non-expansiveness of the projection operator it holds that

$$\begin{aligned}\|\boldsymbol{\pi}^i_{t+1}-\boldsymbol{\pi}^j_{t+1}\|_2^2 =& \|\Pi_{\Delta_{\mathcal{A}}}((1-\tau\eta_t)\boldsymbol{\pi}^i_t+\eta_t\mathbf{r}^i_t)-\Pi_{\Delta_{\mathcal{A}}}((1-\tau\eta_t)\boldsymbol{\pi}^j_t+\eta_t\mathbf{r}^j_t)\|_2^2\\ \leq& \|(1-\tau\eta_t)\boldsymbol{\pi}^i_t+\eta_t\mathbf{r}^i_t-(1-\tau\eta_t)\boldsymbol{\pi}^j_t-\eta_t\mathbf{r}^j_t\|_2^2\\ \leq& \|(1-\tau\eta_t)(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^j_t)+\eta_t(\mathbf{r}^i_t-\mathbf{r}^j_t)\|_2^2\\ =& (1-\tau\eta_t)^2\|\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^j_t\|_2^2+\eta_t^2\|\mathbf{r}^i_t-\mathbf{r}^j_t\|_2^2+2(1-\tau\eta_t)\eta_t(\boldsymbol{\pi}^i_t-\boldsymbol{\pi}^j_t)^\top(\mathbf{r}^i_t-\mathbf{r}^j_t)\end{aligned}$$

Taking the conditional expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}_{t+1}^j\|_2^2|\mathcal{F}_t\right] \leq& (1-\tau\eta_t)^2\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2 + \mathbb{E}\left[\eta_t^2\|\mathbf{r}_t^i - \mathbf{r}_t^j\|_2^2|\mathcal{F}_t\right] \\ &+ 2(1-\tau\eta_t)\eta_t(\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j)^\top \mathbb{E}\left[\mathbf{r}_t^i - \mathbf{r}_t^j|\mathcal{F}_t\right] \\ =& (1-\tau\eta_t)^2\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2 + \eta_t^2\,\mathbb{E}\left[\|\mathbf{n}_t^i - \mathbf{n}_t^j\|_2^2|\mathcal{F}_t\right] \\ =& (1-\tau\eta_t)^2\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2 + 2\eta_t^2 K\sigma^2\end{aligned}$$

almost surely, since we have $\mathbf{r}_t^i := \mathbf{F}(\widehat{\mu}_t) + \mathbf{n}_t^i$. Then, taking the expectation on both sides,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}_{t+1}^j\|_2^2\right] \leq& (1-\tau\eta_t)^2\,\mathbb{E}\left[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2\right] + 2\eta_t^2 K\sigma^2 \\ \leq& \left(1 - \frac{1}{t+2}\right)^2 \mathbb{E}\left[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2\right] + \left(\frac{\tau^{-1}}{t+2}\right)^2 2K\sigma^2 \\ \leq& \left(1 - \frac{2}{t+2}\right)\mathbb{E}\left[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2\right] + \frac{1}{(t+2)^2}\,\mathbb{E}\left[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2\right] + \frac{2\tau^{-2}K\sigma^2}{(t+2)^2} \\ \leq& \left(1 - \frac{2}{t+2}\right)\mathbb{E}\left[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2\right] + \frac{2\tau^{-2}K\sigma^2 + 2}{(t+2)^2}\end{aligned}$$

To bound the recurrence, we can use the recurrence lemma (Lemma C.2, noting $\gamma = 2, a = 2, u_0 = 0, c_0 = 0, c_1 = 2\tau^{-2}K\sigma^2 + 2$ in its statement):

$$\mathbb{E}\left[\|\boldsymbol{\pi}_{t+1}^i - \boldsymbol{\pi}_{t+1}^j\|_2^2\right] \leq 5\frac{2\tau^{-2}K\sigma^2 + 2}{(t+2)^2} + 3\frac{2\tau^{-2}K\sigma^2 + 2}{t+2} + \frac{2\tau^{-2}K\sigma^2 + 2}{(t+2)^2} \leq \frac{14\tau^{-2}K\sigma^2 + 14}{t+2}.$$

Then, the expected values of e_t^i can be bounded using:

$$e_t^i = \|\boldsymbol{\pi}_t^i - \bar{\mu}_t\|_2^2 = \left\|\boldsymbol{\pi}_t^i - \frac{1}{N}\sum_{j=1}^N \boldsymbol{\pi}_t^j\right\|_2^2 \leq \frac{1}{N}\sum_{j=1}^N \|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}_t^j\|_2^2$$

by an application of Jensen's inequality. Then, we have $\mathbb{E}[e_t^i] \leq \frac{14\tau^{-2}K\sigma^2+14}{t+2}$. □

With an explicit bound in expectation on the mean policy deviation e_t^i, we can now proceed to the main recurrence for the expected error terms u_t^i in order to prove our main convergence result. We state our main convergence result for TRPA-Full dynamics in Theorem 4.4 by solving these two recurrences for the monotone and strongly monotone cases.

Theorem 4.4 (Convergence, full feedback). *Assume* $\mathbf{F}$ *is Lipschitz, monotone. Assume* N *agents run the TRPA-Full update rule for* T *time steps with learning rates* $\eta_t := \frac{\tau^{-1}}{t+2}$ *and arbitrary regularization* $\tau > 0$. *Then it holds for any* $i \in [N]$ *that* $\mathbb{E}[\mathcal{E}_{exp}^i(\{\boldsymbol{\pi}_T^j\}_{j=1}^N)] \leq O(\frac{\tau^{-2}}{\sqrt{T}} + \frac{\tau^{-1}}{\sqrt{N}} + \tau)$. *Furthermore, if* $\mathbf{F}$ *is* λ*-strongly monotone, then* $\mathbb{E}[\mathcal{E}_{exp}^i(\{\boldsymbol{\pi}_T^j\}_{j=1}^N)] \leq O(\frac{\tau^{-3/2}\lambda^{-1/2}}{\sqrt{T}} + \frac{\tau^{-1/2}\lambda^{-1/2}}{\sqrt{N}} + \tau)$.

Proof. Note that the exploitability in the main statement of the theorem can be related to u_t^i as follows using Lemma 2.2:

$$\begin{aligned}\mathbb{E}[\mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}_t^j\}_{j=1}^N)] \leq& \mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}^*\}_{j=1}^N) + \sqrt{K}\,\mathbb{E}[\|\boldsymbol{\pi}_t^i - \boldsymbol{\pi}^*\|_2] + \frac{4L\sqrt{2K}}{N}\sum_{j\neq i}\mathbb{E}[\|\boldsymbol{\pi}_t^j - \boldsymbol{\pi}^*\|_2]\\ \leq& \mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}^*\}_{j=1}^N) + \sqrt{K}\sqrt{u_t^i} + \frac{4L\sqrt{2K}}{N}\sum_{j\neq i}\sqrt{u_t^j}\\ \leq& \mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}^*\}_{j=1}^N) + \frac{\max\{\sqrt{K}, 4L\sqrt{2K}\}}{N}\sum_{j}\sqrt{u_t^j}\end{aligned}$$

Hence the bounds on u_t^j will yield the result of the theorem by linearity of expectation, along with an invocation of Theorem 3.3.

Finally, we solve the recurrences for $\lambda = 0$ and $\lambda > 0$ using Lemma 4.2. For the case $\lambda > 0$, if $\eta_t = \frac{\tau^{-1}}{t+2}$, Lemma 4.2 provides the bound

$$\begin{aligned}u_{t+1}^i \leq& \frac{3\tau^{-2}K(1+\sigma^2) + 2\tau^{-2}(L+\tau)^2}{(t+2)^2} + \frac{4\tau^{-1}L^2\lambda^{-1}}{N(t+2)} + \frac{2\tau^{-1}L^2\lambda^{-1}}{t+2}\,\mathbb{E}\left[e_t^i\right]\\ &+ \left(1 - \frac{2\tau^{-1}(\lambda/2 + \tau)}{t+2}\right)u_t^i.\end{aligned}$$

By placing $\mathbb{E}[e_t^i] \leq \frac{14\tau^{-2}K\sigma^2+14}{t+2}$ due to Lemma 4.3, we obtain

$$\begin{aligned}u_{t+1}^i \leq& \frac{3\tau^{-2}K(1+\sigma^2) + 2\tau^{-2}(L+\tau)^2 + 28\tau^{-3}K\sigma^2\lambda^{-1}L^2 + 28\tau^{-1}L^2\lambda^{-1}}{(t+2)^2}\\ &+ \frac{4\tau^{-1}L^2\lambda^{-1}}{N(t+2)} + \left(1 - \frac{2}{t+2}\right)u_t^i.\end{aligned}$$

Invoking a generic recurrence lemma (Lemma C.2 in Appendix C) leads to the main statement of the theorem.

For the monotone case $\lambda = 0$, we have the recursion:

$$\begin{aligned}u_{t+1}^i \leq& \frac{3\tau^{-2}K(1+\sigma^2) + 2\tau^{-2}(L+\tau)^2}{(t+2)^2} + \frac{4\tau^{-2}L^2\delta^{-1}}{N(t+2)} + \frac{\tau^{-2}L^2\delta^{-1}}{t+2}\,\mathbb{E}\left[e_t^i\right]\\ &+ \left(1 - \frac{2(1-\delta)}{t+2}\right)u_t^i.\end{aligned}$$

and once again placing the upper bound on expected policy deviation due to Lemma 4.3,

$$\begin{aligned}u_{t+1}^i \leq& \frac{3\tau^{-2}K(1+\sigma^2) + 2\tau^{-2}(L+\tau)^2 + 28K\tau^{-4}L^2\delta^{-1}\sigma^2 + 28\tau^{-2}L^2\delta^{-1}}{(t+2)^2}\\ &+ \frac{2\tau^{-2}L^2\delta^{-1}}{N(t+2)} + \left(1 - \frac{2(1-\delta)}{t+2}\right)u_t^i.\end{aligned}$$

Another invocation of Lemma C.2 concludes the proof, choosing $\delta = 1/4$. □

This convergence result is stated in terms of exploitability of the unregularized game, leading to an additional $O(\tau)$ term. However, in many cases, the NE of the regularized game itself is of interest, in which case the upper bounds should read $O(\frac{\tau^{-2}}{\sqrt{T}} + \frac{\tau^{-1}}{\sqrt{N}})$ and $O(\frac{\tau^{-3/2}\lambda^{-1/2}}{\sqrt{T}} + \frac{\tau^{-1/2}\lambda^{-1/2}}{\sqrt{N}})$ for the monotone and strongly monotone cases respectively.

In the choice of learning rate η_t above, no intrinsic problem parameter is assumed to be known. Furthermore, due to (1) the regularization τ and (2) a finite population, a non-vanishing exploitability of $\mathcal{O}(\tau + \tau^{-1}/\sqrt{N})$ will be induced in terms of the NE in the monotone case. While Theorem 3.3 readily suggested a bias of order $\mathcal{O}(1/\sqrt{N})$ is fundamental, when learning is conducted with finitely many agents Theorem 4.4 shows this is amplified to $\mathcal{O}(\tau^{-1}/\sqrt{N})$. Since for finite population SMFG, there will always be a non-vanishing exploitability in terms of NE due to the mean-field approximation, in practice τ could be chosen to incorporate an acceptable bias level. Alternatively, if the exact value of the number of players N is known by each agent, one could choose τ optimally, to obtain the following corollary.

Corollary 4.5 (Optimal τ, full feedback). *Assume the conditions of Theorem 4.4. For monotone* $\mathbf{F}$*, choosing regularization parameter* $\tau = 1/\sqrt[4]{N}$ *yields* $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j_T\}_{j=1}^N)] \leq \mathcal{O}(\frac{\sqrt{N}}{\sqrt{T}} + \frac{1}{\sqrt[4]{N}})$ *for all* $i \in \mathcal{N}$*. For* λ*-strongly monotone* $\mathbf{F}$*, choosing* $\tau = 1/\sqrt[4]{N}$ *yields* $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j_T\}_{j=1}^N)] \leq \mathcal{O}(\frac{\lambda^{-1/2}\sqrt{N}}{\sqrt{T}} + \frac{\lambda^{-1/2}}{\sqrt[3]{N}})$.

Even though TRPA-Full solves the regularized (hence strongly monotone) problem, compared with the $\mathcal{O}(1/T)$ rate in classical strongly monotone VI [33] or strongly convex optimization [56], our worse $\mathcal{O}(1/\sqrt{T})$ time dependence is due to independent learning. Intuitively, additional time is required to ensure the policies of independent learners are sufficiently close when "collectively" evaluating $\mathbf{F}$. The additional dependence of the time-vanishing term on $\sqrt{N}$ is also a result of this fact. Furthermore, when learning itself is performed by N agents, we note that the bias as a function of N decreases with $\mathcal{O}(1/\sqrt{N})$ (or $\mathcal{O}(1/\sqrt[4]{N})$ for strongly monotone problems), and not with $\mathcal{O}(1/\sqrt{N})$ as Theorem 3.2 might suggest. We leave the question of whether this gap can be improved and whether knowledge of N is required in Corollary 4.5, as future work.

5 Convergence in the Bandit Feedback Case

We now move on to the more challenging and realistic bandit feedback case, where agents can only observe the payoffs of the actions they have chosen. Once again, we analyze the IL setting (or in bandits terminology, the "no communications" setting) where agents cannot interact or coordinate with each other. One of the main challenges of bandit feedback with IL in our setting is that it is difficult for each agent to identify itself (i.e., assign itself a unique number between $1, \dots, N$) so that exploration of action payoffs can be performed in turns. For instance, in MMAB algorithms, this is typically achieved using variants of the so-called musical chairs algorithm [40], which is not available in our formulation. Instead, we adopt a *probabilistic* exploration scheme where each agent probabilistically decides it is its turn to explore payoffs while the rest of the agents induce the required empirical population distribution on which $\mathbf{F}$ should be evaluated.

Our algorithm, which we call TRPA-Bandit, is straightforward and relies on exploration occurring over epochs, where policies are updated once in between epochs using the estimate of action payoffs constructed during the exploration phase. We use the subscript h to index epochs, which consist of T_h repeated plays indexed by (h, t) for $t = 1, \dots, T_h$. While we formally presented TRPA-Bandit (Algorithm 2), the procedure informally is as follows for each agent, fixing an exploration parameter $\varepsilon \in (0, 1)$ and an agent $i \in \mathcal{N}$:

(1) At each epoch h, for $T_h > 0$ time steps, repeat the following:
   (a) With probability ε, sample uniformly an action $a^i_{h,t}$, observe the payoff $r^i_{h,t}$, and keep the importance sampling estimate $\widehat{\mathbf{r}}^i_h \leftarrow K r^i_{h,t} \mathbf{e}_{a^i_{h,t}}$.
   (b) Otherwise (with probability $1 - \varepsilon$), sample action according to current policy $\boldsymbol{\pi}^i_h$.
(2) Update the policy using TRPA, $\boldsymbol{\pi}^i_{h+1} = \Pi_{\Delta_{\mathcal{A}}}((1 - \tau\eta_h)\boldsymbol{\pi}^i_h + \eta_h \widehat{\mathbf{r}}^i_h)$. If the agent did not explore this epoch, use $\widehat{\mathbf{r}}^i_h = \mathbf{0}$.

ALGORITHM 2: TRPA-Bandit: IL with Bandit Feedback Algorithm for Each Agent $i \in \mathcal{N}$.

Require: Number of actions K, regularization $\tau > 0$, exploration probability $\varepsilon > 0$, number of epochs H, epoch lengths $\{T_h\}_h$, learning rates $\{\eta_h\}_h$
  $\boldsymbol{\pi}_0^i \leftarrow \frac{1}{K}\mathbf{1}$
  for $h = 0, \dots, H-1$ **do**
    $\widehat{\mathbf{r}}_h^i \leftarrow \mathbf{0}$
    for $t = 1, \dots, T_h$ **do** ▷ Exploration for T_h rounds before policy update,
      Sample Bernoulli r.v. $X_{h,t}^i \sim \text{Ber}(\varepsilon)$.
      if $X_{h,t}^i = 1$ **then**
        Play action $a_{h,t}^i \sim \text{Unif}(\mathcal{A})$ uniformly at random. ▷ Explore with prob. ε,
        Observe payoff $r_{h,t}^i$, set $\widehat{\mathbf{r}}_h^i \leftarrow K r_{h,t}^i \mathbf{e}_{a_{h,t}^i}$.
      else if $X_{h,t}^i = 0$ **then**
        Play action with current policy $a_{h,t}^i \sim \boldsymbol{\pi}_h^i$. ▷ Else, play the current policy.
      end if
    end for
    $\boldsymbol{\pi}_{h+1}^i = \Pi_{\Delta_{\mathcal{A}}}((1-\tau\eta_h)\boldsymbol{\pi}_h^i + \eta_h \widehat{\mathbf{r}}_h^i)$ ▷ After each epoch, update policy.
  end for
  Return $\boldsymbol{\pi}_H^i$

Intuitively, the probabilistic sampling scheme allows some agents to build a low-variance estimate of $\mathbf{F}$, while others simply sample actions with their current policy in order to induce the empirical population distribution at which $\mathbf{F}$ should be evaluated.

Similar to the full feedback setting, we introduce useful notation used throughout this chapter. We define the sigma algebra $\mathcal{F}_h := \mathcal{F}(\{\boldsymbol{\pi}_{h'}^i\}_{h'=0,\dots,h}^{i=1,\dots,N})$. Adapting the notation from Section 1.1 to the case with multiple epochs, we use

$$\widehat{\mu}_{h,t} := \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_{a_{h,t}^i}, \qquad \mathbf{r}_{h,t}^i := \mathbf{F}(\widehat{\mu}_{h,t}) + \mathbf{n}_{h,t}^i,$$

where the updated time indices h, t simply refer to the tth round of play in epoch h, and $a_{h,t}^i$ is the action played by player i at epoch h, round t. Under the dynamics of Algorithm 2, we define the following random variables to assist our proofs:

$$\bar{\mu}_h := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}_h^i, \qquad e_t^i := \|\boldsymbol{\pi}_h^i - \bar{\mu}_h\|_2^2, \qquad u_h^i := \mathbb{E}\left[\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2\right],$$

which correspond to the mean policy at epoch h, the policy deviation of agent i from the mean at epoch h and the ℓ_2 distance from the MF-NE. Note that since policies are updated only in between epochs, the above quantities are indexed by epochs h rather than rounds h, t.

Our analysis follows the ideas in the case of expert feedback, the main difference being randomization due to the exploration probabilities and the errors being analyzed per epoch rather than per round. Similar to the full feedback setting, we will proceed in several steps expressed as intermediate lemmas: (1) we bound the added bias and variance due to the importance sampling strategy, (2) we obtain a non-linear recursion for the expectation of the terms e_t^i and possible sampling bias, (3) we bound the expected differences of each agent's action probabilities e_t^i, showing the deviation of the policies of the agents goes to zero in expectation, (4) we solve the recursion to obtain the convergence rate.

The next result, Lemma 5.1, provides an answer to the first step. We show that despite the probabilistic exploration step, the estimates $\widehat{\mathbf{r}}_h^i$ in Algorithm 2 have low bias and variance.

Lemma 5.1 (Exploration bias). *Under the dynamics of TRPA-Bandit, it holds almost surely for each epoch $h \geq 0$ that*

$$\| \mathbb{E}[\widehat{\mathbf{r}}_h^i | \mathcal{F}_h] - \mathbf{F}(\varepsilon \frac{1}{K}\mathbf{1} + (1-\varepsilon)\bar{\mu}_h)\|_2 \leq K^{3/2}\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\} + \frac{2L}{N} + \frac{2L}{\sqrt{N}}.$$

The full proof has been postponed to Appendix D.4. In summary, the proof strategy is to decompose and analyze the bias due to the possibility of no exploration round happening (the term $K^{3/2}\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\}$), the impact of the exploring agent on payoffs (the term $\frac{2L}{N}$), and bias due to the finitely many agents, similar to Theorem 3.2 (the term $\frac{2L}{\sqrt{N}}$). The additional bias due to probabilistic exploration originates from the possibility that no exploration occurs in a given epoch: the probability of this event can be bounded by $\exp\{-\varepsilon T_h\}$.

Lemma 5.1 shows that even if the players do not have full feedback, they can obtain low-bias, low-variance estimates of $\mathbf{F}(\varepsilon \frac{1}{K}\mathbf{1} + (1-\varepsilon)\bar{\mu}_h) \approx \mathbf{F}(\bar{\mu}_h)$ when ε is small. It guarantees that even if the agents do not explore each epoch, in expectation the probabilistic exploration scheme yields a low bias if the epoch lengths T_h are logarithmically large: hence, full feedback can be simulated by paying a logarithmic cost. Therefore, in our epoched exploration scheme, the bias in "querying" the payoff operator $\mathbf{F}$ due to an exploring population can be controlled by tuning ε and T_h.

We next state the error recursion in Lemma 5.2, which uses the result of Lemma 5.1 to construct the main recurrence for the bandit feedback case.

Lemma 5.2 (Main recurrence for TRPA-Bandit). *Under TRPA-Bandit dynamics, it holds for any $i \in \{1, \dots, N\}$ and each epoch $h \geq 0$ that*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{h+1}^i - \boldsymbol{\pi}^*\|_2^2\right] \leq & 4\eta_h^2 K^3(1+\sigma^2) + 8\eta_h^2(L+\tau)^2 + 8K^{3/2}\eta_h\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\} \\ & + 128\eta_h\lambda^{-1}L^2N^{-1} + 16\eta_h\lambda^{-1}L^2\varepsilon^2 + 2\eta_h\lambda^{-1}L^2\,\mathbb{E}\left[e_h^i\right] \\ & + (1 - 2\eta_h(\lambda/2 + \tau))\,\mathbb{E}\left[\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2\right],\end{aligned}$$

for strongly monotone $\lambda > 0$ payoffs, and

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{h+1}^i - \boldsymbol{\pi}^*\|_2^2\right] \leq & 4\eta_h^2 K^3(\sigma^2+1) + 8\eta_h^2(L+\tau)^2 + 8K^{3/2}\eta_h\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\} \\ & + 64\tau^{-1}\eta_h\delta^{-1}L^2N^{-1} + 8\tau^{-1}\eta_h\delta^{-1}L^2\varepsilon^2 + \tau^{-1}\eta_h\delta^{-1}L^2\,\mathbb{E}\left[e_h^i\right] \\ & + (1 - 2\tau\eta_h(1-\delta))\,\mathbb{E}\left[\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2\right],\end{aligned}$$

for monotone payoffs for arbitrary $\delta \in (0,1)$.

Once again, the full proof has been postponed to Appendix D.5. The proof of Lemma 5.2 follows a similar path as in the recurrence in the full feedback case (Lemma 4.2), with the exception that $\mathbf{r}_{h,t}^i$ has been replaced by the importance sampling estimator $\widehat{\mathbf{r}}_h^i$. In the decomposition due to Inequality (4), the analysis of term (a) remains the same, whereas the terms (b), (c) must be further analyzed using Lemma 5.1 to account for deviations between $\mathbf{r}_{h,t}^i$ and $\widehat{\mathbf{r}}_h^i$, as well as the ε fraction of the population now exploring each round.

The recurrence in Lemma 5.2 is similar in form to the full feedback case (Lemma 4.2), apart from the term $8K^{3/2}\eta_h\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\}$ due to the exploration scheme. However, keeping the exploration epoch lengths T_h logarithmically large can make this term small. Furthermore, once again the recursion produces a dependence on expected mean policy deviations, $\mathbb{E}[e_h^i]$. Hence, the expected policy deviation $\mathbb{E}[e_h^i]$ at epoch h must be bounded once again at a rate $O(1/h)$ in order

to obtain a non-vacuous upper bound on exploitability. As in the full feedback case, we employ regularization to ensure $\mathbb{E}[e_h^i]$ is small. The next lemma presents our upper bound.

Lemma 5.3 (Policy deviation under TRPA-Bandit). *Under TRPA-Bandit dynamics, with learning rates $\eta_h := \frac{\tau^{-1}}{h+2}$, arbitrary exploration rate $\varepsilon > 0$ and epoch lengths $T_h := \lceil \varepsilon^{-1} \log(h+2) \rceil$ it holds for any $i, j \in \{1, \dots, N\}, i \neq j$ and each epoch $h \geq 0$ that*

$$\mathbb{E}[e_h^i] \leq \frac{24\tau^{-2}K^3(\sigma^2+2) + 48\tau^{-2} + 24}{h+1} + \frac{16\tau^{-2}L^2}{N^2}.$$

The proof of Lemma 5.3 follows similar ideas to Lemma 4.3, while accounting for (a) the increased variance due to importance sampling, and (b) potential further deviation between agent policies due to the $O(1/N)$ impact of exploration on $\widehat{\mu}_{h,t}$. In particular, an additional source of policy deviation occurs when an agent does not explore in a given epoch, in which case the payoff estimator is uninformative ($\widetilde{\mathbf{r}}_h^i = \mathbf{0}$) causing additional policy deviation. The full proof has been postponed to Appendix D.6.

In the case of TRPA-Bandit, due to the additional variance of probabilistic exploration, the policy deviations between agents might be larger: compare the upper bounds of Lemma 5.3 and Lemma 4.3. In particular, the upper bound of Lemma 5.3 contains a non-vanishing term unlike Lemma 4.3. Nevertheless, they can still be controlled of order $O(1/(h+1) + 1/N^2)$, where the additional $O(1/N^2)$ term compared with TRPA-Full vanishes very quickly when N is large.

With these intermediate lemmas established, we state and prove the main convergence result for TRPA-Bandit Theorem 5.4, the main result of this work. We provide asymptotic rates for brevity, although the proof of the theorem provides explicit bounds.

Theorem 5.4 (Convergence, bandit feedback). *Assume $\mathbf{F}$ is Lipschitz, monotone. Assume N agents run TRPA-Bandit (Algorithm 2) for T time steps with regularization $\tau > 0$ and exploration parameter $\varepsilon > 0$, and agents return policies $\{\boldsymbol{\pi}^i\}_i$ after executing Algorithm 2. Then, for any agent $i \in \mathcal{N}$ that $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)] \leq \widetilde{O}(\frac{\tau^{-2}\varepsilon^{-1/2}}{\sqrt{T}} + \tau^{-1}\varepsilon + \tau + \frac{\tau^{-1}}{\sqrt{N}} + \frac{\tau^{-3/2}}{N})$. If $\mathbf{F}$ is λ-strongly monotone, then $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)] \leq \widetilde{O}(\frac{\tau^{-3/2}\lambda^{-1/2}\varepsilon^{-1/2}}{\sqrt{T}} + \tau^{-1/2}\lambda^{-1/2}\varepsilon + \tau + \frac{\tau^{-1/2}\lambda^{-1/2}}{\sqrt{N}} + \frac{\tau^{-1}\lambda^{-1/2}}{N})$ for all $i \in \mathcal{N}$.*

The proof follows from a straightforward combination of Lemmas 5.1, 5.2, and 5.3 as in the full feedback case. The exact bounds and proof are in the appendix, Section D.7.

Once again, the values of τ and exploration probability ε can be chosen to incorporate tolerable exploitability. In the case where the number of participants N in the game is known, the following corollary indicates the asymptotically optimal choices for the hyperparameters.

Corollary 5.5 (Optimal ε, τ, bandit feedback). *Assume the conditions of Theorem 5.4 for N agents running TRPA-Bandit. For monotone $\mathbf{F}$, choosing $\tau = 1/\sqrt[4]{N}$ and $\varepsilon = 1/\sqrt{N}$ yields $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)] \leq \widetilde{O}(\frac{N^{3/4}}{\sqrt{T}} + \frac{1}{\sqrt[4]{N}})$ for all $i \in \mathcal{N}$. For λ-strongly monotone $\mathbf{F}$, choosing $\tau = 1/\sqrt[3]{N}$ and $\varepsilon = 1/\sqrt{N}$ yields $\mathbb{E}[\mathcal{E}^i_{exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)] \leq \widetilde{O}(\frac{N^{3/4}\lambda^{-1/2}}{\sqrt{T}} + \frac{\lambda^{-1/2}}{\sqrt[3]{N}})$ for all $i \in \mathcal{N}$.*

The dependence of N of the sample complexity in the bandit case is worse compared with the full feedback setting as expected: intuitively the agents must take turns to estimate the payoffs of each action in bandit feedback. Furthermore, while our problem framework is different and a direct comparison is difficult in terms of bounds, we point out that classical MMAB results such as [40] have a linear dependence on N, while in our case the dependence on N scales with $N^{3/4}$. We emphasize that the time-dependence is sublinear in terms of N, up to the non-vanishing finite population bias. As in the full feedback case, the non-vanishing finite population bias in the bandit feedback case scales with $O(1/\sqrt[4]{N})$ or $O(1/\sqrt[3]{N})$, rather than $O(1/\sqrt{N})$ which would match Theorem 3.2.

Note that the dependence of the bias on N varies in various mean-field game results [60], but asymptotically is known to converge to zero as $N \to \infty$, as our explicit finite-agent bound also demonstrates.

Finally, we note that as expected the algorithm for bandit feedback has a worse dependency on the number of actions. This is as expected due to the fact that (i) the importance sampling estimator increases variance on payoff estimators by a factor of K, and (2) in other words, a factor of $O(K)$ is introduced in order to explore all actions.

6 Experimental Results

We validate the theoretical results of our work on numerical and two real-world experiments. The details of the setup are presented in greater detail in the appendix (Section E) and the code is provided in supplementary materials. Our experiments assume bandit feedback, although in the appendix we also include results under the full feedback model. We provide an overview of our setup first.

Numerical problems. Firstly, we formulate three numerical problems, which are based on examples suggested in Section 2. We randomly generate monotone payoff operators that are linear (LINEAR) and a payoff function that admits a KL divergence potential (KL). We also analyze a particular "beach bar process" (BB), a stateless version of the example presented in [53]. For numerical examples, we use $K = 5$, and vary between various population sizes in $N = \{20, 50, 100, 200, 500, 1000\}$ to quantify the effect of finite N and compare with theoretical bias bounds. We set the parameters ε, τ using the theoretical values from Corollary 5.5.

Traffic flow simulation. Using the UTD19 dataset [39], we evaluate our algorithms on traffic congestion data on three different routes for accessing the city center of Zurich (UTD). The UTD19 dataset features many closed loop sensors across various urban roads in Zurich, providing granular measurements of road occupancy and traffic flow. We use these real-world stationary detector measurements to approximate traveling times as a function of route occupancy with a kernelized regression model on three routes. We then evaluate our algorithms on estimated traveling times. The real-world data from the UTD19 dataset suggests the travel duration (and therefore the payoffs) are in fact non-increasing in the occupancy of each segment, and hence can be reasonably modeled by a monotone payoff. Further empirical evaluation of monotonicity is discussed in Appendix E.5.

Access to the Tor network. We also run experiments on accessing the Tor network, which is a large decentralized network for secure communications and an active area of research in computer security [29, 42]. The Tor network consists of many decentralized servers and access to the network is achieved by communicating with one of many entry guard servers (a list advertised publicly [67], some hosted by universities). As the entry guards serve a wide public of users and each user is free to choose an entry point, the network is an ideal setting to test our algorithms. We simulate 100 independent agents accessing Tor by choosing among $K = 5$ entry servers every minute and use the real-world ping delays for bandit feedback. Our empirical evaluations show that monotonicity is also a reasonable assumption for the Tor network experiment due to the fact that response times tend to be non-decreasing in occupation: we provide an empirical argument in Appendix E.6.

Overall, our experiments in both models and real-world use cases support our theory. Firstly, our experiments verify learning in the sense of decreasing maximum exploitability $\max_{i \in \mathcal{N}} \mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)$ and mean distance to MF-NE given by $\frac{1}{N}\sum_i \|\boldsymbol{\pi}^* - \boldsymbol{\pi}^i_t\|_2$ during IL (Figure 1-b,c). As expected from Corollary 5.5, the maximum exploitability oscillates but remains bounded (due to the effect of finite N). Furthermore, the agents converge to policies closer to the MF-NE as N increases (Figure 1-c), empirically demonstrating that the MF-NE is an accurate description of the limiting behavior of SMFG as $N \to \infty$. The scale of the excess exploitability in

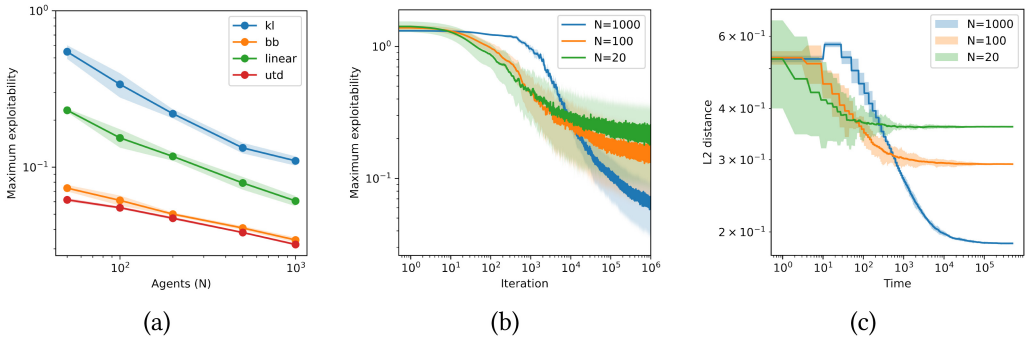

Fig. 1. Results for numerical problems KL, BB, LINEAR, UTD. (a) Maximum exploitability of N agents at convergence as a function of N for different problems, (b) The max. exploitability among N agents during training with linear payoff (LINEAR), for different N, (c) The mean ℓ_2 distance of agent policies during training to the MF-NE in the Zurich traffic flow simulation problem (UTD).

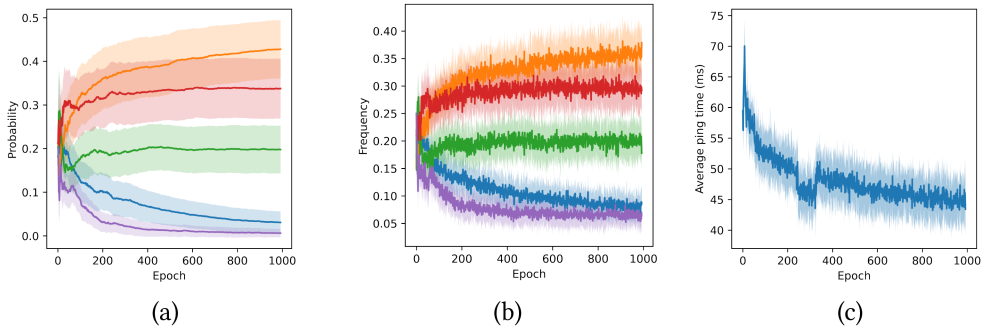

Fig. 2. Results for the Tor network experiment. (a) verage policies (probability distribution) over 5 servers of the 100 agents in the Tor network access experiment, (b) Empirical distribution of agents over Tor entry servers during training on 5 servers (different colors indicate different entry servers), (c) Average waiting times for Tor network access during training.

our experiments decreases rapidly in N, allowing our results to be significant even for hundreds or thousands of agents.

Our experimental results in traffic congestion and server access also support our theoretical claims. Despite having no *a priori* assumption of monotonicity, our methods efficiently converge. We note that in the case of the network access experiments, we refrain from simulating thousands of agents to minimize network impact. This implies that the delays have a high dependence on external factors as the behavior of other users will be dominant. Nevertheless, our experiment indicates that our algorithm can produce competitive results in the wild (Figure 2).

Finally, we note that while there are no algorithms that have theoretical guarantees in the N agent IL setting, we compare our results with a heuristic extension of online mirror descent [52] in the numerical problems, see Appendix E.4. Our results suggest a naive extension of OMD to our setting will lead to divergence or cycling.

7 Discussion and Conclusion

In this work, we proposed a finite-agent, independent learning algorithm for learning SMFGs under bandit feedback, and established finite sample guarantees. Our results hinge on the VI perspective of the mean-field game limit, demonstrating that efficient independent learning is achievable.

As the first study on decentralized learning on SMFGs, our results are based on the standard (strong) monotonicity assumption on VIs. However, research on VIs has identified several more general yet still tractable conditions that might be more applicable to real-world SMFGs, such as the weak Minty condition [13], the variationally stable condition [43], pseudo-monotone operators [31], and generalized monotonicity [33]. More abstractly, an intriguing open question is whether the existence of efficient algorithms for solving the (possibly non-monotone) MF-VI always implies the existence of decentralized algorithms for learning SMFG. At present, several challenges remain when generalizing our results beyond monotonicity: firstly, our current method requires the expected policy deviation to converge to zero, restricting the algorithmic tools that can be deployed. Moreover, our proof strategy relies on the convergence of iterates $\boldsymbol{\pi}_t^i$ to the unique regularized solution $\boldsymbol{\pi}^*$ due to the contractivity of the regularized ascent operator $\Gamma^{\eta,\tau}$. Yet for more general VIs, obtaining such a contractive operator in the ℓ_2-norm might be difficult.

From a game-theoretic and learning perspective, another relevant question is whether our VI approach to decentralized learning can be extended to Markovian MFGs. The current theory does not immediately suggest our analysis could be extended to Markovian games. Specifically, it is not clear if the Markovian NE would still be characterized compactly as a VI, or when the corresponding operator would be monotone. The question of IL in stationary MFGs has been tackled by [73] but with poor sample complexity; future work could address this problem using more general VIs. Finally, while our results focus on expected exploitability, they could potentially be generalized to guarantee convergence to low exploitability policies with high probability.

Acknowledgements

We thank the editor and reviewers for their detailed comments and insightful feedback.

References

[1] K. Ahn, P. Jain, Z. Ji, S. Kale, P. Netrapalli, and G. I. Shamir. 2022. Reproducibility in optimization: Theoretical framework and limits. *Advances in Neural Information Processing Systems*, 35 (2022), 18022–18033.

[2] P. Alatur, A. Barakat, and N. He. 2024. Independent policy mirror descent for markov potential games: Scaling to large number of players. In *The 63rd IEEE Conference on Decision and Control*, 2024.

[3] L. Ambühl, A. Loder, M. Menendez, and K. W. Axhausen. 2017. Empirical macroscopic fundamental diagrams: New insights from loop detector and floating car data. In *TRB 96th Annual Meeting Compendium of Papers*. Transportation Research Board, 17–03331.

[4] B. Anahtarci, C. D. Kariksiz, and N. Saldi. 2022. Q-learning in regularized mean-field games. *Dynamic Games and Applications*. 1–29.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. 2011. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications* 29, 4 (2011), 731–745.

[6] A. Aurell, R. Carmona, G. Dayanıklı, and M. Laurière. 2022. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications* 12, 1 (2022), 49–81.

[7] O. Avner and S. Mannor. 2014. Concurrent bandits and cognitive radio networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*. 66–81.

[8] A. Beznosikov, B. Polyak, E. Gorbunov, D. Kovalev, and A. Gasnikov. 2023. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. *European Mathematical Society Magazine* 127 (2023), 15–28.

[9] S. Bubeck, T. Budzinski, and M. Sellke. 2021. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. In *Conference on Learning Theory*. PMLR, 821–822.

[10] R. Carmona, G. Dayanıklı, and M. Laurière. 2022. Mean field models to regulate carbon emissions in electricity production. *Dynamic Games and Applications* 12, 3 (2022), 897–928.

[11] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. 2009. The complexity of computing a nash equilibrium. *Communications of the ACM* 52, 2 (2009), 89–97.

[12] C. Daskalakis, D. J. Foster, and N. Golowich. 2020. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 5527–5540.

[13] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. 2021. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2746–2754.
[14] D. Ding, C.-Y. Wei, K. Zhang, and M. Jovanovic. 2022. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *Proceedings of the 39th International Conference on Machine Learning*. 5166–5220.
[15] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. 2008. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*. 272–279.
[16] A. El Gamal and T. M. Cover. 1980. Multiple user information theory. *Proceedings of the IEEE* 68, 12 (1980), 1466–1483.
[17] F. Facchinei and J.-S. Pang. 2003. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.
[18] Z. Fu, Z. Yang, Y. Chen, and Z. Wang. 2019. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*, 2019.
[19] M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos, and O. Pietquin. 2022. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 489–497.
[20] N. Geroliminis and C. F. Daganzo. 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological* 42, 9 (2008), 759–770.
[21] D. A. Gomes and J. Saúde. 2021. A mean-field game approach to price formation. *Dynamic Games and Applications* 11, 1 (2021), 29–53.
[22] H. Gu, X. Guo, X. Wei, and R. Xu. 2024. Mean-field multiagent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*.
[23] R. Gummadi, R. Johari, S. Schmit, and J. Y. Yu. 2013. Mean field analysis of multi-armed bandit games. *Available at SSRN 2045842*, 2013.
[24] X. Guo, R. Xu, and T. Zariphopoulou. 2022. Entropy regularization for mean field games with learning. *Mathematics of Operations Research*, 2022.
[25] A. Heliou, J. Cohen, and P. Mertikopoulos. 2017. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*. 30.
[26] S. Hoogendoorn and V. Knoop. 2013. Traffic flow theory and modelling. *The Transport System and Transport Policy: An Introduction*. 125–159.
[27] M. Huang, R. P. Malhamé, and P. E. Caines. 2006. Large population stochastic dynamic games: Closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems* 6, 3 (2006), 221–252.
[28] Y. Huang and Q. Zhu. 2022. Game-theoretic frameworks for epidemic spreading and human decision-making: A review. *Dynamic Games and Applications* 12, 1 (2022), 7–48.
[29] R. Jansen, F. Tschorsch, A. Johnson, and B. Scheuermann. 2014. The sniper attack: Anonymously deanonymizing and disabling the tor network. *Technical report, Office of Naval Research Arlington VA*, 2014.
[30] A. Juditsky, A. Nemirovski, and C. Tauvel. 2011. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* 1, 1 (2011), 17–58.
[31] S. Karamardian. 1976. Complementarity problems over cones with monotone and pseudomonotone maps. *Journal of Optimization Theory and Applications* 18, 4 (1976), 445–454.
[32] G. M. Korpelevich. 1976. The extragradient method for finding saddle points and other problems. *Matecon* 12 (1976), 747–756.
[33] G. Kotsalis, G. Lan, and T. Li. 2022. Simple and optimal methods for stochastic variational inequalities, i: Operator extrapolation. *SIAM Journal on Optimization* 32, 3 (2022), 2041–2073.
[34] D. Kovalev, A. Beznosikov, A. Sadiev, M. Persiianov, P. Richtárik, and A. Gasnikov. 2022. Optimal algorithms for decentralized stochastic variational inequalities. *Advances in Neural Information Processing Systems* 35 (2022), 31073–31088.
[35] J.-M. Lasry and P.-L. Lions. 2007. Mean field games. *Japanese Journal of Mathematics* 2, 1 (2007), 229–260.
[36] M. Laurière, S. Perrin, J. Pérolat, S. Girgin, P. Muller, R. Élie, M. Geist, and O. Pietquin. 2024. Learning in mean field games: A survey. arXiv:2205.12944. Retrieved from https://arxiv.org/abs/2205.12944
[37] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. 2022. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022.
[38] T. Lin and M. I. Jordan. 2024. Perseus: A simple and optimal high-order method for variational inequalities. *Mathematical Programming*. 1–42.
[39] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen. 2019. Understanding traffic capacity of urban networks. *Scientific reports* 9, 1 (2019), 16283.

[40] G. Lugosi and A. Mehrabian. 2022. Multiplayer bandits without observing collision information. *Mathematics of Operations Research* 47, 2 (2022), 1247–1265.
[41] L. Matignon, G. J. Laurent, and N. Le Fort-Piat. 2007. Hysteretic q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 64–69.
[42] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker. 2008. Shining light in dark places: Understanding the tor network. In *Privacy Enhancing Technologies: 8th International Symposium, PETS 2008 Leuven, Belgium, July 23-25, 2008 Proceedings 8*. Springer, 63–76.
[43] P. Mertikopoulos and Z. Zhou. 2019. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* 173 (2019), 465–507.
[44] D. Monderer and L. S. Shapley. 1996. Potential games. *Games and Economic Behavior* 14, 1 (1996), 124–143.
[45] S. Mukherjee and M. Chakraborty. 2020. A decentralized algorithm for large scale min-max problems. In *The 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2967–2972.
[46] J. F. Nash. 1950. Non-cooperative games. 1950.
[47] A. Nemirovski. 2004. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization* 15, 1 (2004), 229–251.
[48] Y. Nesterov. 2007. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming* 109, 2-3 (2007), 319–344.
[49] A. Ozdaglar, M. O. Sayin, and K. Zhang. 2021. Independent learning in stochastic games. arXiv:2111.11743. Retrieved from https://arxiv.org/abs/2111.11743
[50] C. H. Papadimitriou. 1994. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and System Sciences* 48, 3 (1994), 498–532.
[51] J. Perolat, B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, et al. 2022. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science* 378, 6623 (2022), 990–996.
[52] J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin. 2022. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1028–1037.
[53] S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie, and O. Pietquin. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems* 33 (2020), 13199–13213.
[54] J. P. Pinasco, N. Saintier, and M. Kind. 2023. Learning, mean field approximations, and phase transitions in auction models. *Dynamic Games and Applications*. 1–32.
[55] N. Quijano, C. Ocampo-Martinez, J. Barreiro-Gomez, G. Obando, A. Pantoja, and E. Mojica-Nava. 2017. The role of population games and evolutionary dynamics in distributed control systems: The advantages of evolutionary game theory. *IEEE Control Systems Magazine* 37, 1 (2017), 70–97.
[56] A. Rakhlin, O. Shamir, and K. Sridharan. 2011. Making gradient descent optimal for strongly convex stochastic optimization. arXiv:1109.5647. Retrieved from https://arxiv.org/abs/1109.5647
[57] R. T. Rockafellar. 1976. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* 14, 5 (1976), 877–898.
[58] J. Rosenski, O. Shamir, and L. Szlak. 2016. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*. PMLR, 155–163.
[59] R. W. Rosenthal. 1973. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory* 2, 1 (1973), 65–67.
[60] N. Saldi, T. Başar, and M. Raginsky. 2019. Approximate nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research* 44, 3 (2019), 1006–1033.
[61] W. H. Sandholm. 2015. Population games and deterministic evolutionary dynamics. In *Handbook of Game Theory with Economic Applications*. Elsevier, 4 (2015), 703–778.
[62] M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. 2021. Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems* 34 (2021), 18320–18334.
[63] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv:1712.01815. Retrieved from https://arxiv.org/abs/1712.01815
[64] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362, 6419 (2018), 1140–1144.
[65] J. K. Skwirzynski. 1981. *New concepts in multi-user communication. Number 43. Springer Science & Business Media*, 1981.

[66] K. Srivastava, A. Nedić, and D. Stipanović. 2011. Distributed min-max optimization in networks. In *2011 17th International Conference on Digital Signal Processing (DSP)*, 1–8. IEEE, 2011.
[67] The Tor Network. Tor metrics: Relay search. Retrieved from https://metrics.torproject.org/rs.html#search/flag:Guard, 2024. Accessed: 2024-05-01.
[68] Tor Metrics. Tor metrics - users. Retrieved from https://metrics.torproject.org/userstats-relay-country.html, 2024. Accessed: 2024-05-01.
[69] L. Wang, Z. Yang, and Z. Wang. 2020. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*. PMLR, 10092–10103.
[70] X. Wang and R. Jia. 2021. Mean field equilibrium in multi-armed bandit game with continuous reward. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence IJCAI-21*. 3118–3124.
[71] Q. Xie, Z. Yang, Z. Wang, and A. Minca. 2021. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*. PMLR, 11436–11447.
[72] J. Yang, X. Ye, R. Trivedi, H. Xu, and H. Zha. 2018. Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*, 2018.
[73] B. Yardim, S. Cayci, M. Geist, and N. He. 2023. Policy mirror ascent for efficient and independent learning in mean field games. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 202 (2023), 39722–39754.
[74] B. Yardim, A. Goldman, and N. He. 2024. When is mean-field reinforcement learning tractable and relevant? In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2038–2046.
[75] B. Yongacoglu, G. Arslan, and S. Yüksel. 2022. Independent learning in mean-field games: Satisficing paths and convergence to subjective equilibria. arXiv:2209.05703. Retrieved from https://arxiv.org/abs/2209.05703
[76] M. A. U. Zaman, A. Koppel, S. Bhatt, and T. Basar. 2023. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 10178–10206.
[77] F. Zhang, V. Tan, Z. Wang, and Z. Yang. 2024. Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*. 36.
[78] L. Zhang, J. Yang, A. Karbasi, and N. He. 2024. Optimal guarantees for algorithmic reproducibility and gradient complexity in convex optimization. *Advances in Neural Information Processing Systems*. 36.

Appendices

A Table for Notation

General notation:	
$\Delta_{\mathcal{X}}$	probability simplex on discrete set $\mathcal{X}$
$[M]$	$:= \{1, \dots, M\}$, for any $M \in \mathbb{N}_{>0}$
$\Delta_{\mathcal{X},N}$	$:= \{\mathbf{u} = \{u_i\}_i \in \Delta_{\mathcal{X}} \mid Nu_i \in \mathbb{N}_{\geq 0}\}$, for discrete set $\mathcal{X}$, $N \in \mathbb{N}_{>0}$
$\mathbb{M}^{D_1,D_2}$	$D_1 \times D_2$ matrices
$\mathbb{S}^D_{++}$	positive definite $D \times D$ matrices
Π_K	projection onto convex, compact set $K \subset \mathbb{R}^H$
$\mathbf{e}_a$	$\in \mathbb{R}^{\mathcal{A}}$, standard unit vector with coordinate a set to 1
For SMFGs:	
$\mathbf{F}$	payoff function, $\mathbf{F} : \Delta_{\mathcal{A}} \to [0,1]^K$
λ	strong monotonicity modulus of $\mathbf{F}$
L	Lipschitz modulus of $\mathbf{F}$
τ	Tikhonov regularization parameter
$\boldsymbol{\pi}^*$	unique solution of τ regularized (MF-RVI)
K	number of actions
N	number of players
$\mathcal{N}$	$:= \{1, \dots, N\}$, the set of players
$\mathcal{A}$	the set of actions, $\lvert\mathcal{A}\rvert = K$.
σ^2	upper bound of the standard deviation of received payoff
$\widehat{\mu}$	$:= \frac{1}{N}\sum_{i=1}^N \mathbf{e}_{a^i}$, when actions $\{a^i\}_{i=1}^N$ clear in context
$V^i(\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^N)$	expected payoff of player i under strategy profile $(\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^N)$

$\mathcal{E}^i_{\exp}(\{\boldsymbol{\pi}^j\}_{j=1}^N)$	$:= \max_{\boldsymbol{\pi}' \in \Delta_{\mathcal{A}}} V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) - V^i(\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^N)$, exploitability
$\widehat{\mu}(\{a^i\}_{i=1}^N)$	$:= \frac{1}{N}\sum_{i=1}^N \mathbf{e}_{a^i}$, the action distribution induced by particular $\{a^i\}_{i=1}^N \in \mathcal{A}^N$

For full feedback:

t	round of play
a^i_t	$\in \mathcal{A}$ action take by player i at round t
$\widehat{\mu}_t$	$:= \frac{1}{N}\sum_{i=1}^N \mathbf{e}_{a^i_t}$, empirical distribution over actions $\mathcal{A}$ on round t
$\mathbf{r}^i_t$	$:= \mathbf{F}(\widehat{\mu}_t) + \mathbf{n}^i_t$, payoff vector observed by player i
η_t	learning rate
$\bar{\mu}_t$	$:= \frac{1}{N}\sum_{i=1}^N \boldsymbol{\pi}^i_t$, mean policy at round t
e^i_t	$:= \|\boldsymbol{\pi}^i_t - \bar{\mu}_t\|_2^2$, deviation of policy of player i
u^i_t	$:= \mathbb{E}[\|\boldsymbol{\pi}^i_t - \boldsymbol{\pi}^*\|_2^2]$

For bandit feedback:

h	epoch of play
t	round of play in epoch
$a^i_{h,t}$	$\in \mathcal{A}$ action take by player i at round t of epoch h
T_h	number of rounds in epoch h
ε	exploration probability
$\widehat{\mu}_{h,t}$	$:= \frac{1}{N}\sum_{i=1}^N \mathbf{e}_{a^i_{h,t}}$, empirical distribution over actions $\mathcal{A}$ on round t of epoch h
$\mathbf{r}^i_{h,t}$	$:= \mathbf{F}(\widehat{\mu}_{h,t}) + \mathbf{n}^i_{h,t}$, (unobserved) payoff vector by player i
$r^i_{h,t}$	$:= \mathbf{r}^i_{h,t}(a^i_{h,t})$, payoff observed by player i at round t of epoch h
$X^i_{h,t}$	$\sim \mathrm{Ber}(\varepsilon)$ indicator variable, 1 if i explores at round t of epoch h
$\widehat{\mathbf{r}}^i_h$	$:= K r^i_{h,t} \mathbf{e}_{a^i_{h,t}}$ if $X^i_{h,t} = 1$, importance sampling estimate of player i
η_h	learning rate
$\bar{\mu}_h$	$:= \frac{1}{N}\sum_{i=1}^N \boldsymbol{\pi}^i_h$, mean policy at epoch h
e^i_h	$:= \|\boldsymbol{\pi}^i_h - \bar{\mu}_h\|_2^2$, deviation of policy of player i
u^i_h	$:= \mathbb{E}[\|\boldsymbol{\pi}^i_h - \boldsymbol{\pi}^*\|_2^2]$

B Formalizing Learning Algorithms

In this section, we formalize the concept of an independent learning algorithm in the full feedback and bandit feedback setting. In general, we formalize the notion of an algorithm as a map $A^i_t : \mathcal{H}^i_t \to \Delta_{\mathcal{A}}$ that maps the set of past observations of agent i at time t to action selection probabilities. The definition of the set $\mathcal{H}^i_t$ varies between the feedback models.

Definition B.1 (Learning algorithm with full feedback). An independent learning algorithm with full information $\mathbf{A} = \{A^i_t\}_{i,t}$ is a sequence of mappings for each player with

$$A^i_t : \Delta^{t-1}_{\mathcal{A}} \times \mathcal{A}^{t-1} \times [0,1]^{(t-1)\times K} \to \Delta_{\mathcal{A}}, \text{ for all } t > 0,$$
$$A^i_0 \in \Delta_{\mathcal{A}},$$

that maps past $t-1$ observations from previous rounds to a mixed strategy on actions $\mathcal{A}$ at time $t > 0$ for each agent i.

Naturally, we are interested in algorithms that yield a good approximation of the NE in expectation. More explicitly, we will be interested in designing algorithms that converge to policy profiles with low expected explotability.

Definition B.2 (Rational learning algorithm with full feedback). Let $\mathbf{A}$ be an algorithm with full feedback as defined in Definition B.1. We call $\mathbf{A}$ δ-rational if it holds that for all i, the induced mixed strategies $\boldsymbol{\pi}_t^i$ under $\boldsymbol{\pi}_0^i = A_0^i, \boldsymbol{\pi}_t^i = A_t^i(\boldsymbol{\pi}_0^i, \dots, \boldsymbol{\pi}_{t-1}^i, a_0^i, \dots, a_{t-1}^i, \mathbf{r}_0^i, \dots, \mathbf{r}_{t-1}^i)$ satisfy

$$\lim_{t\to\infty} \mathbb{E}[\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}_t^j\}_{j=1}^N)] \leq \delta, \text{ for all } i \in \mathcal{N}.$$

Note that while not specified in the definition above, we will also be interested in the rate of convergence of the exploitability term for a rational algorithm. Since we are solving the SMFG at the finite-agent regime, we will be interested in δ-rational algorithms that have $\delta \to 0$ as $N \to \infty$, that is, the non-vanishing bias should scale inversely with the number of agents.

Finally, we also formalize the concepts of a learning algorithm and δ-rationality in the bandit setting.

Definition B.3 (Algorithm with bandit feedback). An algorithm with bandit feedback $\mathbf{A} = \{A_t^i\}_{i,t}$ is a sequence of mappings for each player with

$$A_t^i : \Delta_{\mathcal{A}}^{t-1} \times \mathcal{A}^{t-1} \times [0,1]^{(t-1)} \to \Delta_{\mathcal{A}}, \text{ for all } t > 0,$$
$$A_0^i \in \Delta_{\mathcal{A}},$$

that maps past $t-1$ observations from previous rounds at all times $t > 0$ (only including the payoffs of the *played* actions) to a probability distribution on actions $\mathcal{A}$.

Definition B.4 (Rational algorithm with bandit feedback). Let $\mathbf{A}$ be an algorithm with bandit feedback as defined in Definition B.3. We call $\mathbf{A}$ δ-rational if it holds that for all i, the induced (random) mixed strategies $\boldsymbol{\pi}_t^i$ under $\boldsymbol{\pi}_0^i = A_0^i, \boldsymbol{\pi}_t^i = A_t^i(\boldsymbol{\pi}_0^i, \dots, \boldsymbol{\pi}_{t-1}^i, a_0^i, \dots, a_{t-1}^i, r_0^i, \dots, r_{t-1}^i)$ satisfy

$$\lim_{t\to\infty} \mathbb{E}[\mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}_t^j\}_{j=1}^N)] \leq \delta, \text{ for all } i \in \mathcal{N}.$$

C Basic Inequalities

In our proofs, we will need to repeatedly bound certain recurrences and sums. In this section, we present useful inequalities to this end.

Lemma C.1 (Harmonic partial sum bound). *For any integers $s, \bar{s}$, constant $c \in \mathbb{R}$ such that $1 \leq \bar{s} < s$, $p \neq -1$, and $a \geq 0$, it holds that*

$$\log(s+a) - \log(\bar{s}+a) + \frac{1}{s+a} \leq \sum_{n=\bar{s}}^{s} \frac{1}{n+a} \leq \frac{1}{\bar{s}+a} + \log(s+a) - \log(\bar{s}+a),$$

$$\frac{(s+a)^{p+1}}{p+1} - \frac{(\bar{s}+a)^{p+1}}{(p+1)} + (\bar{s}+a)^p \leq \sum_{n=\bar{s}}^{s} (n+a)^p \leq \frac{(s+a)^{p+1}}{p+1} - \frac{(\bar{s}+a)^{p+1}}{p+1} + (s+a)^p, \text{ if } p \geq 0$$

$$\frac{(s+a)^{p+1}}{p+1} - \frac{(\bar{s}+a)^{p+1}}{p+1} + (s+a)^p \leq \sum_{n=\bar{s}}^{s} (n+a)^p \leq \frac{(s+a)^{p+1}}{p+1} - \frac{(\bar{s}+a)^{p+1}}{p+1} + (\bar{s}+a)^p, \text{ if } p \leq 0$$

Proof. Let $f_1 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a non-decreasing positive function and $f_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a non-increasing positive function. Then, it holds that

$$\int_{x=\bar{s}}^{s} f_1(x)dx + f_1(\bar{s}) \leq \sum_{n=\bar{s}}^{s} f_1(n) \leq \int_{x=\bar{s}}^{s} f_1(x)dx + f_1(s),$$

$$\int_{x=\bar{s}}^{s} f_2(x)dx + f_2(s) \leq \sum_{n=\bar{s}}^{s} f_2(n) \leq \int_{x=\bar{s}}^{s} f_2(x)dx + f_2(\bar{s}).$$

The result follows from a simple integral bound with $\int \frac{1}{x}dx = \log x$ and $\int x^p dx = \frac{x^{p+1}}{p+1}$. □

We state a certain recurrence inequality that appears several times in our analysis as a lemma, in order to shorten some proofs.

Lemma C.2 (General error recurrence). *Let $c_0 \geq 0, c_1 \geq 0, \gamma > 1, a \geq 0$ be arbitrary constants such that $a \geq \gamma$. Furthermore, let $\{u_t\}_{t=0}^{\infty}$ be a sequence of non-negative numbers such that for all $t \geq 0$, it holds that*

$$u_{t+1} \leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + \left(1 - \frac{\gamma}{t+a}\right) u_t.$$

Then, for all values of $t \geq 0$, it holds that:

$$u_{t+1} \leq \frac{u_0 a^\gamma + c_1(a^{\gamma-2}+1)(1+a^{-1})^\gamma + c_0(1+a^{-1})^\gamma a^{\gamma-1}}{(t+a)^\gamma} + \frac{c_0 + c_1(1+a^{-1})^\gamma(\gamma-1)^{-1}}{t+a} + \frac{c_1}{(t+a)^2} + \gamma^{-1}(1+a^{-1})^\gamma c_0$$

Proof. We note that inductively, we have

$$u_{t+1} \leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + u_0 \prod_{s=0}^{t}\left(1 - \frac{\gamma}{s+a}\right) + \sum_{s=0}^{t-1}\left(\frac{c_0}{s+a} + \frac{c_1}{(s+a)^2}\right)\prod_{s'=s+1}^{t}\left(1 - \frac{\gamma}{s'+a}\right).$$

Using the inequality $1 + x \leq e^x$, we obtain

$$\begin{aligned} u_{t+1} &\leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + u_0 \prod_{s=0}^{t}\exp\left\{-\frac{\gamma}{s+a}\right\} + \sum_{s=0}^{t-1}\left(\frac{c_0}{s+a} + \frac{c_1}{(s+a)^2}\right)\prod_{s'=s+1}^{t}\exp\left\{-\frac{\gamma}{s'+a}\right\} \\ &\leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + u_0 \exp\left\{-\sum_{s=0}^{t}\frac{\gamma}{s+a}\right\} + \sum_{s=0}^{t-1}\left(\frac{c_0}{s+a} + \frac{c_1}{(s+a)^2}\right)\exp\left\{-\sum_{s'=s+1}^{t}\frac{\gamma}{s'+a}\right\}. \end{aligned}$$

Here, using Lemma C.1, since $a - 1 > 0$, it holds that

$$\sum_{s=0}^{t}\frac{\gamma}{s+a} \geq \sum_{s=1}^{t+1}\frac{\gamma}{s+(a-1)} \geq \gamma\log(t+a) - \gamma\log a = \log\left(\frac{(t+1)^\gamma}{a^\gamma}\right)$$

$$\sum_{s'=s+1}^{t}\frac{\gamma}{s'+a} \geq \gamma\log(t+a) - \gamma\log(s+a+1) \geq \log\left\{\frac{(t+a)^\gamma}{(s+a+1)^\gamma}\right\},$$

therefore u_{t+1} can be further upper bounded by

$$\begin{aligned} u_{t+1} &\leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + u_0\frac{a^\gamma}{(t+a)^\gamma} + \sum_{s=0}^{t-1}\left(\frac{c_0}{s+a} + \frac{c_1}{(s+a)^2}\right)\left(\frac{s+a+1}{t+a}\right)^\gamma \\ &\leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + \frac{u_0 a^\gamma}{(t+a)^\gamma} + (t+a)^{-\gamma}\sum_{s=0}^{t-1}\left(\frac{c_0}{s+a} + \frac{c_1}{(s+a)^2}\right)(s+a+1)^\gamma \\ &\leq \frac{c_0}{t+a} + \frac{c_1}{(t+a)^2} + \frac{u_0 a^\gamma}{(t+a)^\gamma} + (t+a)^{-\gamma}(1+a^{-1})^\gamma\sum_{s=0}^{t-1}\left(c_0(s+a)^{\gamma-1} + c_1(s+a)^{\gamma-2}\right). \end{aligned}$$

The last term can be bound with the corresponding integral (see Lemma C.1), yielding (since $\gamma - 1 > 0$):

$$\sum_{s=0}^{t-1}(s+a)^{\gamma-1} \leq \frac{(t+a)^\gamma}{\gamma} + a^{\gamma-1}.$$

For the term $\sum_{s=0}^{t-1}(s+1)^{\gamma-2}$, we analyzing the two cases $1<\gamma\leq 2$ and $\gamma>2$ using Lemma C.1 we obtain

$$\sum_{s=0}^{t-1}(s+a)^{\gamma-2}\leq\frac{(t+a)^{\gamma-1}}{\gamma-1}+a^{\gamma-2}+1.$$

The two inequalities combined yield the stated bound,

$$\begin{aligned}
u_{t+1}\leq&\frac{c_0}{t+a}+\frac{c_1}{(t+a)^2}+\frac{u_0a^\gamma}{(t+a)^\gamma}+(t+a)^{-\gamma}(1+a^{-1})^\gamma\sum_{s=0}^{t-1}\left(c_0(s+a)^{\gamma-1}+c_1(s+a)^{\gamma-2}\right)\\
\leq&\frac{c_0}{t+a}+\frac{c_1}{(t+a)^2}+\frac{u_0a^\gamma}{(t+a)^\gamma}+\gamma^{-1}(1+a^{-1})^\gamma c_0+\frac{c_0(1+a^{-1})^\gamma a^{\gamma-1}}{(t+a)^\gamma}\\
&+\frac{c_1(1+a^{-1})^\gamma}{(t+a)(\gamma-1)}+\frac{c_1(a^{\gamma-2}+1)(1+a^{-1})^\gamma}{(t+a)^\gamma}\\
\leq&\frac{u_0a^\gamma+c_1(a^{\gamma-2}+1)(1+a^{-1})^\gamma+c_0(1+a^{-1})^\gamma a^{\gamma-1}}{(t+a)^\gamma}\\
&+\frac{c_0+c_1(1+a^{-1})^\gamma(\gamma-1)^{-1}}{t+a}+\frac{c_1}{(t+a)^2}+\gamma^{-1}(1+a^{-1})^\gamma c_0.
\end{aligned}$$

□

D Proofs of Technical Results

D.1 Proof of Lemma 2.1

Proof. We introduce an auxiliary random variable $\bar{a}^i$ which is independent from other players' actions $\{a^j\}_{j=1}^N$ and has distribution $\boldsymbol{\pi}^i$, that is, we introduce the random variable $\bar{a}^i$ as an identically distributed independent copy of a^i. Then, it holds by simple computation that

$$\begin{aligned}
V^i(\boldsymbol{\pi}^1,\dots,\boldsymbol{\pi}^N)=&\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\sum_{j=1}^N\mathbf{e}_{a^j},a^i\Big)\right]\\
=&\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big),a^i\Big)\right]\\
&+\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\sum_{j=1}^N\mathbf{e}_{a^j},a^i\Big)-\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big),a^i\Big)\right].
\end{aligned}$$

For the first term above, we observe that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big),a^i\Big)\right]=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big),a^i\Big)\Big|a^i\right]\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{e}_{a^i}^\top\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big)\Big)\Big|a^i\right]\right]\\
=&\mathbb{E}[\mathbf{e}_{a^i}^\top]\,\mathbb{E}\left[\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^N\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big)\Big)\right]\\
=&\boldsymbol{\pi}^{i,\top}\,\mathbb{E}[\mathbf{F}(\widehat{\mu})],
\end{aligned}$$

since $\{a^j\}_{j=1}^N$ and $(\bar{a}^i,\mathbf{a}^{-i})$ are identically distributed by the independence of both $\bar{a}^i$ and a^i from $\mathbf{a}^{-i}$, and since $(\bar{a}^i,\mathbf{a}^{-i})$ is independent of a^i.

The second term above can be bounded using

$$\begin{aligned}
&\left|\mathbf{F}\Big(\frac{1}{N}\sum_{j=1}^{N}\mathbf{e}_{a^j},a^i\Big)-\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^{N}\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big),a^i\Big)\right|\\
=&\left|\mathbf{e}_{a^i}^{\top}\mathbf{F}\Big(\frac{1}{N}\sum_{j=1}^{N}\mathbf{e}_{a^j}\Big)-\mathbf{e}_{a^i}^{\top}\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^{N}\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big)\Big)\right|\\
\leq&\|\mathbf{e}_{a^i}\|_2\left\|\mathbf{F}\Big(\frac{1}{N}\sum_{j=1}^{N}\mathbf{e}_{a^j}\Big)-\mathbf{F}\Big(\frac{1}{N}\Big(\sum_{j=1,j\neq i}^{N}\mathbf{e}_{a^j}+\mathbf{e}_{\bar{a}^i}\Big)\Big)\right\|_2\\
\leq&\frac{L}{N}\|\mathbf{e}_{a^i}-\mathbf{e}_{\bar{a}^i}\|_2\leq\frac{L\sqrt{2}}{N}.
\end{aligned}$$

The last line follows from the fact that $\mathbf{F}$ is L-Lipschitz. □

D.2 Proof of Lemma 2.2

We first prove the fact that V^i is Lipschitz. In the proof, we denote the empirical action distribution induced by actions $\{a^j\}_{j=1}^N\in\mathcal{A}^N$ by $\widehat{\mu}(\{a^j\}_{j=1}^N)\in\Delta_{\mathcal{A}}$ to simplify notation. We analyze two cases of Lipschitz moduli stated in the lemma separately. In the first case, we compute the Lipschitz moduli $L_{i,i}$ for any $i\in\mathcal{N}$ as follows:

$$\begin{aligned}
&|V^i(\boldsymbol{\pi},\boldsymbol{\pi}^{-i})-V^i(\boldsymbol{\pi}',\boldsymbol{\pi}^{-i})|\\
&\leq\Big|\mathbb{E}\left[\mathbf{F}\left(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i\right)\Big|a^j\sim\boldsymbol{\pi}^j,\forall j\neq i,a^i\sim\boldsymbol{\pi}\right]\\
&\qquad-\mathbb{E}\left[\mathbf{F}\left(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i\right)\Big|a^j\sim\boldsymbol{\pi}^j,\forall j\neq i,a^i\sim\boldsymbol{\pi}'\right]\Big|\\
&\leq\Big|\sum_{\substack{a^j\in\mathcal{A}\\j\neq i}}\sum_{a^i\in\mathcal{A}}\boldsymbol{\pi}(a^i)\mathbf{F}(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i)\prod_{j\neq i}\boldsymbol{\pi}^j(a^j)-\sum_{\substack{a^j\in\mathcal{A}\\j\neq i}}\sum_{a^i\in\mathcal{A}}\boldsymbol{\pi}'(a^i)\mathbf{F}(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i)\prod_{j\neq i}\boldsymbol{\pi}^j(a^j)\Big|\\
&\leq\sum_{\substack{a^j\in\mathcal{A}\\j\neq i}}\Big|\sum_{a^i\in\mathcal{A}}\left[\boldsymbol{\pi}(a^i)-\boldsymbol{\pi}'(a^i)\right]\mathbf{F}(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i)\Big|\prod_{j\neq i}\boldsymbol{\pi}^j(a^j)\\
&\leq\sum_{\substack{a^j\in\mathcal{A}\\j\neq i}}\|\boldsymbol{\pi}-\boldsymbol{\pi}'\|_2\sqrt{\sum_{a^i\in\mathcal{A}}\mathbf{F}(\widehat{\mu}(\{a^k\}_{k=1}^N),a^i)^2}\prod_{j\neq i}\boldsymbol{\pi}^j(a^j)\leq\|\boldsymbol{\pi}-\boldsymbol{\pi}'\|_2\sqrt{K}.
\end{aligned}$$

where we use Jensen's inequality in the penultimate step and the Cauchy-Schwartz inequality in the final step.

In the second case, for any $k\neq i$, it holds that

$$\begin{aligned}
&|V^i(\boldsymbol{\pi},\boldsymbol{\pi}^{-k})-V^i(\boldsymbol{\pi}',\boldsymbol{\pi}^{-k})|\\
&\leq\mathbb{E}\left[\mathbf{F}\left(\widehat{\mu}(\{a^l\}_{l=1}^N),a^i\right)\Big|a^j\sim\boldsymbol{\pi}^j,\forall j\neq k,a^k\sim\boldsymbol{\pi}\right]-\mathbb{E}\left[\mathbf{F}\left(\widehat{\mu}(\{a^l\}_{l=1}^N),a^i\right)\Big|a^j\sim\boldsymbol{\pi}^j,\forall j\neq k,a^k\sim\boldsymbol{\pi}'\right]\\
&\leq\Big|\sum_{\substack{a^j\in\mathcal{A}\\j\neq k}}\prod_{j\neq k}\boldsymbol{\pi}^j(a^j)\sum_{a^k\in\mathcal{A}}\boldsymbol{\pi}(a^k)\mathbf{F}(\widehat{\mu}(\{a^l\}_{l=1}^N),a^i)-\sum_{\substack{a^j\in\mathcal{A}\\j\neq k}}\prod_{j\neq k}\boldsymbol{\pi}^j(a^j)\sum_{a^k\in\mathcal{A}}\boldsymbol{\pi}'(a^k)\mathbf{F}(\widehat{\mu}(\{a^l\}_{l=1}^N),a^i)\Big|\\
&\leq\sum_{\substack{a^j\in\mathcal{A}\\j\neq k}}\prod_{j\neq k}\boldsymbol{\pi}^j(a^j)\Big|\sum_{a^k\in\mathcal{A}}\left[\boldsymbol{\pi}(a^k)-\boldsymbol{\pi}'(a^k)\right]\mathbf{F}(\widehat{\mu}(\{a^l\}_{l=1}^N),a^i)\Big|
\end{aligned}$$

In this case, note that for any $a, a' \in \mathcal{A}, \mathbf{a} \in \mathcal{A}^K$, we have $|\mathbf{F}(\widehat{\mu}(a, \mathbf{a}^{-k}), a^i) - \mathbf{F}(\widehat{\mu}(a', \mathbf{a}^{-k}), a^i)| \leq \|\mathbf{F}(\widehat{\mu}(a, \mathbf{a}^{-k})) - \mathbf{F}(\widehat{\mu}(a', \mathbf{a}^{-k}))\|_2 \leq {}^{L\sqrt{2}}/_{N}$. That is, the set $\{\mathbf{F}(\widehat{\mu}(a, \mathbf{a}^{-k}), a^i) : a \in \mathcal{A}\} \subset \mathbb{R}$ has diameter ${}^{2L}/_{\sqrt{N}}$, and there exists a constant $v_k \in \mathbb{R}$ such that $|\mathbf{F}(\widehat{\mu}(a, \mathbf{a}^{-k}), a^i) - v_k| \leq {}^{2L\sqrt{2}}/_{N}$ for all a. Then,

$$
\begin{aligned}
&|V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-k}) - V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-k})| \\
&\leq \sum_{\substack{a^j \in \mathcal{A} \\ j \neq k}} \prod_{j \neq k} \boldsymbol{\pi}^j(a^j) \Big| \sum_{a^k \in \mathcal{A}} \Big[\boldsymbol{\pi}(a^k) - \boldsymbol{\pi}'(a^k)\Big] \Big[\mathbf{F}(\widehat{\mu}(\{a^l\}_{l=1}^N), a^i) - v_k\Big] \Big| \\
&\leq \sum_{\substack{a^j \in \mathcal{A} \\ j \neq k}} \prod_{j \neq k} \boldsymbol{\pi}^j(a^j) \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \sqrt{\sum_{a^k \in \mathcal{A}} \Big[\mathbf{F}(\widehat{\mu}(\{a^l\}_{l=1}^N), a^i) - v_k\Big]^2} \\
&\leq \sum_{\substack{a^j \in \mathcal{A} \\ j \neq k}} \prod_{j \neq k} \boldsymbol{\pi}^j(a^j) \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \frac{2L\sqrt{2K}}{N} \leq \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \frac{2L\sqrt{2K}}{N}.
\end{aligned}
$$

We establish the Lipschitz continuity of $\mathcal{E}^i_{\text{exp}}$ with the above inequalities.

$$
\begin{aligned}
&|\mathcal{E}^i_{\text{exp}}(\boldsymbol{\pi}, \boldsymbol{\pi}^{-i}) - \mathcal{E}^i_{\text{exp}}(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i})| \\
&\quad \leq \left| \max_{\overline{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}} V^i(\overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-i}) - V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-i}) - \max_{\overline{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}} V^i(\overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-i}) + V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) \right| \\
&\quad \leq \left| V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) - V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-i}) \right| \leq \sqrt{K} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2.
\end{aligned}
$$

Similarly, for $k \neq i$,

$$
\begin{aligned}
&|\mathcal{E}^i_{\text{exp}}(\boldsymbol{\pi}, \boldsymbol{\pi}^{-k}) - \mathcal{E}^i_{\text{exp}}(\boldsymbol{\pi}', \boldsymbol{\pi}^{-k})| \\
&\leq \left| \max_{\overline{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}} [V^i(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-k,i}) - V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-k})] - \max_{\overline{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}} [V^i(\boldsymbol{\pi}', \overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-k,i}) - V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-k})] \right| \\
&\leq \max_{\overline{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}} \left| V^i(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-k,i}) - V^i(\boldsymbol{\pi}', \overline{\boldsymbol{\pi}}, \boldsymbol{\pi}^{-k,i}) \right| + \left| V^i(\boldsymbol{\pi}, \boldsymbol{\pi}^{-k}) - V^i(\boldsymbol{\pi}', \boldsymbol{\pi}^{-k}) \right| \\
&\leq \frac{4L\sqrt{2K}}{N} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2.
\end{aligned}
$$

D.3 Extended Definitions for Bandit Feedback

When analyzing the TRPA-Bandit dynamics, several random variables and events will be reused to assist analysis. For brevity, we define them here. We define the following random variables:

$$
\begin{aligned}
\mathbb{1}^i_{h,t} &:= \mathbb{1}\{\text{player } i \text{ explores at round } t \text{ of epoch } h\} = X^i_{h,t} \\
E^i_{h,t} &:= \{\mathbb{1}^i_{h,t} = 1\} \\
\mathbb{1}^i_h &:= \mathbb{1}\{\text{player } i \text{ explores at least once during epoch } h\} = \max_{t=1,\ldots,T_h} \mathbb{1}^i_{h,t} \\
E^i_h &:= \{\mathbb{1}^i_h = 1\} = \bigcup_{t=1}^{T_h} E^i_{h,t} \\
a^i_h &:= \text{Last explored action in epoch } h \text{ by agent } i, \\
&\qquad \text{and } a_0 \text{ if no exploration occurred.} \\
s^i_h &:= \text{Timestep when exploration last occurred in epoch } h \text{ by agent } i, \\
&\qquad \text{and 0 if no exploration occurred.} \in \{1, \ldots, T_h\}
\end{aligned}
$$

D.4 Proof of Lemma 5.1

We will reuse the definitions of Section D.3. By the definition of the events and the probabilistic exploration scheme, we have $\widehat{\mathbf{r}}_h^i = K(\mathbf{F}(\widehat{\mu}_{s_h^i,h}, a_h^i) + \mathbf{n}^i_{s_h^i,h}(a_h^i))\mathbb{1}^i_h$. Firstly, by the law of total expectations and the fact that E_h^i are independent of $\mathcal{F}_h$,

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| \mathcal{F}_h\right] &= \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right]\mathbb{P}(E_h^i) + \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| \overline{E_h^i}, \mathcal{F}_h\right]\mathbb{P}(\overline{E_h^i}) \\ &= \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right] \underbrace{- \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right]\mathbb{P}(\overline{E_h^i}) + \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| \overline{E_h^i}, \mathcal{F}_h\right]\mathbb{P}(\overline{E_h^i})}_{:=\mathbf{b}_h^i} \\ &= \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right] + \mathbf{b}_h^i\end{aligned}$$

for $\mathbf{b}_h^i$ quantifying a bias induced due to the probability of no exploration. We have that $\|\mathbf{b}_h^i\|_2 \leq K\sqrt{K}\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\}$ since $\mathbb{E}[\widehat{\mathbf{r}}_h^i | \overline{E_h^i}, \mathcal{F}_h] = 0$ and exploration probabilities are determined by independent random Bernoulli variables hence $\mathbb{P}(\overline{E_h^i}) = (1-\varepsilon)^{T_h} \leq \exp\{-\varepsilon T_h\}$. To further characterize the bias, we introduce a coupling argument. Define independent random variables $\bar{a}_h^j \sim \varepsilon\boldsymbol{\pi}_{\text{unif}} + (1-\varepsilon)\boldsymbol{\pi}_h^j$ for all $j \in \mathcal{N}$ and $\bar{a}^i_{h,exp} \sim \boldsymbol{\pi}_{\text{unif}}$. By the definition, it holds that (where $\widehat{\mu}(\{\bar{a}_h^j\}_j) \in \Delta_{\mathcal{A}}$ denotes the empirical distribution induced by actions $\{\bar{a}_h^j\}_j$):

$$\begin{aligned}\| \mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right] - \mathbb{E}\left[\mathbf{F}(\widehat{\mu}(\{\bar{a}_h^j\}_j)) \middle| \mathcal{F}_h\right] \|_2 &\leq \| \mathbb{E}\left[\mathbf{F}(\widehat{\mu}(\bar{a}^i_{h,exp}, \bar{a}_h^{-i})) \middle| \mathcal{F}_h\right] - \mathbb{E}\left[\mathbf{F}(\widehat{\mu}(\{\bar{a}_h^j\}_j)) \middle| \mathcal{F}_h\right] \|_2 \\ &\leq \mathbb{E}\left[\|\mathbf{F}(\widehat{\mu}(\bar{a}^i_{h,exp}, \bar{a}_h^{-i})) - \mathbf{F}(\widehat{\mu}(\{\bar{a}_h^j\}_j))\|_2 \middle| \mathcal{F}_h\right] \leq \frac{2L}{N}.\end{aligned}$$

With the additional bound $\mathbb{E}[\|\mathbf{F}(\varepsilon\boldsymbol{\pi}_{\text{unif}} + (1-\varepsilon)\bar{\mu}_h) - \mathbf{F}(\widehat{\mu}(\{\bar{a}_h^j\}_j))\|_2 | \mathcal{F}_h] \leq \frac{2L}{\sqrt{N}}$, we obtain the lemma.

D.5 Proof of Lemma 5.2

We will reuse the definitions of Section D.3. We formulate a recurrence for the main error term of interest, $\|\boldsymbol{\pi}^i_{t+1} - \boldsymbol{\pi}^*\|_2^2$, Repeating the steps presented in Lemma 4.2, (noting $\alpha_h := 1 - \tau\eta_h$), our proof strategy is to analyze the three terms in the following decomposition:

$$\begin{aligned}\|\boldsymbol{\pi}^i_{h+1} - \boldsymbol{\pi}^*\|_2^2 \leq \underbrace{\eta_h^2\|\widehat{\mathbf{r}}_h^i - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2 + 2\eta_h^2(\mathbf{F}(\boldsymbol{\pi}_h^i) - \mathbf{F}(\boldsymbol{\pi}^*)))^\top(\widehat{\mathbf{r}}_h^i - \mathbf{F}(\boldsymbol{\pi}_h^i))}_{(a)} \\ + \underbrace{2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \mathbf{F}(\boldsymbol{\pi}_h^i))}_{(b)} + \underbrace{\|\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*) + \eta_h(\mathbf{F}(\boldsymbol{\pi}_h^i) - \mathbf{F}(\boldsymbol{\pi}^*))\|_2^2}_{(c)}\end{aligned}$$

Once again, we will need to upper bound the three terms above. For the term (a) we have $\mathbb{E}[(a)] \leq 4\eta_t^2K^3(\sigma^2+1)$, noting that $\|\widehat{\mathbf{r}}_h^i\|_2^2 \leq K^3$ almost surely. Likewise, it still holds for the term (c) by Lemma 4.1 that

$$(c) \leq \left(1 - 2(\lambda+\tau)\eta_h + (L+\tau)^2\eta_h^2\right)\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2.$$

However, unlike the bound in Lemma 4.2, the exploration parameter ε will cause additional bias in the term (b). Define the random vector $\widetilde{\mathbf{r}}_h^i = \mathbb{E}[\widehat{\mathbf{r}}_h^i | E_h^i, \mathcal{F}_h]$.

$$
\begin{aligned}
(b) =& 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \mathbf{F}(\boldsymbol{\pi}_h^i)) \\
=& 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i) + 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)) \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)) \\
\leq& 2\eta_h\alpha_h\left(\frac{\lambda}{4}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{\lambda}\|\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)\|_2^2\right) \\
&+ 2\eta_h\alpha_h\left(\frac{\lambda}{4}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{\lambda}\|\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2\right) \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i) \\
\leq& 2\eta_h\frac{\lambda}{2}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{2\eta_h}{\lambda}\|\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)\|_2^2 + \frac{2\eta_h}{\lambda}\|\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2 \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i),
\end{aligned}
$$

and similarly, if $\lambda = 0$, we have

$$
\begin{aligned}
(b) =& 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \mathbf{F}(\boldsymbol{\pi}_h^i)) \\
=& 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i) + 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)) \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)) \\
\leq& 2\eta_h\alpha_h\left(\frac{\tau\delta}{2}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{2\tau\delta}\|\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)\|_2^2\right) \\
&+ 2\eta_h\alpha_h\left(\frac{\tau\delta}{2}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{1}{2\tau\delta}\|\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2\right) \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i) \\
\leq& 2\eta_h\tau\delta\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2^2 + \frac{\eta_h}{\tau\delta}\|\widetilde{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)\|_2^2 + \frac{\eta_h}{\tau\delta}\|\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2 \\
&+ 2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i).
\end{aligned}
$$

Denote $\boldsymbol{\pi}_{unif} := \frac{1}{K}\mathbf{1}_K$. The remaining error terms we can bound by (using the auxiliary coupling random actions $\bar{a}_h^j$ from Lemma 5.1):

$$
\begin{aligned}
|\,\mathbb{E}[2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i)|\mathcal{F}_h]| \leq& |\,\mathbb{E}[2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i)|E_h^i, \mathcal{F}_h]\mathbb{P}(E_h^i) \\
&+ \mathbb{E}[2\eta_h\alpha_h(\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*)^\top(\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i)|\overline{E_h^i}, \mathcal{F}_h]\mathbb{P}(\overline{E_h^i})| \\
\leq& 2\eta_h\,\mathbb{E}\left[\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}^*\|_2\|\widehat{\mathbf{r}}_h^i - \widetilde{\mathbf{r}}_h^i\|_2|\overline{E_h^i}, \mathcal{F}_h\right]\mathbb{P}(\overline{E_h^i}) \\
\leq& 8\eta_h K^{3/2}\sqrt{1+\sigma^2}\,\mathbb{P}(\overline{E_h^i})
\end{aligned}
$$

and $\mathbb{P}(\overline{E_h^i}) \leq \exp\{-\varepsilon T_h\}$. Furthermore,

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{\mathbf{r}}_h^i - \mathbf{F}(\bar{\mu}_h)\|_2^2 | \mathcal{F}_h] \\
&\leq 2\,\mathbb{E}\left[\|\widehat{\mathbf{r}}_h^i - \mathbf{F}(\widehat{\mu}(\bar{a}^i_{h,exp}, \bar{a}_h^{-i}))\|_2^2 \middle| \mathcal{F}_h\right] + 2\,\mathbb{E}\left[\|\mathbf{F}(\widehat{\mu}(\bar{a}^i_{h,exp}, \bar{a}_h^{-i})) - \mathbf{F}(\bar{\mu}_h)\|_2^2 | \mathcal{F}_h\right] \\
&\leq \frac{8L^2}{N} + 2L^2\,\mathbb{E}[\|\widehat{\mu}(\bar{a}^i_{exp}, \bar{a}^{-i}) - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{\pi}_h^j\|_2^2 | \mathcal{F}_h] \\
&\leq \frac{8L^2}{N} + 4L^2\,\mathbb{E}[\|\widehat{\mu}(\bar{a}^i_{exp}, \bar{a}^{-i}) - \frac{1}{N}\sum_{j\neq i} (\varepsilon \boldsymbol{\pi}_{unif} + (1-\varepsilon)\boldsymbol{\pi}_h^j) - \frac{\boldsymbol{\pi}_{unif}}{N}\|_2^2 | \mathcal{F}_h] \\
&\quad + 4L^2\,\mathbb{E}[\|\varepsilon \frac{1}{N}\sum_{j\neq i} (\varepsilon \boldsymbol{\pi}_h^j - \boldsymbol{\pi}_{unif}) + \frac{\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_{unif}}{N}\|_2^2 | \mathcal{F}_h] \\
&\leq \frac{8L^2}{N} + \frac{8L^2}{N} + 4L^2(2\varepsilon^2 + \frac{4}{N}) \\
&\leq \frac{64L^2}{N} + 8L^2\varepsilon^2,
\end{aligned}
$$

and finally, using the trivial Lipschitz continuity property: $\|\mathbf{F}(\bar{\mu}_h) - \mathbf{F}(\boldsymbol{\pi}_h^i)\|_2^2 \leq L^2 \|\bar{\mu}_h - \boldsymbol{\pi}_h^i\|_2^2 = L^2 e_h^i$.

D.6 Proof of Lemma 5.3

We will reuse the definitions of Section D.3. Once again, repeating the derivations from Lemma 4.3, we have that

$$
\|\boldsymbol{\pi}^i_{h+1} - \boldsymbol{\pi}^j_{h+1}\|_2^2 = (1-\tau\eta_h)^2 \|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \eta_h^2 \|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2 + 2(1-\tau\eta_h)\eta_h (\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j)^\top (\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j).
$$

Unlike in the expert feedback case, the last term does not vanish in expectation when bounding policy deviation.

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\pi}^i_{h+1} - \boldsymbol{\pi}^j_{h+1}\|_2^2 | \mathcal{F}_h\right] \leq& (1-\tau\eta_h)^2 \|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \mathbb{E}\left[\eta_h^2 \|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2 | \mathcal{F}_h\right] \\
&+ 2(1-\tau\eta_h)\eta_h (\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j)^\top\, \mathbb{E}\left[\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j | \mathcal{F}_h\right] \\
\leq& (1-\tau\eta_h)^2 \|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \mathbb{E}\left[\eta_h^2 \|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2 | \mathcal{F}_h\right] \\
&+ 2\eta_h (\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j)^\top \left[\mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right] + \mathbf{b}_h^i - \mathbb{E}\left[\widehat{\mathbf{r}}_h^j \middle| E_h^j, \mathcal{F}_h\right] - \mathbf{b}_h^j\right] \\
\leq& (1-\tau\eta_h)^2 \|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \mathbb{E}\left[\eta_h^2 \|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2 | \mathcal{F}_h\right] \\
&+ 2\eta_h (\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j)^\top \left[\mathbb{E}\left[\widehat{\mathbf{r}}_h^i \middle| E_h^i, \mathcal{F}_h\right] - \mathbb{E}\left[\widehat{\mathbf{r}}_h^j \middle| E_h^j, \mathcal{F}_h\right]\right] + 8\eta_h \exp\{-\varepsilon T_h\},
\end{aligned}
$$

where the last line follows from the bound on $\mathbf{b}_h^i$ in Lemma 5.1. Furthermore, using Young's inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{h+1}^i - \boldsymbol{\pi}_{h+1}^j\|_2^2|\mathcal{F}_h\right] \leq& (1-\tau\eta_h)^2\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \mathbb{E}\left[\eta_h^2\|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2|\mathcal{F}_h\right] + 8\eta_h\exp\{-\varepsilon T_h\}\\ &+ \frac{\tau\eta_h}{2}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \frac{\eta_h\tau^{-1}}{2}\left\|\mathbb{E}\left[\widehat{\mathbf{r}}_h^i\middle|E_h^i,\mathcal{F}_h\right] - \mathbb{E}\left[\widehat{\mathbf{r}}_h^j\middle|E_h^j,\mathcal{F}_h\right]\right\|_2^2\\ \leq& (1-\tau\eta_h)^2\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \mathbb{E}\left[\eta_h^2\|\widehat{\mathbf{r}}_h^i - \widehat{\mathbf{r}}_h^j\|_2^2|\mathcal{F}_h\right] + 8\eta_h\exp\{-\varepsilon T_h\}\\ &+ \frac{\tau\eta_h}{2}\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2 + \frac{2\eta_h\tau^{-1}L^2}{N^2}.\end{aligned}$$

With the choice of $T_h = \varepsilon^{-1}\log(h+2)$ and noting that $\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2 \leq 2$, we obtain the recurrence

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\pi}_{h+1}^i - \boldsymbol{\pi}_{h+1}^j\|_2^2\right] \leq& \left(1 - \frac{3/2}{h+2}\right)\mathbb{E}\left[\|\boldsymbol{\pi}_h^i - \boldsymbol{\pi}_h^j\|_2^2\right] + \frac{4\tau^{-2}K^3(\sigma^2+1) + 8\tau^{-1} + 4}{(h+2)^2}\\ &+ \frac{4\tau^{-2}L^2}{N^2(h+2)}.\end{aligned}$$

Hence, by invoking the recurrence lemma (Lemma C.2, with $a = 2, c_0 = \frac{4\tau^{-2}L^2}{N^2}, c_1 = 4\tau^{-2}K^3(\sigma^2+1) + 8\tau^{-1} + 4, \gamma = 3/2, u_0 = 0$), we have

$$\mathbb{E}\left[\|\boldsymbol{\pi}_{h+1}^i - \boldsymbol{\pi}_{h+1}^j\|_2^2\right] \leq \frac{24\tau^{-2}K^3(\sigma^2+1) + 48\tau^{-2} + 24}{h+2} + \frac{16\tau^{-2}L^2}{N^2}.$$

The statement in the lemma follows as in the full feedback case.

D.7 Proof of Theorem 5.4

Using Lemma 5.2, for the strongly monotone case $\lambda > 0$ we have that

$$\begin{aligned}u_{h+1}^i \leq& 4\eta_h^2K^3(1+\sigma^2) + 8\eta_h^2(L+\tau)^2 + 8K^{3/2}\eta_h\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\}\\ &+ 128\eta_h\lambda^{-1}L^2N^{-1} + 16\eta_h\lambda^{-1}L^2\varepsilon^2 + 2\eta_h\lambda^{-1}L^2\,\mathbb{E}\left[e_h^i\right]\\ &+ (1 - 2\eta_h(\lambda/2 + \tau))\,u_h^i,\end{aligned}$$

and for the monotone case it holds that

$$\begin{aligned}u_{h+1}^i \leq& 4\eta_h^2K^3(\sigma^2+1) + 8\eta_h^2(L+\tau)^2 + 8K^{3/2}\eta_h\sqrt{1+\sigma^2}\exp\{-\varepsilon T_h\}\\ &+ 64\tau^{-1}\eta_h\delta^{-1}L^2N^{-1} + 8\tau^{-1}\eta_h\delta^{-1}L^2\varepsilon^2 + \tau^{-1}\eta_h\delta^{-1}L^2\,\mathbb{E}\left[e_h^i\right]\\ &+ (1 - 2\tau\eta_h(1-\delta))\,u_h^i.\end{aligned}$$

By Lemma 5.3, we know that

$$\mathbb{E}[e_h^i] \leq \frac{12\tau^{-2}K^3(\sigma^2+1) + 48\tau^{-2} + 24}{h+2} + \frac{16\tau^{-2}L^2}{N^2}.$$

Placing this bound as well as $T_h = \lceil\varepsilon^{-1}\log(h+2)\rceil$ and $\eta_h = \frac{\tau^{-1}}{h+2}$, we obtain the recurrences which are solved by using Lemma C.2. In the monotone case, we pick $\delta = 1/4$ as before.

The bound in the statement of the theorem in the main body of the article follows from the fact that the lengths of the exploration epochs scale with $T_h = O(\varepsilon^{-1}\log(h+2)) = \widetilde{O}(\varepsilon^{-1})$.

E Details of Experiments

Setup. All experiments were run on single core of an AMD EPYC 7742 CPU, a single experiment with 1,000 independent agents and 100,000 iterations takes roughly 1 hour. For all experiments, we use parameters τ, ϵ as implied by Corollaries 4.5,5.5. Projections to the probability simplex were implemented using the algorithm by Ref. [15].

E.1 Problem Generation Details

We provide further details on how we generate/simulate the SMFG problems.

Linear payoffs. We generate a payoff map

$$\mathbf{F}_{lin}(\mu) := (\mathbf{S} + \mathbf{X})\mu + \mathbf{b}$$

for some $\mathbf{S} \in \mathbb{S}^{K\times K}_{++}$ and $\mathbf{X}$ anti-symmetric matrix, which makes $\mathbf{F}_{lin}$ monotone. We randomly sample $\mathbf{S}$ from a Wishart distribution (which has support contained in positive definite matrices), generate $\mathbf{X}$ by computing $\frac{\mathbf{U}-\mathbf{U}^\top}{2}$ for a random matrix $\mathbf{U}$ with entries sampled uniformly from $[0, 1]$ and $\mathbf{b}$ having entries uniformly sampled from $[0, 1]$.

Payoffs with KL potential. Next, we construct the following payoff operator $\mathbf{F}_{KL}$ for some reference distribution $\mu_{\text{ref}} \in \Delta_{\mathcal{A}}$:

$$\begin{aligned}\Phi_{KL}(\mu) &:= \mathrm{D}_{KL}(\gamma\mu + (1-\gamma)\mu_{\text{ref}}||\mu_{\text{ref}})\\ \mathbf{F}_{KL}(\mu) &:= \nabla\Phi_{KL}(\mu)\\ \mathbf{F}_{KL}(\mu, a) &= \gamma \log\left(\frac{\gamma\mu(a) + (1-\gamma)\mu_{\text{ref}}(a)}{\mu_{\text{ref}}(a)}\right) + \gamma\end{aligned}$$

Note that as Φ_{KL} is convex, $\mathbf{F}_{KL}$ is monotone. In our experiments, we use $\gamma = 0.1$, and we generate μ_{ref} by sampling K uniform random variables in $[0, 1]$ and normalizing.

Beach bar process. Following the example given in Ref. [53], we use the action set $\mathcal{A} = \{1, \dots, K\}$ for potential locations at the beach and assume a bar is located at $x_{bar} := \lfloor \frac{K}{2} \rfloor$. Taking into the proximity to the bar and the occupancy measure over actions (i.e., the crowdedness of locations at the beach), the payoff map is given by

$$\mathbf{F}_{bb}(\mu, a) = 1 - \frac{|a - x_{bar}|}{K} - \alpha \log(1 + \mu(a)).$$

The above payoff map is monotone. We use $\alpha = 1$ for our experiments.

E.2 Learning Curves—Full Feedback

We provide the learning curves under full feedback for various choices of the number of agents $N \in \{20, 100, 1000\}$. The errors in terms of maximum exploitability and distance to MF-NE are presented in Figures 3 and 4, respectively.

As expected, the games with larger number of players N converge to better approximate NE in the sense that the final maximum exploitability is smaller at convergence. Furthermore, in most cases the exploitability converges slightly slower with more agents, also supporting the theoretical finding that there is a dependence on N. As before, the exploitability curves have oscillations at later stages of the training, even though they remain upper bounded as foreseen the theoretical results. This does not contradict our results as long as for larger N, the upper bound on the oscillations us smaller. The confidence intervals plotted in figures support a high-probability upper bound on the maximum exploitability as one would expect.

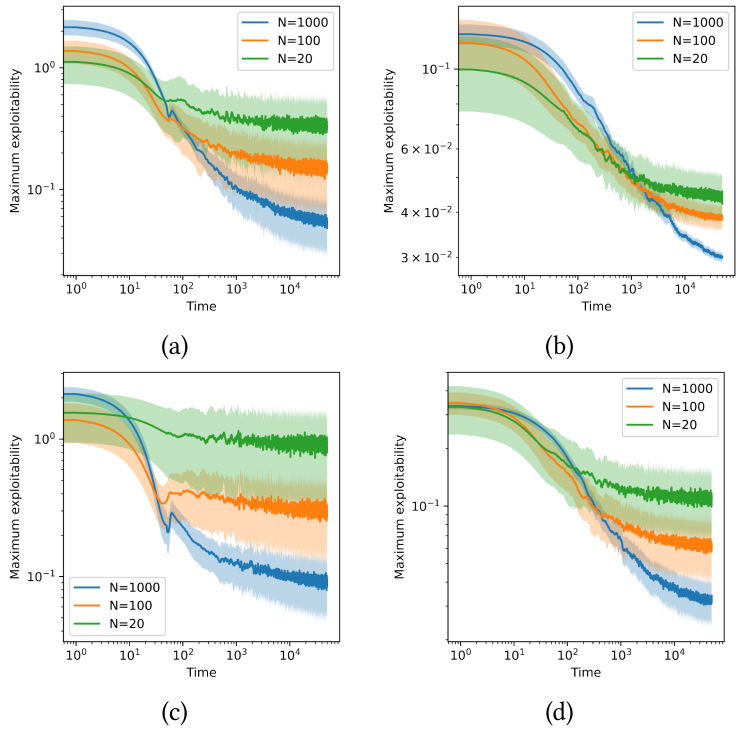

Fig. 3. The (smoothed) maximum exploitability $\max_{i\in\mathcal{N}} \mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^j\}_{j=1}^N)$ among N agents throughout learning with full feedback for three different N, on the problems (a) linear payoffs, (b) exponentially decreasing payoffs, (c) payoffs with KL potential and (d) the beach bar payoffs.

E.3 Learning Curves—Bandit Feedback

We provide the learning curves under bandit feedback for various choices of the number of agents $N \in \{50, 100, 1000\}$. The errors in terms of maximum exploitability and distance to MF-NE are presented in Figures 5 and 6, respectively.

As in the case of full feedback, the curves converge to smaller values as N increases. Furthermore, one straightforward observation is that the variance at early stages of learning is much higher than in the full feedback case. This can be due to the added variance of the importance sampling estimator constructed through exploration epochs. As exploration occurs in shorter duration at early epochs, the variance between agent policies will be high as well, explaining the initial increase in exploitability in certain toy experiments in Figure 5.

Furthermore, comparing the observations for bandit feedback (Figure 5) and full feedback cases (Figure 3), we empirically confirm that learning take more iterations in the bandit case. This is likely due to the fact that exploration occurs probabilistically, inducing additional variance in the policy updates that increases with N and incorporates an additional logarithmic term in the theoretical bounds. However, the number of exploration epochs in the bandit case is comparable to the number of time steps in the full feedback case.

Finally, we also emphasize the fact that in earlier stages of training with bandit feedback, the cases with $N = 1000$ had much higher exploitability and ℓ_2 distance to MF-NE at earlier time steps. This is due to the fact that policy updates occur with larger intervals in between when N is large, as can be seen in Algorithm 2. This can be explicitly observed in Figure 6, as the policies of agents

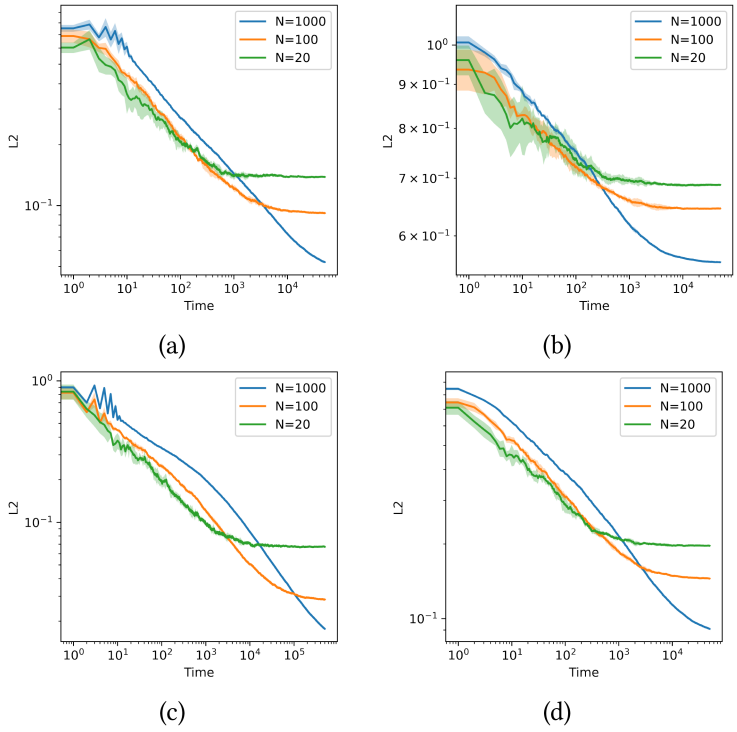

Fig. 4. The mean ℓ_2 distance to MF-NE given by $\frac{1}{N}\sum_{i\in\mathcal{N}}\|\boldsymbol{\pi}^i-\boldsymbol{\pi}^*\|_2$ with N agents throughout learning with full feedback for three different N, on the problems (a) linear payoffs, (b) exponentially decreasing payoffs, (c) payoffs with KL potential and (d) the beach bar payoffs.

are constant in between policy updates. However, at later stages as the time-dependent term in the bound on epxloitability in Corollary 5.5 disappears, we observe that the experiments for larger N converge to a better policy (i.e., one with lower exploitability) than the cases with smaller N as the theory suggests.

Finally, comparing Figures 5 and 6, we see that in certain experiments for some N despite having a lower exploitability we observe a greater ℓ_2 distance to MF-NE. This likely due to fact that the non-vanishing bias term in exploitability and ℓ_2 distance have differing dependence on problem parameters such as L, λ, K. Therefore, for instance for the KL potential payoffs, we observe a greater mean ℓ_2 distance to MF-NE but a smaller exploitability for $N = 1000$.

E.4 Comparison with Multiplicative Weight Updates

The SMFG problem of interactive learning without communications and bandit feedback has not been studied in the literature previously, and there is no alternative algorithm with theoretical guarantees to the best of our knowledge in this setting. However, in Figure 7, we provide comparison with a heuristic extension of the OMD algorithm proposed by Pérolat et al. [52] which was analyzed in the case of an infinite agent game. In general, independent learning with N agents introduces unique challenges that cause the algorithm to cycle or diverge.

E.5 Experiments on Traffic Congestion

UTD19 and closed-loop sensors. The UTD19 dataset contains measurements by closed-loop sensors which report the fraction of the time a particular section of the route remains occupied

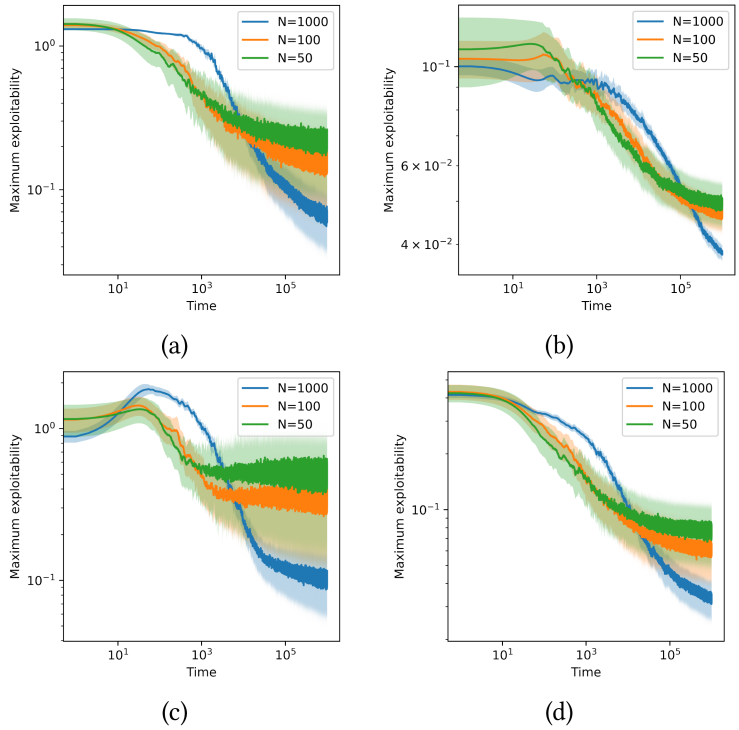

Fig. 5. The (smoothed) maximum exploitability $\max_{i \in \mathcal{N}} \mathcal{E}^i_{\text{exp}}(\{\boldsymbol{\pi}^j\}_{j=1}^N)$ among N agents throughout learning with bandit feedback for three different N, on the problems (a) linear payoffs, (b) exponentially decreasing payoffs, (c) payoffs with KL potential and (d) the beach bar payoffs.

(i.e., a car is located in between sensors placed on the sides). The data consists of measurements every 5 minutes, from various sensors across 41 European cities. The dataset contains 2 weeks of data collected by sensors placed around Zurich.

Payoff models. We fit kernelized ridge regression models to model the flow as a function of occupancy using the UTD19 dataset. We use an RBF kernel and a regularization of $\alpha = 1.0$ for all models. The dependence of traffic speed on route occupancy in two sample paths in the dataset can be seen in Figure 8-(a, b). The measurements motivate the monotonicity assumption as both the original data and the fitted regression model, increased occupancy leads to increased travel time, at least, when ignoring interactions. We compute a proxy for the travel velocity using the flow and occupancy measurements on each route, and a scaling factor c_{dist} due to varying lengths of each route, leading to the estimated travel time

$$T_{travel} = c_{\text{dist}} \frac{\text{flow}}{\text{occupancy}}.$$

We use $-T_{travel}$ as the reward for each agent.

E.6 Experiments on Network Access

For the Tor network access experiment, we randomly chose 5 entry guard servers (the complete list available publicly [67]) in various geographic locations, among the servers that have the longest recorded uptime. To simulate access to each server, we ping each node 5 consecutive times and

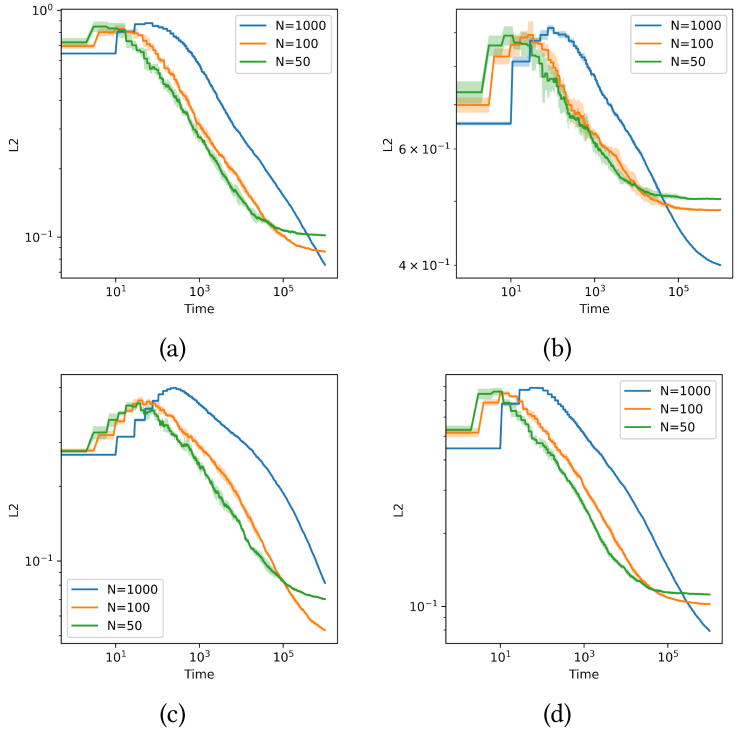

Fig. 6. The mean ℓ_2 distance to MF-NE given by $\frac{1}{N}\sum_{i\in\mathcal{N}}\|\boldsymbol{\pi}^i - \boldsymbol{\pi}^*\|_2$ with N agents throughout learning with bandit feedback for three different N, on the problems (a) linear payoffs, (b) exponentially decreasing payoffs, (c) payoffs with KL potential and (d) the beach bar payoffs.

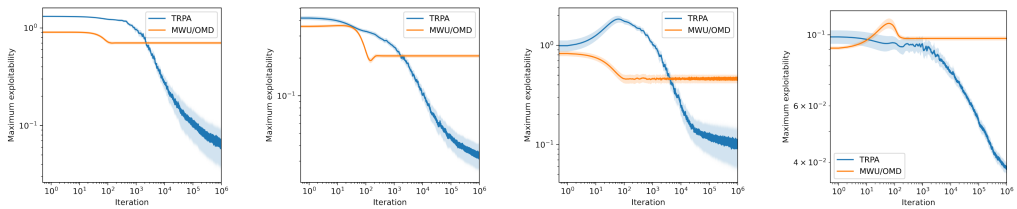

Fig. 7. Comparison of maximum exploitability between our TRPA-Bandit and multiplicative weight updates (equivalently OMD of Pérolat et al. [52]) in benchmarks LINEAR, BB, KL, EXP (left to right). $N = 1000$ agents, averaged over 10 experiments.

average the delays to compute the cost. As expected, due to varying bandwidths/computational power, each server has different sensitivities to load in terms of delay. We plot measured ping times vs the number of agents simultaneously accessing a particular node in Figure 8-(c). The ping times were measured to be stochastic but clearly sensitive to occupancy, motivating a monotone payoff operator.

We use parameters $\tau = 0.01, \varepsilon = 0.3$ for the experiments in this section. The arbitrary choice is due to the fact that in the presence of external agents in the game that do not use TRPA-Bandit (in this case, thousands of other users accessing the Tor network), the theoretically optimal parameters implied by Corollary 5.5 cannot be used. While more realistic simulations that are also closer to

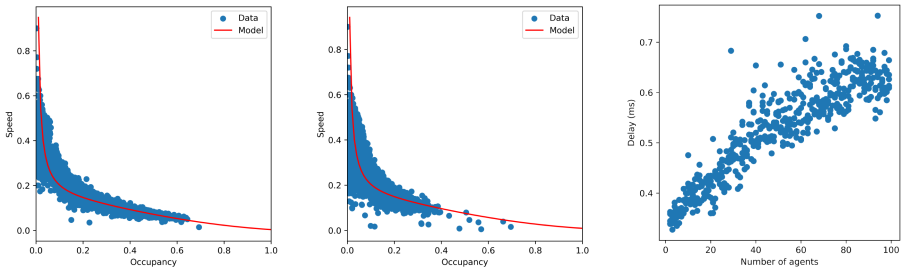

Fig. 8. (a, b) Data and fitted models of occupancy vs. speed on two different urban roads (sensors K324D11 and K418D12) from the UTD19 dataset. The red line indicates the fitted model predictions of the kernelized regression model. (c) Measured ping times of a selected Tor entry node as a function of parallel requests (i.e., number of agents accessing the node).

the theory could be run by keeping N larger and simulating a Tor access rather than simple pings, we refrain from this in order to minimize the footprint of our experiments on the Tor network.

Received 11 September 2024; revised 14 January 2025; accepted 24 February 2025