Learning in Congestion Games with Bandit Feedback

Qiwen Cui[1]*
qwcui@cs.washington.edu

Zhihan Xiong[1]*
zhihanx@cs.washington.edu

Maryam Fazel[2]
mfazel@uw.edu

Simon S. Du[1]
ssdu@cs.washington.edu

[1] Paul G. Allen School of Computer Science & Engineering, University of Washington
[2] Department of Electrical & Computer Engineering, University of Washington

Abstract

In this paper, we investigate Nash-regret minimization in congestion games, a class of games with benign theoretical structure and broad real-world applications. We first propose a centralized algorithm based on the optimism in the face of uncertainty principle for congestion games with (semi-)bandit feedback, and obtain finite-sample guarantees. Then we propose a decentralized algorithm via a novel combination of the Frank-Wolfe method and G-optimal design. By exploiting the structure of the congestion game, we show the sample complexity of both algorithms depends only polynomially on the number of players and the number of facilities, but not the size of the action set, which can be exponentially large in terms of the number of facilities. We further define a new problem class, Markov congestion games, which allows us to model the non-stationarity in congestion games. We propose a centralized algorithm for Markov congestion games, whose sample complexity again has only polynomial dependence on all relevant problem parameters, but not the size of the action set.

1 Introduction

Nash equilibrium (NE) is a widely adopted concept in game theory community, used to describe the behavior of multi-agent systems with selfish players [Roughgarden, 2010]. At the Nash equilibrium, no player has the incentive to change its own strategy unilaterally, which implies it is a steady state of the game dynamics. For a general-sum game, computing the Nash equilibrium is PPAD-hard [Daskalakis, 2013] and the query complexity is exponential in the number of players [Rubinstein, 2016]. To help address these issues, a natural approach is to consider games with special structures. In this paper, we focus on congestion games.

Congestion games are general-sum games with *facilities* (resources) shared among players [Rosenthal, 1973]. During the game, each player will decide what combination of facilities to utilize, and popular facilities will become congested, which results in a possibly higher cost on each user. One example of congestion game is the routing game [Fotakis et al., 2002], where each player needs to travel from a given starting point to a destination point through some shared routes. These routes are represented as a traffic graph and the facilities are the edges. Each player will decide her path to go, and the more players use the same edge, the longer the edge travel time will be. Congestion games also have wide applications in electrical grids [Ibars et al., 2010], internet routing [Al-Kashoash et al., 2017] and

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

rate allocation [Johari and Tsitsiklis, 2004]. In many real-world scenarios, players can only have (semi-)bandit feedback, i.e., players know only the payoff of the facilities they choose. This kind of learning under uncertainty has been widely studied in bandits and in reinforcement learning for the single-agent setting, while theoretical understanding for the multi-agent case is still largely missing.

There are two types of algorithms in multi-agent systems, namely centralized algorithms and decentralized algorithms. For centralized algorithms, there exists a central authority that can control and receive feedback from all players in the game. As we have global coordination, centralized algorithms usually have favorable performance. On the other hand, such a central authority may not always be available in practice, and thus people turn to decentralized algorithms, i.e., each player makes decisions individually and can only observe her own feedback. However, decentralized algorithms are vulnerable to *nonstationarity* because each player is making decisions in a nonstationary environment as others' strategies are changing [Zhang et al., 2021a]. In this paper, we will study both centralized and decentralized algorithms in congestion games with bandit feedback, and we will provide motivating scenarios for both algorithms in Section 1.2.

The main challenge in designing algorithms for m-player congestion games with bandit feedback is the curse of exponential action set, i.e., the number of actions can be exponential in the number of facilities F because every subset of facilities can be an action. As a result, an efficient algorithm should have sample complexity polynomial in m and F and has no dependence on the size of the action space. One closely related type of general-sum game is the potential game, in which each individual's payoff changes, resulting from strategy modification, can be quantified by a common potential function. It is well-known that all congestion games are potential games, and each potential game has an equivalent congestion game formulation [Monderer and Shapley, 1996]. However, existing algorithms designed for potential games all have sample complexity scaling at least linearly in the number of actions [Leonardos et al., 2021, Ding et al., 2022], which is inefficient for congestion games. This motivates the following question:

Can we design provably sample-efficient centralized and decentralized learning algorithms for congestion games with bandit feedback?

We provide an affirmative answer to this question. To be precise, we use Nash-regret minimization (formally defined in Section 3) as our objective for learning in congestion games. This regret-like objective commonly appears in the literature of online learning and reinforcement learning [Orabona, 2019, Ding et al., 2022, Liu et al., 2021], which focuses on finite-time analysis and accumulative rewards throughout the learning process instead of the asymptotic behavior. In general, a sublinear Nash regret implies a best-iterate convergence, meaning that the algorithm has reached the approximate Nash equilibrium at least once, while the converse does not hold.

We highlight our contributions below and compare our results with previous algorithms in Table 1. Our algorithms are shaded and we prove sublinear Nash regrets for all of them. In Table 1, sample complexity refers to the number of samples required to reach best-iterate convergence to an ϵ-approximate Nash equilibrium and the results are obtained by standard online-to-batch conversion as in Section 3.1 of [Jin et al., 2018].

1.1 Main Novelties and Contributions

1. Centralized algorithm for congestion game. We adapt the principle of optimism in the face of uncertainty in stochastic bandits to ensure sufficient exploration in congestion games. We begin with congestion games with semi-bandit feedback, in which each player can observe the reward of every facility in the action. Instead of estimating the action reward as in stochastic multi-armed bandits, we estimate the facility rewards directly, which *removes the dependence on the size of action space*. Furthermore, we consider congestion games with bandit feedback, in which each player can only observe the overall reward. In this setting, we borrow ideas from linear bandits to estimate the reward function and analyze the algorithm. The algorithm is provably sample efficient in both cases.

2. Decentralized algorithm for congestion game. Our decentralized algorithm is a Frank-Wolfe method with exploration, in which each player only observes her own actions and rewards. To efficiently explore in the congestion game, we utilize G-optimal design allocation for bandit feedback and a specific distribution for semi-bandit feedback. As a result, the sample complexity does not depend on the number of actions. In addition, the L_1 smoothness parameter of the potential function does not depend on the number of actions, which is exploited by the Frank-Wolfe method. With the

Table 1: Comparison of algorithms for congestion games in terms of sample complexity and Nash regret, where "IPPG" stands for "independent projected policy gradient", "IPGA" stands for "independent policy gradient ascent", "I" represents the setting of semi-bandit feedback and "II" represents the setting of bandit feedback. Bandit feedback is assumed for algorithms from previous work. Here, A_i is the size of player i's action space, m is the number of players, $A_{\max} = \max_{i\in[m]} A_i$, F is the number of facilities and T is the number of samples collected. Our algorithms are shaded.

Algorithms	Sample complexity	Nash regret	Decentralized
Nash-VI [Liu et al., 2021]	$(\prod_{i=1}^m A_i)F/\epsilon^2$	$\sqrt{(\prod_{i=1}^m A_i)FT}$	No
V-learning [Jin et al., 2021a]	$A_{\max}F/\epsilon^2$ (CCE)	NA	Yes
IPPG [Leonardos et al., 2021]	$A_{\max}mF/\epsilon^6$	NA	Yes
IPGA [Ding et al., 2022]	$A_{\max}^2 m^3F^5/\epsilon^5$	$mF^{4/3}\sqrt{A_{\max}}T^{4/5}$	Yes
Nash-UCB I	mF^2/ϵ^2	$F\sqrt{mT}$	No
Nash-UCB II	m^2F^3/ϵ^2	$mF^{3/2}\sqrt{T}$	No
Frank-Wolfe with Exploration I	$m^{12}F^9/\epsilon^6$	$m^2F^{3/2}T^{5/6}$	Yes
Frank-Wolfe with Exploration II	$m^{12}F^{12}/\epsilon^6$	$m^2F^2T^{5/6}$	Yes

help of these two specific algorithmic designs for congestion games, we give the first decentralized algorithm for both semi-bandit feedback and bandit feedback that has no dependence on the size of the action space in congestion games.

3. Centralized algorithm for independent Markov congestion game. We extend the formulation of congestion game into a Markov setting and propose the independent Markov congestion game (IMCG), in which each facility has its own internal state and state transition happens independently among all the facilities. In Section 1.2, we give some examples that fit in this model. By utilizing techniques from factored MDPs, we extend our centralized algorithms for congestion games to efficiently solve IMCGs, with both semi-bandit and bandit feedback.

1.2 Motivating Examples

We provide an exmple here to motivate our proposed models. See Section 3 for the formal definition of (semi-)bandit feedback and (Markov) congestion games and Appendix A for additional examples.

Example 1 (**Routing Games**). For a routing game, there are multiple players in a traffic graph travelling from starting points to destination points, and the facilities are the edges (roads). The cost of each edge is the waiting time, which depends on the number of players using that edge.

• **Centralized algorithm for routing games:** Imagine each player is using Google Maps to navigate. Then Google Maps can serve as a center that knows the starting points and the destination points, as well as the real-time feedback of the waiting time on each edge of all the players. Google Maps itself also has the incentive to assign paths according to the Nash equilibrium strategy as then each player will find out that deviating from the navigation has no benefit and thus sticks to the app.

• **Decentralized algorithm for routing games:** Consider the case where players are still using Google Maps but due to privacy concerns or limited bandwidth, they only use the offline version, which has access only to the information of each single user. Then Google Maps needs to use decentralized algorithms so that it can still assign Nash equilibrium strategy to each user after repeated plays.

• **Markov routing games:** For Markov routing games, the time cost on each edge will change among different timesteps, which is a more accurate model of the real-world. For instance, some roads are prone to car accidents, which will result in an increasing cost on the next timestep, and the chance of accidents also depends on the number of players using that edge currently. This is modeled by the Markovian facility state transition in independent Markov congestion games.

2 Related Work

Potential Games. Potential games are general-sum games that admit a common potential function to quantify the changes in individual's payoff [Monderer and Shapley, 1996]. Algorithmic game theory community has studied how different dynamics converge to the Nash equilibium, e.g., best

response dynamics [Durand, 2018, Swenson et al., 2018] and no-regret dynamics [Heliou et al., 2017, Cheung and Piliouras, 2020], while usually they provide only asymptotic convergence, with either full information setting or bandit feedback setting. Recently, reinforcement learning community studied Markov potential games with bandit feedback, which can be applied to standard potential games. See the Markov Games part below for more details.

Congestion Games. Congestion games are developed in the seminal work [Rosenthal, 1973], and later Monderer and Shapley [1996] builds a close connection between congestion games and potential games. Congestion games are divided into atomic and non-atomic congestion games depending on whether each player is separable. Many papers consider non-atomic congestion games with non-decreasing cost function, which implies a convex potential function [Roughgarden and Tardos, 2004]. We consider the more difficult atomic congestion game where the potential function can be non-convex. For online non-atomic case, [Krichene et al., 2015] considers partial information setting while they provide convergence in the sense of Cesaro means. [Kleinberg et al., 2009, Krichene et al., 2014] show that some no-regret online learning algorithms asymptotically converges to Nash equilibrium. [Chen and Lu, 2015, 2016] are two closely related works that consider bandit feedback in atomic congestion games and provide non-asymptotic convergence. However, they still assume a convex potential function and the sample complexity has exponential dependence on the number of facilities, which is far from ideal.

Markov Games. Markov games are widely studied since the seminal work [Shapley, 1953]. Recently, the topic has received much attention due to advances in reinforcement learning theory. Liu et al. [2021] provides a centralized algorithm for learning the Nash equilibrium in general-sum Markov games, and [Jin et al., 2021a, Song et al., 2021] provide decentralized algorithms for learning the (coarse) correlated equilibrium. One closely related line of research is on Markov potential games [Leonardos et al., 2021, Zhang et al., 2021b, Fox et al., 2021, Cen et al., 2022, Ding et al., 2022]. However, applying their algorithms to congestion games leads to explicit dependence on the number of actions, which would be exponentially worse than our algorithms. See Table 1 for comparisons. Our independent Markov congestion game is motivated by the state-based potential games studied in Marden [2012] and Macua et al. [2018], and its transition kernel is closely related to the factored MDPs, for which single agent algorithms are studied in [Osband and Van Roy, 2014, Chen et al., 2020, Xu and Tewari, 2020, Tian et al., 2020, Rosenberg and Mansour, 2021].

Learning in Games. Different from our paper, learning in games in traditional literature of game theory mainly considers players' asymptotic behavior [Leslie and Collins, 2005, Cominetti et al., 2010, Coucheney et al., 2015]. In early literature, Leslie [2004] investigates actor-critic learning and Q-learning algorithms in games with bandit feedback and their connection to best-response dynamics. Leslie and Collins [2005] proposes individual Q-learning algorithm and shows that it converges to the NE almost surely in two-player zero-sum game and Leslie and Collins [2006] studies learning the NE from the perspective of a fictitious play-like process. Later, Cominetti et al. [2010] considers payoff-based learning rules and shows convergence to NE in traffic games, while another payoff-based learning model for continuous games is developed in Bervoets et al. [2020]. Coucheney et al. [2015] derives a new penalty-regulated dynamics and proposes a corresponding learning algorithms that converges to NE in potential games with bandit feedback. Bravo et al. [2018] proposes that in monotone games with bandit feedback, as long as all players are using some no-regret learning algorithm, the dynamics will converge to the NE. An improved analysis of the same algorithm is then proposed in Drusvyatskiy and Ratliff [2021]. In contrast, our learning objective focuses on finite-time cumulative rewards, which is more widely used in current multi-agent reinforcement learning literature [Ding et al., 2022, Liu et al., 2021].

3 Preliminaries

General-sum Matrix Games. We consider the model of general-sum matrix games, defined by the tuple $\mathcal{G} = (\{\mathcal{A}_i\}_{i=1}^m, R)$, where m is the number of players, $\mathcal{A}_i$ is the action space of player i and $R(\cdot|\boldsymbol{a})$ is the reward distribution on $[0, r_{\max}]^m$ with mean $\boldsymbol{r}(\boldsymbol{a})$. Let $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be the whole action space and denote an element as $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A}$. After all players take actions $\boldsymbol{a} \in \mathcal{A}$, a reward vector is sampled $\boldsymbol{r} \sim R(\cdot|\boldsymbol{a})$ and player i will receive reward $r_i \in [0, r_{\max}]$ with mean $r_i(\boldsymbol{a})$. Each player's objective is to maximize her own reward.

A general policy π is defined as a vector in $\Delta(\mathcal{A})$, the probability simplex over the action space $\mathcal{A}$. A product policy $\pi = (\pi_1, \ldots, \pi_m)$ is defined as a tuple in $\Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_m)$, in which $\boldsymbol{a} = (a_1, \ldots, a_m) \sim \pi$ represents $a_i \overset{\text{i.i.d.}}{\sim} \pi_i$. The value of policy π for player i is $V_i^{\pi} = \mathbb{E}_{\boldsymbol{a}\sim\pi}[r_i(\boldsymbol{a})]$.

Nash Equilibrium and Nash Regret. Given a general policy π, let π_{-i} be the marginal joint policy of players $1, \ldots, i-1, i+1, \ldots, m$. Then, the best response of player i under policy π is $\pi_i^{\dagger} = \operatorname{argmax}_{\mu\in\Delta(\mathcal{A}_i)} V_i^{\mu,\pi_{-i}}$ and the corresponding value is $V_i^{\dagger,\pi_{-i}} := V_i^{\pi_i^{\dagger},\pi_{-i}}$. Our goal is to find the approximate Nash equilibrium of the matrix game, which is defined below.

Definition 1. A product policy π is an *ϵ-approximate Nash equilibrium* if $\max_i(V_i^{\dagger,\pi_{-i}} - V_i^{\pi}) \leq \epsilon$.

An ϵ-approximate Nash equilibrium can be obtained by achieving a sublinear Nash regret, which is defined below. See Section 3 in Ding et al. [2022] for a more detailed discussion.

Definition 2. With π^k being the policy at k-th episode, the *Nash regret* after K episodes is define as

$$\text{Nash-Regret}(K) = \sum_{k=1}^{K} \max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k} \right).$$

Remark 1. Here, if we replace $\max_{i\in[m]}$ by $\sum_{i=1}^m$ in the definition of Nash regret, the single-step Nash regret at episode k will become the Nikaido-Isoda (NI) function evaluated at π^k, which is a popular objective for equilibrium computation [Nikaidô and Isoda, 1955, Raghunathan et al., 2019]. Replacing $\max_{i\in[m]}$ by $\sum_{i=1}^m$ will multiply our regret bounds by a factor of m, while our conclusion will not be affected.

Potential Games. A potential game is a general-sum game such that there exists a potential function $\Phi : \Delta(\mathcal{A}) \to [0, \Phi_{\max}]$ such that for any player $i \in [m]$ and policies π_i, π_i', π_{-i}, it satisfies

$$\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = V_i^{\pi_i,\pi_{-i}} - V_i^{\pi_i',\pi_{-i}}.$$

We can immediately see that a policy that maximizes the potential function is a Nash equilibrium.

Congestion Games. A congestion game is defined by $\mathcal{G} = (\mathcal{F}, \{\mathcal{A}_i\}_{i=1}^m, \{R^f\}_{f\in\mathcal{F}})$, where $\mathcal{F} = [F]$ is called the facility set and $R^f(\cdot|n) \in [0,1]$ is the reward distribution for facility f with mean $r^f(n)$, where $n \in [m]$. Each action $a_i \in \mathcal{A}_i$ is a subset of $\mathcal{F}$ (i.e., $a_i \subseteq \mathcal{F}$). Suppose the joint action chosen by all the players is $\boldsymbol{a} \in \mathcal{A}$, then a random reward is sampled $r^f \sim R^f(\cdot|n^f(\boldsymbol{a}))$ for each facility f, where $n^f(\boldsymbol{a}) = \sum_{i=1}^m \mathbb{1}\{f \in a_i\}$ is the number of players using facility f. The reward collected by player i is $r_i = \sum_{f\in a_i} r^f$ with mean $r_i(\boldsymbol{a}) = \sum_{f\in a_i} r^f(n^f(\boldsymbol{a})) \in [0, F]$.

Connection to Potential Games [Monderer and Shapley, 1996]. As a special class of potential game, all congestion games have the potential function: $\Phi(\boldsymbol{a}) = \sum_{f\in\mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} r^f(i)$. To see this, we can easily verify that $\Phi(a_i, a_{-i}) - \Phi(a_i', a_{-i}) = r_i(a_i, a_{-i}) - r_i(a_i', a_{-i})$ holds. Then, by defining $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a}\sim\pi}[\Phi(\boldsymbol{a})]$, we can have $\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = V_i^{\pi_i,\pi_{-i}} - V_i^{\pi_i',\pi_{-i}}$.

Types of feedback. There are in general two types of reward feedback for the congestion games, semi-bandit feedback and bandit feedback, both of which are reasonable under different scenarios. In semi-bandit feedback, after taking the action, player i will receive reward information r^f for each $f \in a_i$; in bandit feedback, after taking the action, player i will only receive the reward $r_i = \sum_{f\in a_i} r^f$ with no knowledge about each r^f. In this paper, we will address both of them, with more focus on the bandit feedback, which can be directly generalized to semi-bandit feedback.

4 Centralized Algorithms for Congestion Games

In this section, we introduce two centralized algorithms for congestion games – one for the semi-bandit feedback and one for the bandit feedback. We will see that both of them can achieve sublinear Nash regret with polynomial dependence on both m and F.

4.1 Algorithm for Semi-bandit Feedback

Summarized in Algorithm 1, Nash upper confidence bound (Nash-UCB) for congestion games is developed based on optimism in the face of uncertainty. In particular, the algorithm estimates the

reward matrices optimistically in line 4, computes its Nash equilibrium policy in line 5 and then follows this policy.

For convenience, we define the empirical counter $N^{k,f}(n) = \sum_{k'=1}^{k} \mathbb{1}\left\{n^f(\boldsymbol{a}^{k'}) = n\right\}$ and $\tilde{\iota} = 2\log(4(m+1)K/\delta)$. Then, the reward estimator for f and the bonus term are defined as

$$\hat{r}^{k,f}(n) = \frac{\sum_{k'=1}^{k} r^{k',f}\mathbb{1}\left\{n^f(\boldsymbol{a}^{k'}) = n\right\}}{N^{k,f}(n) \vee 1}, \quad b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \sum_{f \in a_i} \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a})) \vee 1}}, \tag{1}$$

where $r^{k,f} \in [0,1]$ is the random reward realization of $r^f(n^f(\boldsymbol{a}^k))$. Naturally, the reward estimator for player i is $\hat{r}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} \hat{r}^{k,f}(n^f(\boldsymbol{a}))$.

Algorithm 1 Nash-UCB for Congestion Games

1: **Input:** ϵ, accuracy parameter for Nash equilibrium computation
2: **for** episode $k = 1, \ldots, K$ **do**
3: **for** player $i = 1, \ldots, m$ **do**
4: $\overline{Q}_i^k(\boldsymbol{a}) \leftarrow \hat{r}_i^k(\boldsymbol{a}) + b_i^{k,\mathrm{r}}(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathcal{A}$
5: $\pi^k \leftarrow \epsilon$-NASH$(\overline{Q}_1^k(\cdot), \cdots, \overline{Q}_m^k(\cdot))$ (Algorithm 2)
6: Take action $\boldsymbol{a}^k \sim \pi^k$ and observe reward $r^{k,f}$
7: Update reward estimators $\hat{r}_i^k$ and bonus term $b_i^{k,\mathrm{r}}$

Algorithm 1 is motivated by the Nash-VI algorithm in [Liu et al., 2021] plus a deliberate utilization of the special reward structure in the congestion games. Moreover, notice that a matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ forms a potential game (see Lemma 1). As a result, in line 5, we can *efficiently compute* the ϵ-approximate Nash equilibrium π^k for that matrix game by utilizing Algorithm 2, (see Lemma 2). It is a simple greedy algorithm such that in each round, it modifies one player's policy whose modification can increase the potential function most. In addition, Algorithm 2 always outputs a deterministic product policy.

Algorithm 2 ϵ-approximate Nash Equilibrium for Potential Games

1: **Input:** ϵ, accuracy parameter; full information potential game $(\{\mathcal{A}_i\}_{i=1}^m, \{r_i\}_{i=1}^m)$ such that $r_i \in [0, r_{\max}]$ for all $i \in [m]$
2: **Initialize:** $\pi^1 = \boldsymbol{a}^1$, arbitrary deterministic product policy
3: **for** round $k = 1, \ldots, \left\lceil \frac{m r_{\max}}{\epsilon} \right\rceil$ **do**
4: **for** player $i = 1, \ldots, m$ **do**
5: $\Delta_i = \max_{a_i \in \mathcal{A}_i} r_i(a_i, \pi_{-i}^k) - r_i(\pi^k)$
6: $a_i^{k+1} = \mathrm{argmax}_{a \in \mathcal{A}_i} r_i(a_i, \pi_{-i}^k) - r_i(\pi^k)$
7: **if** $\max_{i \in [m]} \Delta_i \leq \epsilon$ **then**
8: **return** π^k
9: $j = \mathrm{argmax}_{i \in [m]} \Delta_i$
10: $\pi^{k+1}(j) = a_j^{k+1}$, $\pi^{k+1}(i) = \pi^k(i)$, for all $i \neq j$

4.2 Algorithm for Bandit Feedback

When the players can only receive bandit feedback, estimating $\hat{r}^{k,f}$ directly for each $f \in \mathcal{F}$ is no longer feasible. However, notice that the reward function $r_i(\boldsymbol{a}) = \sum_{f \in a_i} r^f(n^f(\boldsymbol{a}))$ can be seen as an inner product between vectors characterized by action $\boldsymbol{a}$ and reward function $r^f(\cdot)$. Therefore, under bandit feedback, we can treat it as a linear bandit and use ridge regression to build the reward estimator $\tilde{r}_i^k$ and corresponding bonus term $\tilde{b}^{k,\mathrm{r}}$, whose index i is dropped since it is the same for all players. The new algorithm will use these two terms to replace $\hat{r}_i^k$ and $b_i^{k,\mathrm{r}}$ in line 4 of Algorithm 1.

In particular, define $\theta \in [0,1]^{\tilde{d}}$ with $\tilde{d} = mF$ to be the vector such that $r^f(n) = \theta_{n+m(f-1)}$. Meanwhile, for player $i \in [m]$, define $A_i : \mathcal{A} \mapsto \{0,1\}^{\tilde{d}}$ to be the vector-valued function such that

$$[A_i(\boldsymbol{a})]_j = \mathbb{1}\left\{j = n + m(f-1), f \in a_i, n = n^f(\boldsymbol{a})\right\}.$$

In other words, $A_i(\boldsymbol{a})$ is a 0-1 vector with element 1 only at indices corresponding to those in θ that represents $r^f(n)$ for $f \in a_i$ and $n = n^f(\boldsymbol{a})$. Now, with these definitions, the reward function can be written as $r_i(\boldsymbol{a}) = \langle A_i(\boldsymbol{a}), \theta \rangle$. Then, we build the reward estimator and the bonus term through ridge regression and corresponding confidence bound, which are defined as the following:

$$\tilde{r}_i^k(\boldsymbol{a}) = \left\langle A_i(\boldsymbol{a}), \widehat{\theta}^k \right\rangle, \quad \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}) = \max_{i \in [m]} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k}, \tag{2}$$

where $\widehat{\theta}^k = \left(V^k\right)^{-1} \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(\boldsymbol{a}^{k'}) r_i^{k'}$, $V^k = I + \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(\boldsymbol{a}^{k'}) A_i(\boldsymbol{a}^{k'})^\top$ and $\sqrt{\tilde{\beta}_k} = \sqrt{\tilde{d}} + \sqrt{\tilde{d} \log\left(1 + \frac{mkF}{\tilde{d}}\right) + \tilde{\iota}}$. Note that we cannot bound the sum of this bonus terms by directly applying the elliptical potential lemma. We instead prove its variant in Lemma 4.

4.3 Regret Analysis

The Nash regret bounds for the two versions of Algorithm 1 are formally presented in Theorem 1. The proof details are deferred to Appendix D.

Theorem 1. *Let $\epsilon = 1/K$. For congestion games with semi-bandit feedback, by running Algorithm 1 with reward estimator and bonus term in (1), with probability at least $1-\delta$, we can achieve that*

$$\textit{Nash-Regret}(K) \leq \widetilde{\mathcal{O}}\left(F\sqrt{mK}\right).$$

Furthermore, if we only have bandit feedback, then by running Algorithm 1 with reward estimator and bonus term in (2), with probability at least $1-\delta$, we can achieve that

$$\textit{Nash-Regret}(K) \leq \widetilde{\mathcal{O}}\left(mF^{3/2}\sqrt{K}\right).$$

Remark 2. Since each action is a subset of $\mathcal{F}$, the size of each player's action space can be 2^F. As a result, directly applying Nash-VI in [Liu et al., 2021] leads to a regret bound exponential in F.

Remark 3. Note that we assume $r^f \in [0,1]$, which implies $r_i \in [0,F]$ for each player $i \in [m]$.

5 Decentralized Algorithms for Congestion Games

In this section, we present a decentralized algorithm for congestion games. Due to limited space, we only introduce the version of bandit feedback as in Section 4.2. The algorithmic details for the semi-bandit feedback setting are deferred into Appendix E.3. We will show that under both settings, even though each player can only observe her own actions and rewards, our decentralized algorithm still enjoys sublinear Nash regret with polynomial dependence on m and F.

We first define the vector-valued function $\phi_i : \mathcal{A}_i \mapsto \{0,1\}^{F_i}$ to be the feature map of player i such that $[\phi_i(a_i)]_f = \mathbb{1}\{f \in a_i\}$ for $a_i \in \mathcal{A}_i$ and $f \in \bigcup_{a_i \in \mathcal{A}_i} a_i$. Here, F_i is the size of $\bigcup_{a_i \in \mathcal{A}_i} a_i \subseteq \mathcal{F}$ and we can immediately see that $F_i \leq F$ for any $i \in [m]$.

The core idea of our algorithm is that the Nash equilibrium can be found by reaching the stationary points of the potential function since all congestion games are potential games. Here, the UCB-like algorithms used in the centralized setting are not applicable because their policy computation requires value functions for all players (e.g., line 5 of Algorithm 1), which are not available in the decentralized setting. Summarized in Algorithm 3, the decentralized algorithm is developed based on the Frank-Wolfe method and has the following three major components.

Gradient Estimator. In line 7, the algorithm builds the estimator $\widehat{\nabla}_i^k \Phi$ defined in (4) by using the τ reward samples collected from line 5. Here, $\widehat{\nabla}_i^k \Phi$ estimates the gradient of potential function Φ with respect to the policy π_i^k. Recall that for a congestion game, we have $\Phi(\boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} r^f(i)$

Algorithm 3 Frank-Wolfe with Exploration for Congestion Game

1: **Input:** γ, ν, mixture weights; π_i^1, initial policy.
2: **Initialize:** ρ_i, the G-optimal design for player i, defined in (5).
3: **for** episode $k = 1, \cdots, K$ **do**
4: **for** round $t = 1, \cdots, \tau$ **do**
5: Each player takes action $a_i^{k,t} \sim \pi_i^k$, observes reward $r_i^{k,t}$.
6: **for** player $i = 1, \cdots, m$ **do**
7: Compute $\widehat{\nabla}_i^k \Phi(a_i)$ by the formula in (4) for all $a_i \in \mathcal{A}_i$
8: Compute $\widetilde{\pi}_i^{k+1} \leftarrow \operatorname{argmax}_{\pi_i \in \Delta(\mathcal{A}_i)} \left\langle \pi_i, \widehat{\nabla}_i^k \Phi \right\rangle$
9: Update $\pi_i^{k+1} \leftarrow (1-\gamma)(\nu \widetilde{\pi}_i^{k+1} + (1-\nu)\pi_i^k) + \gamma \rho_i$

and $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a} \sim \pi}[\Phi(\boldsymbol{a})]$. Then we can define $\nabla_i \Phi := \nabla_{\pi_i} \Phi$ as a vector of dimension $|\mathcal{A}_i|$. For the component indexed by some $a_i \in \mathcal{A}_i$, we can see that $\Phi(\pi) = \pi_i(a_i) \mathbb{E}_{a_{-i} \sim \pi_{-i}}[r_i(a_i, a_{-i})] + \text{const}$, where const does not depend on $\pi_i(a_i)$. Therefore, we have

$$\nabla_i \Phi(a_i) = \mathbb{E}_{a_{-i} \sim \pi_{-i}}[r_i(a_i, a_{-i})] = \mathbb{E}_{a_{-i} \sim \pi_{-i}} \left[\sum_{f \in a_i} r^f(n^f(a_i, a_{-i})) \right] = \langle \phi_i(a_i), \theta_i(\pi) \rangle, \quad (3)$$

where $[\theta_i(\pi)]_f = \mathbb{E}_{a_{-i} \sim \pi_{-i}} \left[r^f(n^f(a_{-i}) + 1) \right]$. Meanwhile, the mean of the t-th reward that player i received at episode k satisfies

$$\mathbb{E}\left[r_i^{k,t} \mid \boldsymbol{a}^{k,t} \right] = r_i(\boldsymbol{a}^{k,t}) = \sum_{f \in a_i^{k,t}} r^f(n^f(\boldsymbol{a}^{k,t})) = \left\langle \phi_i(a_i^{k,t}), \theta_i^{k,t}(a_{-i}^{k,t}) \right\rangle,$$

where $[\theta_i^{k,t}(a_{-i}^{k,t})]_f = r^f(n^f(a_{-i}^{k,t}) + 1)$ and its mean is $[\theta_i(\pi^k)]_f$. Therefore, we can use linear regression to estimate $\theta_i(\pi^k)$. In particular, we have $\widehat{\theta}_i^k(\pi^k) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left(\Sigma_i^k\right)^{-1} \phi_i(a_i^{k,t}) r_i^{k,t}$, with the covariance matrix $\Sigma_i^k = \mathbb{E}_{a_i \sim \pi_i^k} \left[\phi_i(a_i) \phi_i(a_i)^\top \right]$. Then, we have the unbiased gradient estimate

$$\widehat{\nabla}_i^k \Phi(a_i) = \left\langle \phi_i(a_i), \widehat{\theta}_i^k(\pi^k) \right\rangle = \frac{1}{\tau} \sum_{t=1}^{\tau} \phi_i(a_i)^\top \left(\Sigma_i^k\right)^{-1} \phi_i(a_i^{k,t}) r_i^{k,t}. \quad (4)$$

Remark 4. One difference between Algorithm 3 (decentralized) and Algorithm 1 (centralized) is that in the decentralized algorithm, each player is required to play the same policy for τ times before an update can be applied. An episode is thus defined for convenience as the time period during which the players' policies are fixed. We make this artificial design mainly for controlling the variance of the gradient estimator $\widehat{\nabla}_i^k \Phi(a_i)$. However, we conjecture that with more careful design and analysis, it should be possible to improve Algorithm 3 so that only one sample is required per episode [Zhang et al., 2020].

G-optimal Design. In line 8 and 9, the algorithm performs standard Frank-Wolfe update and mixes the updated policy with an exploration policy ρ_i, which is defined as the G-optimal allocation for features $\{\phi_i(a_i)\}_{a_i \in \mathcal{A}_i}$. To be specific, we have

$$\rho_i = \operatorname*{argmin}_{\lambda \in \Delta(\mathcal{A}_i)} \max_{a_i \in \mathcal{A}_i} \|\phi_i(a_i)\|^2_{\mathbb{E}_{a_i' \sim \lambda}\left[\phi_i(a_i')\phi_i(a_i')^\top\right]^{-1}}. \quad (5)$$

Here ρ_i guarantees that Σ_i^k is invertible and the variance of $\widehat{\nabla}_i^k \Phi(a_i) = \left\langle \phi_i(a_i), \widehat{\theta}_i^k(\pi^k) \right\rangle$ depends only on F instead of the size of action space (Lemma 9) because by Kiefer-Wolfowitz theorem, we have $\max_{a_i \in \mathcal{A}_i} \|\phi_i(a_i)\|^2_{\mathbb{E}_{a_i' \sim \rho_i}\left[\phi_i(a_i')\phi_i(a_i')^\top\right]^{-1}} = F_i \leq F$ [Lattimore and Szepesvári, 2020].

Frank-Wolfe Update. Finally, we emphasize that it is crucial to use Frank-Wolfe update because it is compatible with L_1 *norm* and we can show that Φ is mF-smooth with respect to the L_1 norm (Lemma 11). In contrast, its smoothness for L_2 norm will depend on the size of the action space.

Before the game starts, each player i can compute her ρ_i based on her own action set $\mathcal{A}_i$. During the game, all players only have access to their own actions and rewards, which means that Algorithm 3 is fully decentralized. The Nash regret bound for this algorithm is formally stated in Theorem 2 and the proof details are given in Appendix E.1 and E.2.

Theorem 2. *Let $T = K\tau$. For congestion game with bandit feedback, by running Algorithm 3 with gradient estimator $\widehat{\nabla}_i^k \Phi$ and exploration distribution ρ_i defined in (4), (5), with probability at least $1-\delta$, we have*

$$\textit{Nash-Regret}(T) := \sum_{k=1}^{K} \tau \max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k} \right) \leq \widetilde{\mathcal{O}} \left(m^2 F^2 T^{5/6} + m^3 F^3 T^{2/3} \right).$$

For congestion game with semi-bandit feedback, by running Algorithm 3 with gradient estimator $\widetilde{\nabla}_i^k \Phi(a_i)$ and exploration distribution $\tilde{\rho}_i$ in Appendix E.3, with probability at least $1-\delta$, we have

$$\textit{Nash-Regret}(T) \leq \widetilde{\mathcal{O}} \left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3} \right).$$

6 Extension to Independent Markov Congestion Games

In this section, we propose and analyze a Markov extension of the congestion games, called the independent Markov congestion games (IMCGs).

6.1 Problem Formulation

General-sum Markov Games. A finite-horizon time-inhomogeneous tabular general-sum Markov game is defined by $\mathcal{M} = \{\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, H, P, R, s_0\}$, where $\mathcal{S}$ is the state space, m is the number of players, $\mathcal{A}_i$ is the action space of player i, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ is the whole action space, H is the time horizon, s_0 is the initial state[1], $P = (P_1, P_2, \cdots, P_H)$ with $P_h \in [0,1]^{S\times A\times S}$ as the transition kernel at timestep h, $R = \{R_h(\cdot|s_h, \boldsymbol{a}_h)\}_{h=1}^H$ with $R_h(\cdot|s_h, \boldsymbol{a}_h)$ as the reward distribution on $[0, r_{\max}]^m$ with mean $\boldsymbol{r}_h(s_h, a_h) \in [0, r_{\max}]^m$ at timestep $h \in [H]$. At timestep h, all players choose their actions simultaneously and a reward vector is sampled $\boldsymbol{r}_h \sim R_h(\cdot|s_h, \boldsymbol{a}_h)$, where s_h is the current state and $\boldsymbol{a}_h = (a_{h,1}, a_{h,2}, \cdots, a_{h,m})$ is the joint action. Each player i receives reward $r_{h,i}$ and the state transits to $s_{h+1} \sim P_h(\cdot|s_h, \boldsymbol{a}_h)$. The objective for each player is to maximize her own total reward. We assume that the initial state s_1 is fixed.

A (Markov) policy π is a collection of H functions $\{\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h=1}^H$, each of which maps a state to a distribution over the action space. π is a product policy if $\pi_h(\cdot \mid s)$ is a product policy for each $(h, s) \in [H] \times \mathcal{S}$. The value function and Q-value function of player i at timestep h under policy π are defined as

$$V_{h,i}^{\pi}(s) = \mathbb{E}_{\pi} \left[\sum_{h'=h}^{H} r_{h',i}(s_{h'}, \boldsymbol{a}_{h'}) \mid s_h = s \right], \; Q_{h,i}^{\pi}(s, \boldsymbol{a}) = \mathbb{E}_{\pi} \left[\sum_{h'=h}^{H} r_{h',i}(s_{h'}, \boldsymbol{a}_{h'}) \mid s_h = s, \boldsymbol{a}_h = a \right].$$

The best responses and Nash regret can be defined similarly as those for matrix games. In particular, given a policy π, player i's best response policy is $\pi_{h,i}^{\dagger}(\cdot \mid s) = \operatorname{argmax}_{\mu\in\Delta(\mathcal{A}_i)} V_{h,i}^{\mu,\pi_{-i}}(s)$ and the corresponding value function is denoted as $V_{h,i}^{\dagger,\pi_{-i}}$.

Definition 3. With π^k being the policy at kth episode, the *Nash regret* after K episodes is define as

$$\text{Nash-Regret}(K) = \sum_{k=1}^{K} \max_{i\in[m]} \left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k} \right)(s_1).$$

Independent Markov Congestion Game. A general-sum Markov game is an independent Markov congestion game (IMCG) if there exists a facility set $\mathcal{F}$ such that $a_i \subseteq \mathcal{F}$ for any $a_i \in \mathcal{A}_i$, a state space $\mathcal{S} = \prod_{f\in\mathcal{F}} \mathcal{S}^f$, a set of facility reward distributions $\{R_h^f\}_{h\in[H], f\in\mathcal{F}}$ such that if the joint action at s_h is $\boldsymbol{a}$, we have $r_{h,i} = \sum_{f\in a_i} r_h^f$, where $r_h^f \sim R_h^f(\cdot|s_h, n^f(\boldsymbol{a}))$ with support on $[0,1]$

[1] An episode is defined as running H steps from the initial state s_0, which is common for the episodic MDP.

and mean $r_h^f(s_h, n^f(\boldsymbol{a}))$, and a set of transition matrices $\{P_h^f\}_{h\in[H],f\in\mathcal{F}}$ such that $P_h(s'|s,\boldsymbol{a}) = \prod_{f\in\mathcal{F}} P_h^f(s'^f|s^f, n^f(\boldsymbol{a}))$. In other words, at each timestep h and state $s \in \mathcal{S}$, the players are in a congestion game. Meanwhile, each facility has its own state and independent state transition, which only depends on its current state and number of players using that facility. This transition kernel can be viewed as a special case of that in factored MDPs [Szita and Lőrincz, 2009]. The IMCG also admits two types of feedback, semi-bandit feedback and bandit feedback, just like the congestion game. In this paper, we will consider both types of feedback.

6.2 Theoretical Guarantee

Summarized in Algorithm 4, our centralized algorithm for IMCGs is naturally extended from the Nash-UCB (Algorithm 1) by incorporating transition kernel estimators, corresponding bonus terms and Bellman backward update. The key idea is to utilize the independent transition structure to remove the dependence on the exponential size of the state space $S = \prod_{f\in\mathcal{F}} S^f$. We tackle this issue by adapting technique from factored MDP [Chen et al., 2020]. The algorithmic details for both types of feedback are deferred into Appendix C. The Nash regret bounds for the two versions of Algorithm 4 are stated in Theorem 3 and the proof details are deferred to Appendix F.

Theorem 3. *For independent Markov congestion game with semi-bandit feedback, by running the centralized Algorithm 4, with probability at least $1-\delta$, we can achieve that*

$$\textit{Nash-Regret}(K) \le \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} FS^f\sqrt{mH^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).$$

Furthermore, if we only have bandit feedback, then by running Algorithm 4 with reward estimator and bonus term in (11) and (12), with probability at least $1-\delta$, we can achieve that

$$\textit{Nash-Regret}(K) \le \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} FS^f\sqrt{m^2H^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).$$

The regret bound in [Liu et al., 2021] is $\widetilde{O}(\sqrt{H^3S^2(\Pi_{i=1}^m A_i)T})$, where both A_i and $S = \prod_{f\in\mathcal{F}} S^f$ can be exponential in F. Our bounds have polynomial dependence on all the parameters.

7 Conclusion

In this paper, we study sample-efficient learning in congestion games by utilizing the special reward structure. We propose both centralized and decentralized algorithms for congestion games with two types of feedback, all achieving sample complexities only polynomial in the number of facilities. To the best of our knowledge, each one of them is the first sample-efficient learning algorithm for congestion games in its own setting. We further define the independent Markov congestion game (IMCG) as a natural extension of the congestion game into the Markov setting together with a sample-efficient centralized algorithm for both types of feedback.

One promising future direction is to find a sample-efficient decentralized algorithm such that from each player's own perspective, the algorithm is still no-regret. In other words, diminishing regret is guaranteed for the player by running this algorithm even though other players may use policies from different algorithms. Another important future direction is to find sample-efficient centralized/decentralized algorithms that can explicitly find an approximate Nash equilibrium policy.

Acknowledgements

This work was supported in part by NSF TRIPODS II-DMS 2023166, NSF CCF 2007036, NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844.

References

Hayder AA Al-Kashoash, Maryam Hafeez, and Andrew H Kemp. Congestion control for 6lowpan networks: A game theoretic framework. *IEEE internet of things journal*, 4(3):760–771, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

Sebastian Bervoets, Mario Bravo, and Mathieu Faure. Learning with minimal information in continuous games. *Theoretical Economics*, 15(4):1471–1508, 2020.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. *arXiv preprint arXiv:2204.05466*, 2022.

Po-An Chen and Chi-Jen Lu. Playing congestion games with bandit feedbacks. In *AAMAS*, pages 1721–1722, 2015.

Po-An Chen and Chi-Jen Lu. Generalized mirror descents in congestion games. *Artificial Intelligence*, 241:217–243, 2016.

Xiaoyu Chen, Jiachen Hu, Lihong Li, and Liwei Wang. Efficient reinforcement learning in factored mdps with application to constrained rl. *arXiv preprint arXiv:2008.13319*, 2020.

Yun Kuen Cheung and Georgios Piliouras. Chaos, extremism and optimism: Volume analysis of learning in games. *Advances in Neural Information Processing Systems*, 33:9039–9049, 2020.

Roberto Cominetti, Emerson Melo, and Sylvain Sorin. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83, 2010.

Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence, 2022.

Dmitriy Drusvyatskiy and Lillian J Ratliff. Improved rates for derivative free play in convex games. *arXiv preprint arXiv:2111.09456*, 2021.

Stéphane Durand. *Analysis of Best Response Dynamics in Potential Games*. PhD thesis, Université Grenoble Alpes, 2018.

Dimitris Fotakis, Spyros Kontogiannis, Elias Koutsoupias, Marios Mavronicolas, and Paul Spirakis. The structure and complexity of nash equilibria for a selfish routing game. In *International Colloquium on Automata, Languages, and Programming*, pages 123–134. Springer, 2002.

Roy Fox, Stephen McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. *arXiv preprint arXiv:2110.10614*, 2021.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Christian Ibars, Monica Navarro, and Lorenza Giupponi. Distributed demand management in smart grid with a congestion game. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 495–500. IEEE, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021a.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent rl, 2021b.

Ramesh Johari and John N Tsitsiklis. Efficiency loss in a network resource allocation game. *Mathematics of Operations Research*, 29(3):407–435, 2004.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 533–542, 2009.

Walid Krichene, Benjamin Drighès, and Alexandre Bayen. On the convergence of no-regret learning in selfish routing. In *International Conference on Machine Learning*, pages 163–171. PMLR, 2014.

Walid Krichene, Benjamin Drighès, and Alexandre M Bayen. Online learning of nash equilibria in congestion games. *SIAM Journal on Control and Optimization*, 53(2):1056–1081, 2015.

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

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games, 2021.

David S Leslie. *Reinforcement learning in games*. PhD thesis, University of Bristol, 2004.

David S Leslie and Edmund J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

David S Leslie and Edmund J Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Jason R Marden. State based potential games. *Automatica*, 48(12):3075–3088, 2012.

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(S1):807–815, 1955.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Arvind Raghunathan, Anoop Cherian, and Devesh Jha. Game theoretic optimization via gradient-based nikaido-isoda function. In *International Conference on Machine Learning*, pages 5291–5300. PMLR, 2019.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored mdps with unknown structure. *Advances in Neural Information Processing Systems*, 34, 2021.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Tim Roughgarden and Éva Tardos. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and economic behavior*, 47(2):389–403, 2004.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265. IEEE, 2016.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Brian Swenson, Ryan Murray, and Soummya Kar. On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, 56(4):2734–2767, 2018.

István Szita and András Lőrincz. Optimistic initialization and greediness lead to polynomial time learning in factored mdps. In *Proceedings of the 26th annual international conference on machine learning*, pages 1001–1008, 2009.

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907, 2020.

Ziping Xu and Ambuj Tewari. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *Advances in Neural Information Processing Systems*, 33:18226–18236, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021a.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Motivating Examples

In this section, we present two additional motivating examples of our proposed models.

Example 2 (**Web Advertisements**). Consider a set of websites as the facility set and companies who want to advertise their products as the players. Due to budget constraints, each company may only choose some of these websites to put its product ad. For each website, the probability that a user will click on a certain ad (and then buy the product) depends on how many ads are put on the website. If a website receives too many ads, the probability that a user can see a certain ad will decrease, thus making it congested.[2] The reward each company will receive is measured by the amount of products sold during certain period of time, which is bandit feedback.

Example 3 (**Server Usage**). Consider a set of servers in a company as the facility set and server users as the players. Each user needs to request several servers to finish her computation task and the cost triggered from each server depends on the number of users requesting that server. Each user will try to minimize the total cost incurred from the servers she requested. As each user can see the cost from all the servers she requested, this is semi-bandit feedback.

B Compute ϵ-approximate Nash Equilibrium in Potential Games

In this section, we show that the ϵ-NASH$(\cdot)$ operation in Algorithm 1 can be computed efficiently by using Algorithm 2.

In particular, we first show that the matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ used in Algorithm 1 is a potential game in Lemma 1. Then, we show that Algorithm 2 can efficiently compute an ϵ-approximate Nash equilibrium for potential games and output a product policy as shown in Lemma 2.

Lemma 1. *In line 5 of Algorithm 1, the matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ forms a potential game for both settings of semi-bandit feedback and bandit feedback.*

Proof. In the setting of semi-bandit feedback, since $\overline{Q}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} (\hat{r}^{k,f} + b^{k,f,\mathrm{r}})(\boldsymbol{a})$, the reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ form a congestion game, which we know is a potential game [Monderer and Shapley, 1996].

In the setting of bandit feedback, notice that by defining $\tilde{r}^{k,f}(i) = \widehat{\theta}^k_{i+m(f-1)}$ for $(i, f) \in [m] \times \mathcal{F}$, we can have $\tilde{r}_i^k(\boldsymbol{a}) = \left\langle A_i(\boldsymbol{a}), \widehat{\theta}^k \right\rangle = \sum_{f \in a_i} \tilde{r}^{k,f}(n^f(\boldsymbol{a}))$. Therefore, we claim that the desired potential function is

$$\Phi^k(\boldsymbol{a}) = \widetilde{\Phi}^k(\boldsymbol{a}) + \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}), \quad \text{where} \quad \widetilde{\Phi}^k(\boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} \tilde{r}^{k,f}(i).$$

To see this, by referring to the definition of potential function in congestion game [Monderer and Shapley, 1996], since $\tilde{r}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} \tilde{r}^{k,f}(n^f(\boldsymbol{a}))$, we have that

$$\widetilde{\Phi}^k(a_i, a_{-i}) - \widetilde{\Phi}^k(a_i', a_{-i}) = \tilde{r}_i(a_i, a_{-i}) - \tilde{r}_i(a_i', a_{-i}).$$

As a result, we have

$$\begin{aligned}
&\Phi^k(a_i, a_{-i}) - \Phi^k(a_i', a_{-i}) \\
=&\left(\tilde{r}_i(a_i, a_{-i}) + \tilde{b}^{k,\mathrm{r}}(a_i, a_{-i})\right) - \left(\tilde{r}_i(a_i', a_{-i}) + \tilde{b}^{k,\mathrm{r}}(a_i', a_{-i})\right) \\
=&\overline{Q}_i^k(a_i, a_{-i}) - \overline{Q}_i^k(a_i', a_{-i}),
\end{aligned}$$

which means that $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ form a potential game. □

Lemma 2. *Algorithm 2 can output an ϵ-approximate Nash equilibrium.*

[2] Although the website's intelligent recommendation system may more or less mitigate this effect, it can be considered as a part of the reward function's property.

Proof. Note that if at round k, we have $\max_{i\in[m]} \Delta_i \le \epsilon$, then π^k is an ϵ-approximate Nash equilibrium. So we only need to prove that $\max_{i\in[m]} \Delta_i \le \epsilon$ is satisfied at some round $k \in \{1, \dots, \lceil \frac{mr_{\max}}{\epsilon} \rceil\}$.

Suppose the potential game $(\{\mathcal{A}_i\}_{i=1}^m, \{r_i\}_{i=1}^m)$ is associated with potential function $\Phi \in [0, \Phi_{\max}]$. Set $\pi^* = \operatorname{argmax}_{\pi\in\prod_{i\in[m]}\Delta(\mathcal{A}_i)} \Phi(\pi)$. Then for any $\pi \in \prod_{i\in[m]} \Delta(\mathcal{A}_i)$, we have

$$\begin{aligned}
\Phi(\pi^*) - \Phi(\pi) &= \sum_{i\in[m]} \left(\Phi(\pi^*_{1:i}, \pi_{i+1:m}) - \Phi(\pi^*_{1:i-1}, \pi_{i:m})\right) \\
&= \sum_{i\in[m]} \left(V_i^{\pi^*_{1:i}, \pi_{i+1:m}} - V_i^{\pi^*_{1:i-1}, \pi_{i:m}}\right) \\
&\le mr_{\max}.
\end{aligned}$$

As a result, we can set $\Phi_{\max} = mr_{\max}$. On the other hand, if $j = \operatorname{argmax}_{i\in[m]} \Delta_i$ for round k, we have

$$\begin{aligned}
\Phi(\pi^{k+1}) - \Phi(\pi^k) =& \Phi(\pi_j^{k+1}, \pi_{-j}^k) - \Phi(\pi^k) \\
=& V_j^{\pi_j^{k+1}, \pi_{-j}^k} - V_j^{\pi^k} \\
=& r_j(a_j^{k+1}, \pi_{-j}^k) - r_j(\pi^k) && (\pi^k \text{ is deterministic}) \\
=& \Delta_j \\
=& \max_{i\in[m]} \Delta_i.
\end{aligned}$$

So there must exist $k \in \{1, \dots, \lceil \frac{mr_{\max}}{\epsilon} \rceil\}$ such that $\max_{i\in[m]} \Delta_i \le \epsilon$, otherwise $\Phi(\pi^k)$ increase at least ϵ at each round, which contradicts $\Phi \in [0, mr_{\max}]$. □

C Algorithm Details for Independent Markov Congestion Games

In this section, present missing details of our centralized algorithm for independent Markov congestion games, which is summarized in Algorithm 4. The proof of its theoretical guarantee is given in Appendix F.

C.1 Algorithm for Semi-bandit Feedback

Under the semi-bandit feedback, the players can receive reward information from all facilities they choose. Therefore, we can similarly define

$$\begin{aligned}
N_h^{k,f}(s^f, n) &= \sum_{k'=1}^{k} \mathbb{1}\left\{(s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s^f, n)\right\}, \\
\hat{r}_h^{k,f}(s^f, n) &= \frac{\sum_{k'=1}^k r_h^{k',f} \mathbb{1}\left\{(s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s^f, n)\right\}}{N_h^{k,f}(s^f, n) \vee 1}, \\
\widehat{P}_h^{k,f}(s'^f \mid s^f, n) &= \frac{\sum_{k'=1}^k \mathbb{1}\left\{(s_{h+1}^{k',f}, s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s'^f, s^f, n)\right\}}{N_h^{k,f}(s^f, n) \vee 1}.
\end{aligned}$$

Then, the estimators for the reward function and transition kernel can be defined as

$$\hat{r}_{h,i}^k(s, \boldsymbol{a}) = \sum_{f\in a_i} \hat{r}_h^{k,f}(s^f, n^f(\boldsymbol{a})), \quad \widehat{P}_h^k(s' \mid s, \boldsymbol{a}) = \prod_{f\in\mathcal{F}} \widehat{P}_h^{k,f}(s'^f \mid s^f, n^f(\boldsymbol{a})) \tag{6}$$

Then, with $\iota = 2\log(4(m+1)(\sum_{f\in\mathcal{F}} S^f)T/\delta)$, we define the bonus term to be $b_h^k(s, \boldsymbol{a}) = b_h^{k,\mathrm{pv}}(s, \boldsymbol{a}) + b_h^{k,\mathrm{r}}(s, \boldsymbol{a})$, which is a sum of transition bonus and reward bonus. In particular, we have

$$b_h^{k,\mathrm{pv}}(s, \boldsymbol{a}) = \sum_{f\in\mathcal{F}} \sqrt{\frac{4H^2F^2S^f\iota}{N_h^{k,f}(s^f, n^f(\boldsymbol{a})) \vee 1}} + \sum_{f\neq f'} \sqrt{\frac{4H^2F^2\left(S^fS^{f'}\iota\right)^2}{N_h^{k,f}(s^f, n^f(\boldsymbol{a}))N_h^{k,f'}(s^{f'}, n^{f'}(\boldsymbol{a})) \vee 1}}, \tag{7}$$

Algorithm 4 Nash-VI for IMCGs

1: **Input:** ϵ, accuracy parameter for Nash equilibrium computation
2: **Initialize:** $\overline{V}^k_{H+1,i}(s) = 0$ for all $(i,k,s) \in [m] \times [K] \times \mathcal{S}$
3: **for** episode $k = 1, \ldots, K$ **do**
4: **for** step $h = H, H-1, \ldots, 1$ **do**
5: **for** player $i = 1, \ldots, m$ **do**
6: $\overline{Q}^k_{h,i}(s, \boldsymbol{a}) \leftarrow \min\left\{(\hat{r}^k_{h,i} + \widehat{\mathbb{P}}^k_h \overline{V}^k_{h+1,i} + b^k_h)(s, \boldsymbol{a}), HF\right\}$ for all $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$
7: **for** $s \in \mathcal{S}$ **do**
8: $\pi^k_h(\cdot \mid s) \leftarrow \epsilon$-NASH$(\overline{Q}^k_{h,1}(s,\cdot), \cdots, \overline{Q}^k_{h,m}(s,\cdot))$
9: **for** player $i = 1, \ldots, m$ **do**
10: $\overline{V}^k_{h,i}(s) \leftarrow \mathbb{E}_{\boldsymbol{a} \sim \pi^k_h}[\overline{Q}^k_{h,i}(s, \boldsymbol{a})]$
11: **for** step $h = 1, \ldots, H$ **do**
12: Take action $\boldsymbol{a}^k_h \sim \pi^k_h(\cdot \mid s^k_h)$, observe reward $r^{k,f}_h$ and next state s^k_{h+1}
13: Update reward estimator $\hat{r}^k_{h,i}$, transition estimator $\widehat{P}^k_h$ and bonus term b^k_h

$$b^{k,\mathrm{r}}_h(s, \boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sqrt{\frac{\iota}{N^{k,f}_h(s^f, n^f(\boldsymbol{a})) \vee 1}}. \tag{8}$$

For convenience, we define $(\widehat{\mathbb{P}}^k_h V)(s, \boldsymbol{a}) = \mathbb{E}_{s' \sim \widehat{P}^k_h(\cdot \mid s, \boldsymbol{a})}[V(s')]$ with value function $V : \mathcal{S} \mapsto \mathbb{R}$.

Remark 5. Unlike Algorithm 1 for congestion game, here, $\overline{Q}^k_{h,1}(s,\cdot), \ldots, \overline{Q}^k_{h,m}(s,\cdot)$ in line 6 of Algorithm 4 in general does not form a potential game. Therefore, we cannot use Algorithm 2 and ϵ-NASH is not always computationally efficient.

C.2 Algorithm for Bandit Feedback

In bandit feedback scenario, since players' observation about state transitions remains unaffected, we only need to modify the reward estimator $\hat{r}^k_{h,i}$ defined in (6) and reward bonus term $b^{k,\mathrm{r}}_h(s, \boldsymbol{a})$ defined in (8).

Similar to the congestion game with bandit feedback introduced in Section 4.2, for IMCGs, we can also write its reward function as $r_{h,i}(s, \boldsymbol{a}) = \langle A_i(s, \boldsymbol{a}), \theta_h \rangle$, where θ_h is unknown and $A_i(s, \boldsymbol{a})$ is a 0-1 vector.

In particular, define $\theta_h \in [0,1]^d$ with $d = m \sum_{f \in \mathcal{F}} S^f$ to be the vector such that $\theta_{h,i} = r^f_h(s^f, n)$ for some $f \in \mathcal{F}$ and $(s^f, n) \in \mathcal{S}^f \times [m]$. Then, we can similarly build estimator $\hat{r}^k_{h,i}$ through ridge regression as the following.[3]

$$\text{design matrix:} \quad V^k_h = I + \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(s^{k'}_h, \boldsymbol{a}^{k'}_h) A_i(s^{k'}_h, \boldsymbol{a}^{k'}_h)^\top, \tag{9}$$

$$\theta_h \text{ estimator:} \quad \widehat{\theta}^k_h = \left(V^k_h\right)^{-1} \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(s^{k'}_h, \boldsymbol{a}^{k'}_h) r^{k'}_{h,i}, \tag{10}$$

$$\text{reward estimator:} \quad \tilde{r}^k_{h,i}(s, \boldsymbol{a}) = \left\langle A_i(s, \boldsymbol{a}), \widehat{\theta}^k_h \right\rangle, \tag{11}$$

$$\text{reward bonus:} \quad \tilde{b}^{k,\mathrm{r}}_h(s, \boldsymbol{a}) = \max_{i \in [m]} \|A_i(s, \boldsymbol{a})\|_{\left(V^k_h\right)^{-1}} \sqrt{\beta_k}, \tag{12}$$

where $\sqrt{\beta_k} = \sqrt{d} + \sqrt{d \log\left(1 + \frac{mkF}{d}\right) + \iota}$.

[3] For the same reason, we take the regularization parameter in ridge regression to be 1.

D Analysis for Algorithm 1

Recall that the update rule in Algorithm 1 is $\overline{Q}_i^k(\boldsymbol{a}) = \hat{r}_i^k(\boldsymbol{a}) + b_i^{k,\mathrm{r}}(\boldsymbol{a})$, where we have

$$b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \sum_{f \in a_i} b^{k,f,\mathrm{r}}(\boldsymbol{a}), \quad \text{and} \quad b^{k,f,\mathrm{r}}(\boldsymbol{a}) = \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a})) \vee 1}}.$$

For proof convenience, we define auxiliary value functions

$$\begin{aligned}
&\underline{Q}_i^k(\boldsymbol{a}) = \hat{r}_i^k(\boldsymbol{a}) - b_i^{k,\mathrm{r}}(\boldsymbol{a}), \\
&\overline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\overline{Q}_i^k(\boldsymbol{a})] \quad \text{and} \quad \underline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\underline{Q}_i^k(\boldsymbol{a})].
\end{aligned}$$

With these definitions, we now begin to prove Theorem 1.

Proof of Theorem 1. **Semi-bandit Feedback.** By the update rules in Algorithm 1, in the setting of semi-bandit feedback, with probability at least $1-\delta$, simultaneously for all $(k, i, \boldsymbol{a}) \in [K] \times [m] \times \mathcal{A}$, we have

$$\overline{Q}_i^k(\boldsymbol{a}) - r_i(\boldsymbol{a}) = \sum_{f \in a_i} \left[(\hat{r}^{k,f} - r^f)(\boldsymbol{a}) + b^{k,f,\mathrm{r}}(\boldsymbol{a})\right] \geq 0.$$

The second inequality above is obtained by using standard Hoeffding's inequality and union bound, Therefore, we have $\overline{Q}_i^k(\boldsymbol{a}) \geq r_i(\boldsymbol{a})$.

Then, since π^k is the ϵ-approximate Nash equilibrium policy of $\overline{Q}_1^k, \ldots, \overline{Q}_m^k$, we have

$$\begin{aligned}
\overline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\overline{Q}_i^k(\boldsymbol{a})] &= \max_{\nu \in \Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim(\nu,\pi_{-i}^k)}[\overline{Q}_i^k(\boldsymbol{a})] - \epsilon \\
&\geq \max_{\nu \in \Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim(\nu,\pi_{-i}^k)}[r_i(\boldsymbol{a})] - \epsilon = V_i^{\dagger,\pi_{-i}^k} - \epsilon.
\end{aligned}$$

Meanwhile, by definition of $\underline{Q}_i^k(\boldsymbol{a})$ and $\underline{V}_i^k$, we can similarly show that $\underline{Q}_i^k(\boldsymbol{a}) \leq r_i(\boldsymbol{a})$ and $\underline{V}_i^k \leq V_i^{\pi^k}$. Therefore, we can have $V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k} \leq \overline{V}_i^k - \underline{V}_i^k + \epsilon$.

Now, we define $\widetilde{Q}^k(\boldsymbol{a}) = \max_{i\in[m]} 2b_i^{k,\mathrm{r}}(\boldsymbol{a})$ and $\widetilde{V}^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\widetilde{Q}^k(\boldsymbol{a})]$. Then, we can notice that

$$\max_{i\in[m]}(\overline{Q}_i^k - \underline{Q}_i^k)(\boldsymbol{a}) \leq \max_{i\in[m]} 2b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \widetilde{Q}^k(\boldsymbol{a}),$$

$$\max_{i\in[m]}(\overline{V}_i^k - \underline{V}_i^k) \leq \mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\max_{i\in[m]}(\overline{Q}_i^k - \underline{Q}_i^k)(\boldsymbol{a})\right] \leq \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\widetilde{Q}^k(\boldsymbol{a})] = \widetilde{V}^k.$$

We further define $\mathcal{M}^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\widetilde{Q}^k(\boldsymbol{a})\right] - \widetilde{Q}^k(\boldsymbol{a}^k) = \widetilde{V}^k - \widetilde{Q}^k(\boldsymbol{a}^k)$. It is not hard to verify that $\mathcal{M}^k$ is a martingale difference sequence with respect to the history from episode 1 to $k-1$. Meanwhile, since $\left|b^{k,\mathrm{r}}(\boldsymbol{a})\right| = \sum_{f\in\mathcal{F}} \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a}))\vee 1}} \leq F\sqrt{\tilde{\iota}}$. Thus, by Azuma-Hoeffding inequality, we have $\sum_{k=1}^K \mathcal{M}^k = \widetilde{\mathcal{O}}\left(F\sqrt{K}\right)$. Therefore, we have

$$\begin{aligned}
\text{Nash-Regret}(K) &= \sum_{k=1}^K \max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k}\right) \\
&= \sum_{k=1}^K \min\left\{\max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k}\right), F\right\} \\
&\qquad\qquad \text{(Since the value is always bounded by } F.) \\
&\leq \sum_{k=1}^K \min\left\{\max_{i\in[m]} \left(\overline{V}_i^k - \underline{V}_i^k\right), F\right\} + K\epsilon
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \min\left\{\widetilde{V}^k, F\right\} + K\epsilon \\
&= \sum_{k=1}^{K} \left(\min\left\{\widetilde{Q}^k(\boldsymbol{a}^k), F\right\} + \mathcal{M}^k\right) + K\epsilon \\
&\leq \widetilde{\mathcal{O}}\left(F\sqrt{K}\right) + 2\sum_{k=1}^{K}\left\{\max_{i\in[m]} b_i^{k,\mathrm{r}}(\boldsymbol{a}^k), F\right\} && \text{(By taking } \epsilon = 1/K.\text{)} \\
&\leq \widetilde{\mathcal{O}}\left(F\sqrt{K}\right) + 2\sum_{f\in\mathcal{F}}\sum_{k=1}^{K}\sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a}^k))\vee 1}} \\
&\leq \widetilde{\mathcal{O}}\left(F\sqrt{mK}\right) && \text{(By Lemma 6.)}
\end{aligned}
$$

Bandit Feedback. By using Lemma 3, which guarantees optimistic estimation, we can similarly show that

$$
\text{Nash-Regret}(K) \leq \sum_{k=1}^{K}\mathcal{M}^k + \sum_{k=1}^{K}\min\left\{2\tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}^k), F\right\} + K\epsilon.
$$

To have an upper bound on $\mathcal{M}^k$ here, recall that $\tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}) = \max_{i\in[m]}\|A_i(\boldsymbol{a})\|_{(V^k)^{-1}}\sqrt{\tilde{\beta}_k}$ and $\sqrt{\tilde{\beta}_K} = \widetilde{\mathcal{O}}\left(\sqrt{\tilde{d}}\right) = \widetilde{\mathcal{O}}\left(\sqrt{mF}\right)$. Meanwhile, we have $\|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \leq \|A_i(\boldsymbol{a})\|_I = \|A_i(\boldsymbol{a})\|_2 \leq \sqrt{F}$. Thus, we have $\left|\mathcal{M}^k\right| \leq \widetilde{\mathcal{O}}\left(\sqrt{mF^2}\right)$, which by Azuma-Hoeffding inequality implies $\sum_{k=1}^{K}\mathcal{M}^k = \widetilde{\mathcal{O}}\left(\sqrt{mF^2K}\right)$.

Then the sum of the bonus terms can be bounded by using Lemma 4. In particular, with $\epsilon = 1/K$, we have

$$
\begin{aligned}
\text{Nash-Regret}(K) &\leq \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right) + 2\sum_{k=1}^{K}\min\left\{\max_{i\in[m]}\left\|A_i(\boldsymbol{a}^k)\right\|_{(V^k)^{-1}}\sqrt{\tilde{\beta}_k}, F\right\} \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{mF^2K}\right) + 2\sqrt{K\sum_{k=1}^{K}\min\left\{\max_{i\in[m]}\|A_i(\boldsymbol{a}^k)\|^2_{(V^k)^{-1}}\tilde{\beta}_k, F^2\right\}} \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{mF^2K}\right) + \sqrt{\widetilde{\mathcal{O}}\left(mF^2K\right)\sum_{k=1}^{K}\min\left\{\max_{i\in[m]}\|A_i(\boldsymbol{a}^k)\|^2_{(V^k)^{-1}}\tilde{\beta}_k, 1\right\}} \\
&\qquad\qquad \text{(Since } \tilde{\beta}_k = \widetilde{\mathcal{O}}\left(d\right) = \widetilde{\mathcal{O}}\left(mF\right).\text{)} \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{mF^2K}\right) + \widetilde{\mathcal{O}}\left(\sqrt{mF^2K\cdot mF}\right) && \text{(By Lemma 4.)} \\
&\leq \widetilde{\mathcal{O}}\left(mF^{3/2}\sqrt{K}\right).
\end{aligned}
$$

□

D.1 Lemmas for Bandit Feedback

The following lemma, as a direct corollary of the confidence bound for least square estimators, shows that the reward estimation error can be bounded by the reward bonus term.

Lemma 3. *With probability at least $1-\delta$, simultaneously for all $(i, k, \boldsymbol{a})$, it holds that $|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| \leq \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a})$, where $\tilde{r}_i^k$ and $\tilde{b}^{k,\mathrm{r}}$ are defined in (2).*

Proof. By construction, we have

$$
|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| = \left|\left\langle A_i(\boldsymbol{a}), \widehat{\theta} - \theta\right\rangle\right|
$$

$$
\begin{aligned}
&\leq \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \left\|\widehat{\theta}-\theta\right\|_{V^k} \\
&\leq \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \left(\|\theta\|_2 + \sqrt{\log\left(\det(V^k)\right)+\tilde{\iota}}\right). \\
&\qquad\qquad \text{(By Theorem 20.5 in Lattimore and Szepesvári [2020].)}
\end{aligned}
$$

Since each element in θ is bounded in $[0,1]$ by construction, we have $\|\theta\|_2 \leq \sqrt{\tilde{d}}$.

Then, by Lemma 4, we have $\det\left(V^k\right) \leq \left(1+\frac{mkF}{\tilde{d}}\right)^{\tilde{d}}$ since by construction $\|A_i(\boldsymbol{a})\|_2^2 \leq F$.

Finally, to make this bound valid for all player $i \in [m]$, we only need to take maximization over $i \in [m]$. Therefore, with probability at least $1-\delta$, we have

$$
|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| \leq \max_{i\in[m]} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k} = \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}),
$$

where $\sqrt{\tilde{\beta}_k} = \sqrt{\tilde{d}} + \sqrt{\tilde{d}\log\left(1+\frac{mkF}{\tilde{d}}\right)+\tilde{\iota}}$. □

The following is a variant of the famous elliptical potential lemma, which helps bound the sum of reward bonus under bandit feedback. Here, we apply some techniques from the proof of Lemma 19.4 in Lattimore and Szepesvári [2020].

Lemma 4. *Let $K, m \geq 1$ be integers. Suppose $V^k = I + \sum_{k'=1}^{k-1}\sum_{i=1}^{m} A_i^{k'}\left(A_i^{k'}\right)^\top$, where $A_i^{k'} \in \mathbb{R}^d$ and $\left\|A_i^{k'}\right\|_2^2 \leq F$. Then, it holds that*

$$
\det\left(V^k\right) \leq \left(1+\frac{mkF}{d}\right)^d, \quad \textit{and} \quad \sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i^k\right\|_{(V^k)^{-1}}^2, 1\right\} \leq 2d\log\left(1+\frac{mKF}{d}\right).
$$

Proof. For the first upper bound about $\det\left(V^k\right)$, we have

$$
\begin{aligned}
\det\left(V^k\right) &= \prod_{j=1}^{d} \lambda_j && (\lambda_1,\ldots,\lambda_d \text{ are eigenvalues of } V^k) \\
&\leq \left(\frac{\operatorname{tr}\left(V^k\right)}{d}\right)^d && \text{(By AM-GM inequality)} \\
&= \left(\frac{\operatorname{tr}(I) + \sum_{k'=1}^{k-1}\sum_{i=1}^{m}\left\|A_i^{k'}\right\|_2^2}{d}\right)^d \\
&\leq \left(1+\frac{mkF}{d}\right)^d. && \text{(Since } \left\|A_i^{k'}\right\|_2^2 \leq F.)
\end{aligned}
$$

For the second upper bound. First, we notice that $\min\{1,x\} \leq 2\log(1+x)$ for any $x \geq 0$. Thus, we have

$$
\sum_{k=1}^{K} \min\left\{1, \max_{i\in[m]} \left\|A_i^k\right\|_{(V^k)^{-1}}^2\right\} \leq 2\sum_{k=1}^{K} \log\left(1+\max_{i\in[m]} \left\|A_i^k\right\|_{(V^k)^{-1}}^2\right).
$$

Then, for $k \geq 2$, we can notice that

$$
\begin{aligned}
V^k &= V^{k-1} + \sum_{i=1}^{m} A_i^{k-1}\left(A_i^{k-1}\right)^\top \\
&= \left(V^{k-1}\right)^{1/2}\left(I + \left(V^{k-1}\right)^{-1/2}\left(\sum_{i=1}^{m} A_i^{k-1}\left(A_i^{k-1}\right)^\top\right)\left(V^{k-1}\right)^{-1/2}\right)\left(V^{k-1}\right)^{1/2} \\
&= \left(V^{k-1}\right)^{1/2}\left(I + \sum_{i=1}^{m}\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)^\top\right)\left(V^{k-1}\right)^{1/2}.
\end{aligned}
$$

Therefore, we have

$$\begin{aligned}
\det\left(V^k\right) &= \det\left(V^{k-1}\right)\det\left(I+\sum_{i=1}^{m}\left(\left(V^{k-1}\right)^{-1/2}A_i^{k-1}\right)\left(\left(V^{k-1}\right)^{-1/2}A_i^{k-1}\right)^\top\right)\\
&\geq \det\left(V^{k-1}\right)\left(1+\max_{i\in[m]}\left\|A_i^{k-1}\right\|^2_{(V^{k-1})^{-1}}\right) && \text{(By Lemma 5.)}\\
&\geq \prod_{k'=1}^{k-1}\left(1+\max_{i\in[m]}\left\|A_i^{k'}\right\|^2_{\left(V^{k'}\right)^{-1}}\right). && \text{(Since by definition, } V^1=I.\text{)}
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\sum_{k=1}^{K}\min\left\{\max_{i\in[m]}\left\|A_i^k\right\|^2_{(V^k)^{-1}},1\right\} &\leq 2\sum_{k=1}^{K}\log\left(1+\max_{i\in[m]}\left\|A_i^k\right\|^2_{(V^k)^{-1}}\right)\\
&\leq 2\log\left(\det\left(V^{K+1}\right)\right)\\
&\leq 2d\log\left(1+\frac{mKF}{d}\right).
\end{aligned}$$

□

D.2 Technical Lemmas

Lemma 5. *Let $y_1,\ldots,y_m\in\mathbb{R}^d$ be a set of vectors. Then, it holds that*

$$\det\left(I+\sum_{i=1}^{m}y_iy_i^\top\right)\geq 1+\max_{i\in[m]}\|y_i\|_2^2.$$

Proof. Since $I+\sum_{i=1}^{m}y_iy_i^\top \succeq I+y_iy_i^\top$ for any $i\in[m]$, we have $\det\left(I+\sum_{i=1}^{m}y_iy_i^\top\right)\geq \det\left(I+y_iy_i^\top\right)$ for any $i\in[m]$. That is, we have

$$\det\left(I+\sum_{i=1}^{m}y_iy_i^\top\right)\geq\max_{i\in[m]}\det\left(I+y_iy_i^\top\right)=1+\max_{i\in[m]}\|y_i\|_2^2.$$

The last line above holds because the matrix $I+y_iy_i^\top$ has eigenvalues $1+\|y_i\|_2^2$ and 1. □

Lemma 6. *For any $f\in\mathcal{F}$, it holds that*

$$\sum_{k=1}^{K}\sqrt{\frac{1}{\mathcal{N}^{k,f}(n^f(\boldsymbol{a}^k))\vee 1}}\leq\widetilde{\mathcal{O}}\left(\sqrt{mK}\right).$$

Proof. Here, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sqrt{\frac{1}{N^{k,f}(n^f(\boldsymbol{a}^k))\vee 1}} &= \sum_{n=0}^{m}\sum_{\ell=1}^{N^{K,f}(n)}\sqrt{\frac{1}{\ell}}\\
&\leq 2\sum_{n=0}^{m}\sqrt{N^{K,f}(n)} && \text{(By standard technique)}\\
&\leq 2\sqrt{(m+1)\sum_{n=0}^{m}N^{K,f}(n)}\\
&= \widetilde{\mathcal{O}}\left(\sqrt{mK}\right).
\end{aligned}$$

The last equality above is based on a pigeon-hold principle argument similar to Lemma 20. □

E Analysis for Algorithm 3

E.1 Exploration Distribution and Smoothness

We choose the exploration distribution to be the G-optimal design and we have the following properties.

Lemma 7. *(Unbiasedness) For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\mathbb{E}_k\left[\widehat{\nabla}_i^k \Phi(a)\right] = \nabla_i^k \Phi(a),$$

where $\mathbb{E}_k[\cdot]$ is taken over all the randomness before episode k.

Proof. By the definition of $\widehat{\nabla}_i^k \Phi(a)$, we have

$$\begin{aligned}
\mathbb{E}_k\left[\widehat{\nabla}_i^k \Phi(a)\right] =& \mathbb{E}_k\left\langle \phi_i(a), \widehat{\theta}_i^k(\pi^k)\right\rangle \\
=& \mathbb{E}_k\left[\frac{1}{\tau}\sum_{t=1}^{\tau} \phi_i(a)^\top [\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t}) r_i^{k,t}\right] \\
=& \mathbb{E}_k\left[\phi_i(a)^\top [\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,1}) r_i^{k,1}\right] \\
=& \mathbb{E}_k\left[\phi_i(a)^\top [\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,1})\phi_i(a_i^{k,1})^\top \theta_i^{k,1}(\pi^k)\right] \\
=& \sum_{a_i^k \in \mathcal{A}_i} \pi_i^k(a_i^{k,1})\phi_i^\top(a)[\Sigma_i^k]^{-1}\phi_i(a_i^{k,1})\phi_i(a_i^{k,1})^\top \theta_i(\pi^k) \\
& \qquad (a_i^{k,1} \text{ only depends on } \pi_i^k \text{ and } \theta_i^{k,1}(\pi^k) \text{ only depends on } \pi_{-i}^k) \\
=& \phi_i^\top(a)[\Sigma_i^k]^{-1}\left[\sum_{a_i^k \in \mathcal{A}_i} \pi_i^k(a_i^{k,1})\phi_i(a_i^{k,1})\phi_i(a_i^{k,1})^\top\right]\theta_i(\pi^k) \\
=& \phi_i^\top(a_i)\theta_i(\pi^k) \\
=& \nabla_i^k \Phi(a).
\end{aligned}$$

□

Lemma 8. *For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\left|\phi_i(a)^\top [\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t}) r_i^{k,t}\right| \le \frac{F^2}{\gamma}.$$

Proof. As $\pi_i^k = (1-\gamma)(\nu\widetilde{\pi}_i^k + (1-\gamma)\pi_i^{k-1}) + \gamma\rho_i$, we have

$$\Sigma_i^k = \mathbb{E}_{a_i \sim \pi_i^k}\phi_i(a_i)\phi_i(a_i)^\top \succeq \gamma \mathbb{E}_{a_i \sim \rho_i}\phi_i(a_i)\phi_i(a_i)^\top,$$

and ρ_i is the G-optimal design with respect to $\phi_i(\cdot)$, for any action $a \in \mathcal{A}_i$ we have

$$\|\phi_i(a)\|^2_{[\Sigma_i^k]^{-1}} \le \frac{1}{\gamma}\|\phi_i(a)\|^2_{[\mathbb{E}_{a_i \sim \rho_i}\phi_i(a_i)\phi_i(a_i)^\top]^{-1}} \le \frac{F}{\gamma}.$$

Then for any $t \in [\tau]$, since $|r_i^{k,t}| \le F$, we have

$$\left|r_i^{k,t}\phi_i^\top(a)[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})\right| \le \left|r_i^{k,t}\right| \|\phi_i(a)\|_{[\Sigma_i^k]^{-1}} \left\|\phi_i(a_i^{k,t})\right\|_{[\Sigma_i^k]^{-1}} \le \frac{F^2}{\gamma}.$$

As a result, we have

$$\left|\widehat{\nabla}_i^k \Phi(a)\right| = \left|\frac{1}{\tau}\sum_{t=1}^{\tau}\phi_i(a)^\top [\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t}) r_i^{k,t}\right| \le \frac{F^2}{\gamma}$$

□

Lemma 9. *For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t})r_i^{k,t}\right)^2\right] \le \frac{F^3}{\gamma}.$$

Proof. We first show that for any $t \in [\tau]$, we have

$$\begin{aligned}
&\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t})r_i^{k,t}\right)^2\right]\\
\le& F^2\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t})\right)^2\right]\\
\le& F^2\mathbb{E}_k\left[\phi_i(a)^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a_i^{k,t})\phi_i(a_i^{k,t})^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a)^\top\right]\\
=& F^2\phi_i(a)^\top[\Sigma_i^{k,t}]^{-1}\phi_i(a)\\
\le& \frac{F^3}{\gamma}.
\end{aligned}$$

□

Lemma 10. *With probability $1-\delta$, for all $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\left|\widehat{\nabla}_i^k\Phi(a) - \nabla_i^k\Phi(a)\right| \le c\sqrt{\frac{F^4\log(mK/\delta)}{\gamma\tau}} + \frac{cF^3\log(mK/\delta)}{\gamma\tau}$$

Proof. Recall that

$$\widehat{\nabla}_i^k\Phi(a_i) = \frac{1}{\tau}\sum_{t=1}^{\tau}\phi_i^\top(a_i)[C_i^k]^{-1}r_i^{k,t}\phi_i(a_i^{k,t}),$$

and $(a_i^{k,t}, r_i^{k,t})$ are drawn independently at each $t \in [\tau]$. Lemma 7 shows that $\widehat{\nabla}_i^k\Phi(a_i)$ is an unbiased estimate of $\nabla_i^k\Phi(a_i)$ In addition, Lemma 8 shows that $\phi_i^\top(a_i)[C_i^k]^{-1}r_i^{k,t}\phi_i(a_i^{k,t})$ is bounded by F^2/γ and Lemma 9 shows that its second moment is bounded by F^3/γ. Then by Bernstein's inequality, for a fixed $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, with probability $1-\delta$, we have

$$\left|\widehat{\nabla}_i^k\Phi(a) - \nabla_i^k\Phi(a)\right| \le \sqrt{\frac{2F^3\log(2/\delta)}{\gamma\tau}} + \frac{3F^2\log(2/\delta)}{2\gamma\tau}.$$

The argument holds by applying the union bound and the fact that $|\mathcal{A}_i| \le 2^F$.

□

Lemma 11. *$\Phi(\cdot)$ is mF-Lipschitz and mF-smooth with respect to the L1 norm $\|\cdot\|_1$.*

Proof. Recall that $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a}\sim\pi}\Phi(\boldsymbol{a})$ and $\Phi(\boldsymbol{a}) \in [0, mF]$.

$$\begin{aligned}
\Phi(\pi) - \Phi(\pi') =& \mathbb{E}_{\boldsymbol{a}\sim\pi}\Phi(\boldsymbol{a}) - \mathbb{E}_{\boldsymbol{a}\sim\pi'}\Phi(\boldsymbol{a})\\
=& \sum_{i\in[m]}\mathbb{E}_{a_{1:i-1}\sim\pi'_{1:i-1}, a_{i:m}\sim\pi_{i:m}}\Phi(\boldsymbol{a}) - \mathbb{E}_{a_{1:i}\sim\pi'_{1:i}, a_{i+1:m}\sim\pi_{i+1:m}}\Phi(\boldsymbol{a})\\
\le& \sum_{i\in[m]}\|\pi_i - \pi'_i\|_1 \cdot \|\Phi\|_\infty\\
\le& mF\|\pi - \pi'\|_1.
\end{aligned}$$

Similarly we have $\nabla_\pi\Phi(a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\Phi(a_i, a_{-i})$. As a result, we have

$$\|\nabla_\pi\Phi - \nabla_{\pi'}\Phi\|_\infty \le mF\|\pi - \pi'\|_1.$$

□

Definition 4. (Frank Wolfe Gap) The Frank Wolfe gap of a joint strategy π for $\Phi(\cdot)$ is defined as

$$G(\pi) = \max_{\pi'} \langle \pi' - \pi, \nabla_\pi \Phi \rangle .$$

Lemma 12. *Suppose the Frank Wolfe gap of π is ϵ. Then π is an ϵ-Nash policy.*

Proof. For a fixed player i, suppose player i change her strategy to π_i'.

$$\begin{aligned} V_i^{\pi_i', \pi_{-i}} - V_i^{\pi} &= \Phi(\pi_i', \pi_{-i}) - \Phi(\pi) \\ &= \langle \pi_i' - \pi_i, \nabla_{\pi_i} \Phi \rangle \\ &\leq \max_{\pi'} \langle \pi' - \pi, \nabla_\pi \Phi \rangle \\ &\leq \epsilon. \end{aligned}$$

□

E.2 Analysis for Frank Wolfe in Bandit Feedback

Theorem 4. *Let $T = K\tau$. For the congestion game with bandit feedback, by running Algorithm 3 with gradient estimator $\widehat{\nabla}_i^k \Phi$ and exploration distribution ρ_i defined in (4) and (5), setting parameters $\nu = \frac{F}{m\sqrt{K}}$, $\gamma = \frac{F}{mK}$ and $\tau = K^2$, with probability $1-\delta$, we have*

$$\textit{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^2 T^{5/6} + m^3 F^3 T^{2/3}\right).$$

Proof. Set $\nabla^k \Phi = \nabla \Phi(\Pi^k) \in \mathbb{R}^A$ and $\nabla_i^k \Phi = \nabla^k \Phi(\pi_i) \in \mathbb{R}^{A_i}$. As we have $\Phi(\cdot)$ is mF-smooth w.r.t. $\|\cdot\|_1$, we have

$$\begin{aligned} \Phi(\pi^{k+1}) \geq& \Phi(\pi^k) + \left\langle \nabla \Phi(\pi^k), \pi^{k+1} - \pi^k \right\rangle - \frac{mF}{2} \|\pi^{k+1} - \pi^k\|_1^2 \\ =& \Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle + \gamma \left\langle \nabla^k \Phi, \rho - \pi^k \right\rangle \\ & - \frac{mF}{2}(2\nu^2 \left\|\widetilde{\pi}^k - \pi^k\right\|_1^2 + 2\gamma^2 \left\|\rho - \pi^k\right\|_1^2) \\ \geq& \Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle - \gamma \left\|\nabla^k \Phi\right\|_\infty \left\|\rho - \pi^k\right\|_1 \\ & - \frac{mF}{2}(2\nu^2 \left\|\widetilde{\pi}^k - \pi^k\right\|_1^2 + 2\gamma^2 \left\|\rho - \pi^k\right\|_1^2) \\ \geq& \Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle - 2\gamma m^2 F - 4m^3 F(\nu^2 + \gamma^2). \end{aligned}$$

(By Lemma 11.)

Define the true target policy at episode k

$$\widehat{\pi}_i^{k+1} = \operatorname*{argmax}_{\pi_i} \left\langle \pi_i, \nabla_i \Phi(\pi_i^k) \right\rangle,$$

and the Frank Wolfe gap of joint strategy π

$$G(\pi) = \max_{\pi'} \langle \pi' - \pi, \nabla \Phi(\pi) \rangle .$$

Then we have

$$\begin{aligned} \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle =& \left\langle \widehat{\nabla}^k \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla \Phi(\pi^k) - \widehat{\nabla}^k \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle \\ \geq& \left\langle \widehat{\nabla}^k \Phi(\pi^k), \widehat{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla \Phi(\pi^k) - \widehat{\nabla}^k \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle \\ =& \left\langle \nabla \Phi(\pi^k), \widehat{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla \Phi(\pi^k) - \widehat{\nabla}^k \Phi(\pi^k), \widetilde{\pi}^{k+1} - \widehat{\pi}^{k+1} \right\rangle \\ \geq& G(\pi^k) - 2m \left\| \nabla \Phi(\pi^k) - \widehat{\nabla}^k \Phi(\pi^k) \right\|_\infty \\ \geq& G(\pi^k) - c\sqrt{\frac{m^2 F^4 \log(mK/\delta)}{\gamma \tau}} - \frac{cmF^3 \log(mK/\delta)}{\gamma \tau} \end{aligned}$$

Apply it to the previous bound and we have

$$\Phi(\pi^{k+1}) \geq \Phi(\pi^k) + \nu G(\pi^k) - c\frac{\nu}{\sqrt{\gamma\tau}}\sqrt{m^2F^4\log(mK/\delta)} - c\frac{\nu}{\gamma\tau}mF^3\log(mK/\delta) - \gamma 2m^2F - 4m^3F(\nu^2+\gamma^2).$$

Summing over $k \in [K]$ and we get

$$\sum_{k=1}^{K} G(\pi^k) \leq \frac{\Phi(\pi^{K+1}) - \Phi(\pi^1)}{\nu} + c\frac{K}{\sqrt{\gamma\tau}}\sqrt{m^2F^4\log(mK/\delta)} + c\frac{K}{\gamma\tau}mF^3\log(mK/\delta) + \frac{2m^2FK\gamma}{\nu} + \frac{4(\nu^2+\gamma^2)m^3FK}{\nu}.$$

Since $\Phi(\cdot)$ is bounded in [0, mF]. Set $\nu = \frac{F}{m\sqrt{K}}$, $\gamma = \frac{F}{mK}$, $\tau = K^2$, and we have

$$\sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2F^2K^{1/2} + m^3F^3\right).$$

Then by Lemma 12, for $T = K\tau$, we have

$$\text{Nash-Regret}(T) = \tau\sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2F^2T^{5/6} + m^3F^3T^{2/3}\right).$$

□

E.3 Algorithm and Analysis for Semi-bandit Feedback

In the setting of semi-bandit feedback, we will need a different gradient estimator $\widetilde{\nabla}_i^k\Phi(a_i)$ and a different exploration distribution $\tilde{\rho}_i$ to utilize the extra reward information from each chosen facility.

Based on the analysis in Section 5, using (3), we have $\nabla_i^k\Phi(a_i) = \sum_{f\in a_i}[\theta_i(\pi^k)]_f$, where $[\theta_i(\pi^k)]_f = \mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[r^f(n^f(a_{-i})+1)\right]$. Meanwhile, in semi-bandit feedback, the mean of t-th reward player i received for facility f at episode k is $r^f(n^f(a_i^{k,t}, a_{-i}^{k,t}))$. Therefore, we can use inverse propensity score (IPS) estimator to estimate $[\theta_i(\pi^k)]_f$. In particular, we have

$$[\widetilde{\theta}_i^k(\pi^k)]_f = \frac{1}{\tau}\sum_{t=1}^{\tau}[\widetilde{\theta}_i^{k,t}(\pi^k)]_f, \quad \text{where} \quad [\widetilde{\theta}_i^{k,t}(\pi^k)]_f = \frac{r^{k,t,f}\mathbb{1}\left\{f\in a_i^{k,t}\right\}}{\mathbb{P}_{a_i\sim\pi_i^k}(f\in a_i)}.$$

Then, we can naturally have

$$\widetilde{\nabla}_i^k\Phi(a_i) = \sum_{f\in a_i}[\widetilde{\theta}_i^k(\pi^k)]_f. \tag{13}$$

Furthermore, by Lemma 14, we can see that by using $\tilde{\rho}_i$ computed by Algorithm 5, for all players, we have $\mathbb{P}_{a_i\sim\pi_i^k}(f\in a_i) \geq \frac{\gamma}{2F}$ for all $f \in \bigcup_{a_i\in\mathcal{A}_i} a_i$.

Properties of the IPS estimator are summarized in Lemma 15. By using these properties, we can have the following lemma.

Lemma 13. *With probability $1-\delta$, for all $k\in[K]$, $i\in[m]$ and $a_i\in\mathcal{A}_i$, we have*

$$\left|\widetilde{\nabla}_i^k\Phi(a_i) - \nabla_i^k\Phi(a_i)\right| \leq \sqrt{\frac{4F^3\log(2mFK/\delta)}{\gamma\tau}} + \frac{2F^2\log(2mFK/\delta)}{\gamma\tau}.$$

Proof. By Lemma 15 and Bernstein's inequality, simultaneously for all $(i,k,f)\in[m]\times[K]\times\mathcal{F}$, with probability at least $1-\delta$, we have

$$\left|[\widetilde{\theta}_i^k(\pi^k)]_f - [\theta_i(\pi^k)]_f\right| \leq \sqrt{\frac{4F\log(2mFK/\delta)}{\gamma\tau}} + \frac{2F\log(2mFK/\delta)}{\gamma\tau}.$$

Since $\widetilde{\nabla}_i^k \Phi(a_i) = \sum_{f \in a_i} [\widetilde{\theta}_i^k(\pi^k)]_f$, by triangle inequality, we have

$$\left| \widetilde{\nabla}_i^k \Phi(a_i) - \nabla_i^k \Phi(a_i) \right| \leq \sqrt{\frac{4F^3 \log(2mFK/\delta)}{\gamma\tau}} + \frac{2F^2 \log(2mFK/\delta)}{\gamma\tau}.$$

□

With this more refined gradient estimator, we can now have the following theorem.

Theorem 5. *Let $T = K\tau$. For the congestion game with semi-bandit feedback, by running Algorithm 3 with gradient estimator $\widetilde{\nabla}_i^k \Phi$ in (13) and exploration distribution $\tilde{\phi}_i$ in Algorithm 5, setting parameters $\nu = \frac{\sqrt{F}}{m\sqrt{K}}$, $\gamma = \frac{\sqrt{F}}{mK}$ and $\tau = K^2$, with probability $1-\delta$, we have*

$$\textit{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3}\right).$$

Proof. By following the proof of Theorem 4 and applying the concentration inequality in Lemma 13, we can have

$$\begin{aligned}\Phi(\pi^{k+1}) \geq & \Phi(\pi^k) + \nu G(\pi^k) - \frac{\nu}{\sqrt{\gamma\tau}} \sqrt{4m^2F^3 \log(2mK/\delta)} \\ & - \frac{2\nu}{\gamma\tau} mF^2 \log(mK/\delta) - \gamma 2m^2 F - 4m^3 F(\nu^2 + \gamma^2).\end{aligned}$$

Summing over $k \in [K]$ and we get

$$\begin{aligned}\sum_{k=1}^{K} G(\pi^k) \leq & \frac{\Phi(\pi^{K+1}) - \Phi(\pi^1)}{\nu} + \frac{K}{\sqrt{\gamma\tau}} \sqrt{4m^2F^3 \log(mK/\delta)} + \frac{2K}{\gamma\tau} mF^2 \log(mK/\delta) \\ & + \frac{2m^2FK\gamma}{\nu} + \frac{4(\nu^2+\gamma^2)m^3FK}{\nu}.\end{aligned}$$

Set $\nu = \frac{\sqrt{F}}{m\sqrt{K}}$, $\gamma = \frac{\sqrt{F}}{mK}$, $\tau = K^2$ and we can have

$$\sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} K^{1/2} + m^3 F^2\right).$$

Then by Lemma 12, for $T = K\tau$, we have

$$\text{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3}\right).$$

□

E.4 Lemmas for Semi-bandit Feedback

Lemma 14. *Let $\mathcal{F}_i = \bigcup_{a_i \in \mathcal{A}_i} a_i$. For any player i, if $\tilde{\rho}_i$ is the output of Algorithm 5 and π_i^k contains a mixture of $\tilde{\rho}_i$ with weight γ, then we have $\mathbb{P}_{a_i \sim \pi_i^k}(f \in a_i) \geq \frac{\gamma}{2F}$ for any $f \in \mathcal{F}_i$.*

Proof. By Algorithm 5, whenever a new action is added into $\widetilde{\mathcal{A}}_i$, it contains facility not appeared in current $\widetilde{\mathcal{A}}_i$. Then, since there are at most $|\mathcal{F}_i| \leq F$ distinct facilities in the action set $\mathcal{A}_i$, the final $\widetilde{\mathcal{A}}_i$ must satisfy $|\widetilde{\mathcal{A}}_i| \leq F$. Therefore, $\tilde{\rho}_i$ is a valid distribution over $\mathcal{A}_i$.

Since π_i^k contains a mixture of $\tilde{\rho}_i$ with weight γ, for any $a_i \in \mathcal{A}_i$, we have $\pi_i^k(a_i) \geq \gamma \tilde{\rho}_i(a_i)$. Thus, we have

$$\mathbb{P}_{a_i \sim \pi_i^k}(f \in a_i) = \sum_{a_i \in \mathcal{A}_i} \pi_i^k(a_i) \mathbb{1}\{f \in a_i\}$$

Algorithm 5 Compute Exploration Distribution $\tilde{\rho}_i$

Input: $\mathcal{A}_i$, player i-th action set
Initialize $\widetilde{\mathcal{A}}_i \leftarrow \emptyset$
for a_i in $\mathcal{A}_i$ **do**
  if $\exists f \in a_i$ such that $f \notin \bigcup_{a_i' \in \widetilde{\mathcal{A}}_i} a_i'$ **then**
    $\widetilde{\mathcal{A}}_i \leftarrow \widetilde{\mathcal{A}}_i \cup \{a_i\}$
  if $\mathcal{F}_i = \bigcup_{a_i' \in \widetilde{\mathcal{A}}_i} a_i'$ **then**
    break
Assign $\tilde{\rho}_i(a_i) \leftarrow \frac{1}{2F}$ for each $a_i \in \widetilde{\mathcal{A}}_i$
Assign remaining probability mass arbitrarily to actions in $\mathcal{A} \setminus \widetilde{\mathcal{A}}_i$
return $\tilde{\rho}_i$

$$
\begin{aligned}
&\geq \gamma \sum_{a_i \in \mathcal{A}_i} \tilde{\rho}_i(a_i) \mathbb{1}\{f \in a_i\} \\
&\geq \gamma \sum_{a_i \in \widetilde{\mathcal{A}}_i} \tilde{\rho}_i(a_i) \mathbb{1}\{f \in a_i\} \\
&= \frac{\gamma}{2F} \sum_{a_i \in \widetilde{\mathcal{A}}_i} \mathbb{1}\{f \in a_i\} \geq \frac{\gamma}{2F}.
\end{aligned}
$$

The last inequality above holds since by construction, $\widetilde{\mathcal{A}}_i$ contains all facilities contained in $\mathcal{A}_i$.

□

Lemma 15. *If π_i^k contains a mixture of $\tilde{\rho}_i$ given in Algorithm 5 with weight γ. Then, the IPS estimator $[\widetilde{\theta}_i^k(\pi^k)]_f$ satisfies*

$$
\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f\right] = [\theta_i(\pi^k)]_f, \quad |[\widetilde{\theta}_i^{k,t}(\pi^k)]_f| \leq \frac{2F}{\gamma}, \quad \textit{and} \quad \mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f^2\right] \leq \frac{2F}{\gamma}.
$$

Proof. For the first property, since $\mathbb{E}_k\left[r^{k,t,f} \mid \boldsymbol{a}^{k,t}\right] = r^f(n^f(a_i^{k,t}, a_{-i}^{k,t}))$ and $\boldsymbol{a}^{k,t} \sim \pi^k$, We have

$$
\begin{aligned}
&\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f\right] \\
=&\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\frac{r^f(n^f(a_i, a_{-i}))\mathbb{1}\{f \in a_i\}}{\mathbb{P}_{a_i'\sim\pi_i^k}(f \in a_i')}\right] \\
=&\frac{1}{\mathbb{P}_{a_i'\sim\pi_i^k}(f \in a_i')} \cdot \mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i, a_{-i}))\mathbb{1}\{f \in a_i\} \mid a_{-i}\right]\right] \\
=&\frac{1}{\mathbb{P}_{a_i'\sim\pi_i^k}(f \in a_i')} \cdot \mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i, a_{-i})) \mid a_{-i}, f \in a_i\right] \mathbb{P}_{a_i\sim\pi_i^k}(f \in a_i \mid a_{-i})\right] \\
\overset{\text{(i)}}{=}&\frac{\mathbb{P}_{a_i\sim\pi_i^k}(f \in a_i)}{\mathbb{P}_{a_i'\sim\pi_i^k}(f \in a_i')} \cdot \mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[r^f(n^f(a_{-i}) + 1)\right] \\
=&[\theta_i(\pi^k)]_f.
\end{aligned}
$$

The equality (i) above holds because $\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i, a_{-i})) \mid a_{-i}, f \in a_i\right] = r^f(n^f(a_{-i}) + 1)$ and $f \in a_i$ does not depend on a_{-i}.

For the second property, since $\mathbb{P}_{a_i\sim\pi_i^k}(f \in a_i) \geq \frac{\gamma}{2F}$ by Lemma 14 and $r^{k,t,f} \in [0, 1]$, we can immediately have $|[\widetilde{\theta}_i^{k,t}(\pi^k)]_f| \leq \frac{2F}{\gamma}$.

For the third property, we have

$$
\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f^2\right] = \frac{\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[r^f(n^f(a_i, a_{-i}))^2\mathbb{1}\{f \in a_i\}\right]}{\mathbb{P}_{a_i'\sim\pi_i^k}(f \in a_i')^2}
$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\mathbb{1}\left\{f \in a_i\right\}\right]}{\mathbb{P}_{a_i'\sim\pi_i^k}\left(f \in a_i'\right)^2} \\
&= \frac{\mathbb{P}_{a_i\sim\pi_i^k}\left(f \in a_i\right)}{\mathbb{P}_{a_i'\sim\pi_i^k}\left(f \in a_i'\right)^2} \\
&\leq \frac{2F}{\gamma}.
\end{aligned}
$$

□

F Analysis for Algorithm 4

F.1 Bellman Equations for Genera-sum Markov Games

Before analyzing Algorithm 4, we first give a brief review of the Bellman equations for general-sum Markov games. These equations are well-known among the literature Bai and Jin [2020], Liu et al. [2021], Jin et al. [2021b].

Fixed policies. Given a fixed policy π, for any $(h, i, s, \boldsymbol{a}) \in [H] \times [m] \times \mathcal{S} \times \mathcal{A}$, it holds that

$$
Q_{h,i}^{\pi}(s, \boldsymbol{a}) = (r_{h,i} + \mathbb{P}_h V_{h+1,i}^{\pi})(s, \boldsymbol{a}), \quad V_{h,i}^{\pi} = \mathbb{E}_{\boldsymbol{a}'\sim\pi_h(\cdot|s)}\left[Q_{h,i}^{\pi}(s, \boldsymbol{a}')\right], \tag{14}
$$

where $V_{H+1,i}^{\pi}(s) = 0$ for any $(i, s) \in [m] \times \mathcal{S}$.

Best responses. Given a fixed policy π, define the best response value functions for player i as $Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}) = \max_{\pi_i\in\Delta(\mathcal{A}_i)} Q_{h,i}^{\pi_i,\pi_{-i}}(s, \boldsymbol{a})$ and $V_{h,i}^{\dagger,\pi_{-i}}(s) = \max_{\pi_i\in\Delta(\mathcal{A}_i)} V_{h,i}^{\pi_i,\pi_{-i}}(s)$. Then, for any $(h, i, s, \boldsymbol{a}) \in [H] \times [m] \times \mathcal{S} \times \mathcal{A}$, it holds that

$$
\begin{aligned}
&Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}) = (r_{h,i} + \mathbb{P}_h V_{h+1,i}^{\dagger,\pi_{-i}})(s, \boldsymbol{a}), \\
&V_{h,i}^{\dagger,\pi_{-i}}(s) = \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}'\sim(\nu,\pi_{h,-i})(\cdot|s)}\left[Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}')\right],
\end{aligned} \tag{15}
$$

where $V_{H+1,i}^{\dagger,\pi_{-i}}(s) = 0$ for any $(i, s) \in [m] \times \mathcal{S}$.

F.2 Proof of Theorem 3

Recall that the update rule in Algorithm 4 is

$$
\overline{Q}_{h,i}^{k}(s, \boldsymbol{a}) \leftarrow \min\left\{(\hat{r}_{h,i}^{k} + \widehat{\mathbb{P}}_h^k \overline{V}_{h+1,i}^{k} + b_h^k)(s, \boldsymbol{a}), HF\right\}, \quad \overline{V}_{h,i}^{k}(s) \leftarrow \mathbb{E}_{\boldsymbol{a}\sim\pi_h^k}[\overline{Q}_{h,i}^{k}(s, \boldsymbol{a})].
$$

Similar to the proof of Theorem 1, we define auxiliary value functions

$$
\underline{Q}_{h,i}^{k}(s, \boldsymbol{a}) \leftarrow \max\left\{(\hat{r}_{h,i}^{k} + \widehat{\mathbb{P}}_h^k \underline{V}_{h+1,i}^{k} - b_h^k)(s, \boldsymbol{a}), 0\right\}, \quad \underline{V}_{h,i}^{k}(s) \leftarrow \mathbb{E}_{\boldsymbol{a}\sim\pi_h^k}[\underline{Q}_{h,i}^{k}(s, \boldsymbol{a})]. \tag{16}
$$

We now begin to prove the first part of Theorem 3.

Proof of Theorem 3. **Step 1.** We first consider the setting of semi-bandit feedback. Assume the result in Lemma 17 holds since it is a high-probability event. Then, for any $(k, s) \in [K] \times \mathcal{S}$, it holds that

$$
\max_{i\in[m]}\left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k}\right)(s) \leq \max_{i\in[m]}\left(\overline{V}_{1,i}^{k} - \underline{V}_{1,i}^{k}\right)(s) + H\epsilon.
$$

By the update rules in Algorithm 4, we can notice the following recursive relations

$$
\begin{aligned}
&(\overline{Q}_{h,i}^{k} - \underline{Q}_{h,i}^{k})(s, \boldsymbol{a}) \leq \min\left\{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^{k} - \underline{V}_{h+1,i}^{k})(s, \boldsymbol{a}) + 2b_h^k(s, \boldsymbol{a}), HF\right\}, \\
&(\overline{V}_{h,i}^{k} - \underline{V}_{h,i}^{k})(s) = \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s)}\left[(\overline{Q}_{h,i}^{k} - \underline{Q}_{h,i}^{k})(s, \boldsymbol{a}')\right].
\end{aligned}
$$

Thus, we define $\widetilde{V}_{H+1}^k(s) = 0$ for any $s \in \mathcal{S}$ and $\widetilde{Q}_h^k$, $\widetilde{V}_h^k$ recursively as

$$\widetilde{Q}_h^k(s, \boldsymbol{a}) = \min\left\{(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s, \boldsymbol{a}) + 2b_h^k(s, \boldsymbol{a}), HF\right\}, \quad \widetilde{V}_h^k(s) = \mathbb{E}_{\boldsymbol{a}' \sim \pi_h^k(\cdot|s)}\left[\widetilde{Q}_h^k(s, \boldsymbol{a}')\right]. \tag{17}$$

Obviously, we have $\max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \le \widetilde{V}_{H+1}^k$. Then, by inductively assuming the same relation holds for $h+1$, we can have

$$\begin{aligned}
\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s, \boldsymbol{a}) =& \min\left\{\max_{i\in[m]} \widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^k - \underline{V}_{h+1,i}^k)(s, \boldsymbol{a}) + 2b_h^k(s, \boldsymbol{a}), HF\right\} \\
\le& \min\left\{(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s, \boldsymbol{a}) + 2b_h^k(s, \boldsymbol{a}), HF\right\} \\
=& \widetilde{Q}_h^k(s, \boldsymbol{a}), \\
\max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \le& \mathbb{E}_{\boldsymbol{a}' \sim \pi_h^k(\cdot|s)}\left[\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s, \boldsymbol{a}')\right] \\
\le& \mathbb{E}_{\boldsymbol{a}' \sim \pi_h^k(\cdot|s)}\left[\widetilde{Q}_h^k(s, \boldsymbol{a}')\right] \\
=& \widetilde{V}_h^k(s).
\end{aligned}$$

Therefore, by induction, for any $h \in [H]$, we have

$$\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s, \boldsymbol{a}) \le \widetilde{Q}_h^k(s, \boldsymbol{a}), \quad \max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \le \widetilde{V}_h^k(s).$$

As a result, we have

$$\text{Nash-Regret}(K) = \sum_{k=1}^K \max_{i\in[m]}\left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k}\right)(s) \le \sum_{k=1}^K \widetilde{V}_1^k(s_1) + HK\epsilon.$$

Step 2, Semi-bandit Feedback. We define the martingale difference sequences

$$\begin{aligned}
\mathcal{M}_h^k(\widetilde{Q}) &= \mathbb{E}_{\boldsymbol{a}' \sim \pi_h^k(\cdot|s_h^k)}\left[\widetilde{Q}_h^k(s_h^k, \boldsymbol{a}')\right] - \widetilde{Q}_h^k(s_h^k, \boldsymbol{a}_h^k), \\
\mathcal{M}_h^k(\widetilde{V}) &= (\mathbb{P}_h \widetilde{V}_{h+1}^k)(s_h^k, \boldsymbol{a}_h^k) - \widetilde{V}_{h+1}^k(s_{h+1}^k).
\end{aligned}$$

It is not hard to check that $\mathcal{M}_h^k(\widetilde{Q})$ and $\mathcal{M}_h^k(\widetilde{V})$ are both indeed martingale difference sequences with respect to the history till episode k and time step h.

With these definitions, we can now decompose the regret bound as

$$\begin{aligned}
\widetilde{V}_h^k(s_h^k) =& \mathbb{E}_{\boldsymbol{a}' \sim \pi_h^k(\cdot|s_h^k)}\left[\widetilde{Q}_h^k(s_h^k, \boldsymbol{a}')\right] && \text{(By (17))} \\
=& \mathcal{M}_h^k(\widetilde{Q}) + \widetilde{Q}_h^k(s_h^k, \boldsymbol{a}_h^k) \\
\le& \mathcal{M}_h^k(\widetilde{Q}) + 2b_h^k(s_h^k, \boldsymbol{a}_h^k) + (\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s_h^k, \boldsymbol{a}_h^k) && \text{(By (17))} \\
\overset{\text{(i)}}{\le}& \mathcal{M}_h^k(\widetilde{Q}) + 3b_h^k(s_h^k, \boldsymbol{a}_h^k) + (\mathbb{P}_h \widetilde{V}_{h+1}^k)(s_h^k, \boldsymbol{a}_h^k) \\
=& \mathcal{M}_h^k(\widetilde{Q}) + \mathcal{M}_h^k(\widetilde{V}) + 3b_h^k(s_h^k, \boldsymbol{a}_h^k) + \widetilde{V}_{h+1}^k(s_{h+1}^k)
\end{aligned}$$

The above inequality (i) holds by applying Lemma 17 and the fact $\widetilde{V}_h^k(s) \le HF$, which comes from the definition in (17). Then, by unrolling this relation from $h = 1$ to $h = H$ and noticing $\widetilde{V}_{H+1}^k = \mathbf{0}$, we can have

$$\begin{aligned}
&\text{Nash-Regret}(K) \le \sum_{k=1}^K \widetilde{V}_1^k(s_1) + HK\epsilon \\
\le& \sum_{k=1}^K \sum_{h=1}^H \left(\mathcal{M}_h^k(\widetilde{Q}) + \mathcal{M}_h^k(\widetilde{V}) + 3b_h^k(s_h^k, \boldsymbol{a}_h^k)\right) + HK\epsilon && (18) \\
\le& \widetilde{\mathcal{O}}\left(HF\sqrt{T}\right) + 3\sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, \boldsymbol{a}_h^k) && \text{(By Azuma-Hoeffding inequality and taking } \epsilon = 1/T.)
\end{aligned}$$

$$
\begin{aligned}
&\leq \widetilde{\mathcal{O}}\left(HF\sqrt{T}\right) + 6HF\sum_{f\in\mathcal{F}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sqrt{\frac{S^f\iota}{N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))\vee 1}} + \sqrt{\frac{\iota}{N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))\vee 1}}\right)\\
&\qquad + 6HF\sum_{f\neq f'}S^fS^{f'}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\iota^2}{\left(N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'},n^{f'}(\boldsymbol{a}_h^{k,f'}))\right)\vee 1}}\\
&\leq \widetilde{\mathcal{O}}\left(HF\sqrt{T}\right) + \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}HFS^f\sqrt{mHT}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(By Lemma 20 and 21)}\\
&\leq \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}FS^f\sqrt{mH^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).
\end{aligned}
$$

Step 3, Bandit Feedback. In the setting of bandit feedback, we only modify the reward estimator $\tilde{r}_{h,i}^k$ and its corresponding bonus term $\tilde{b}_h^{k,\mathrm{r}}$. Thus, by going through the proof of Lemma 17, we can notice that to have the same result for bandit feedback, it suffice to use Lemma 18 to show that the reward estimation error is bounded by the reward bonus term.

Then, by the inequality (18), we can notice that to achieve the final Nash-regret bound, we only need to bound the summation $\sum_{k=1}^K\sum_{h=1}^H\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k)$, which is

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k) &\leq \sqrt{\beta_K}\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{i\in[m]}\left\|A_i(s_h^k,\boldsymbol{a}_h^k)\right\|_{\left(V_h^k\right)^{-1}} \qquad \text{(By definition of } \tilde{b}_h^{k,\mathrm{r}} \text{ in (12).)}\\
&\leq \left(\sqrt{d}+\sqrt{d\log\left(1+\frac{mKF}{d}\right)+\iota}\right)\widetilde{\mathcal{O}}\left(H\sqrt{dFK}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(By definition of } \beta_k \text{ and Lemma 19.)}\\
&\leq \widetilde{\mathcal{O}}\left(d\sqrt{HFT}\right)\\
&= \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}mS^f\sqrt{HFT}\right). \qquad \text{(Since } d = m\textstyle\sum_{f\in\mathcal{F}}S^f\text{.)}
\end{aligned}
$$

Therefore, by (18), with $\epsilon = 1/T$, under bandit feedback, we have

$$
\begin{aligned}
&\text{Nash-Regret}(K)\\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathcal{M}_h^k(\widetilde{Q}) + \mathcal{M}_h^k(\widetilde{V}) + 3b_h^k(s_h^k,\boldsymbol{a}_h^k)\right)\\
&\leq \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}FS^f\sqrt{mH^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right) + \sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k)\\
&\leq \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}\left(\sqrt{mH^3}F + m\sqrt{HF}\right)S^f\sqrt{T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).
\end{aligned}
$$

□

F.3 Lemmas for Semi-bandit Feedback

The following two lemmas shows that our value function estimations are indeed optimistic.

Lemma 16. *With probability at least $1-\delta$, simultaneously for arbitrary value function $V\in[0,HF]^{\mathcal{S}}$ and any tuple $(k,h,s,\boldsymbol{a})$, it holds that $|(\widehat{\mathbb{P}}_h^k-\mathbb{P}_h)V(s,\boldsymbol{a})|\leq b_h^{k,\mathrm{pv}}(s,\boldsymbol{a})$, where $b_h^{k,\mathrm{pv}}(s,\boldsymbol{a})$ is defined in (7).*

Proof. We define $\mathbb{P}_h^f$ to be the operator such that for some value function $V^f:\mathcal{S}^f\mapsto\mathbb{R}$, we have $(\mathbb{P}_h^fV^f)(s,\boldsymbol{a}) = \mathbb{E}_{s'^f\sim P_h^f(\cdot|s^f,n^f(\boldsymbol{a}))}\left[V^f(s'^f)\right]$. We also define $\widehat{\mathbb{P}}_h^{k,f}$ similarly. Then, by definition of

our transition kernel, for operators $\mathbb{P}_h$ and $\widehat{\mathbb{P}}_h^k$, it holds that

$$\mathbb{P}_h = \prod_{f\in\mathcal{F}} \mathbb{P}_h^f \quad \text{and} \quad \widehat{\mathbb{P}}_h^k = \prod_{f\in\mathcal{F}} \widehat{\mathbb{P}}_h^{k,f}.$$

Therefore, by Lemma E.1 in Chen et al. [2020], since $\|V\|_\infty \le HF$, we have

$$\begin{aligned}|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V(s,\boldsymbol{a})| \le \sum_{f\in\mathcal{F}} &\left|(\widehat{\mathbb{P}}_h^{k,f} - \mathbb{P}_h^f)\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})\right| \\ &+ 2HF\sum_{f\neq f'} \mathrm{errp}_h^{k,f}(s,\boldsymbol{a})\cdot \mathrm{errp}_h^{k,f'}(s,\boldsymbol{a}),\end{aligned} \tag{19}$$

where $\mathrm{errp}_h^{k,f}(s,\boldsymbol{a}) = \|\widehat{P}_h^{k,f}(\cdot \mid s^f, n^f(\boldsymbol{a})) - P_h^f(\cdot \mid s^f, n^f(\boldsymbol{a}))\|_1$.

Now, notice that $\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})$ can be seen as some value function from $\mathcal{S}^f$ to $[0, HF]$. Therefore, by Lemma 12 in Bai and Jin [2020], with probability at least $1-\frac{\delta}{2}$, simultaneously for any V and $(k,h,s,\boldsymbol{a})$, it holds that

$$\left|(\widehat{\mathbb{P}}_h^{k,f} - \mathbb{P}_h^f)\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})\right| \le 2HF\sqrt{\frac{S^f\iota}{N_h^{k,f}(s^f,n^f(\boldsymbol{a}))\vee 1}},$$

where $\iota = 2\log(4(m+1)(\sum_{f\in\mathcal{F}}S^f)T/\delta)$. Meanwhile, by standard Hoeffding's inequality and union bound, with probability at least $1-\frac{\delta}{2}$, simultaneously for any $(k,h,s,\boldsymbol{a})$, it holds that

$$\mathrm{errp}_h^{k,f} \le S^f\sqrt{\frac{\iota}{N_h^{k,f}(s^f,n^f(\boldsymbol{a}))\vee 1}}.$$

Finally, by plugging above two concentration inequalities back into (19), we can have

$$|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V(s,\boldsymbol{a})| \le b_h^{k,\mathrm{pv}}(s,\boldsymbol{a}).$$

□

Lemma 17. *With probability at least $1-\delta$, for any $(k,h,i,s,\boldsymbol{a}) \in [K]\times[H]\times[m]\times\mathcal{S}\times\mathcal{A}$, it holds that*

$$\overline{Q}_{h,i}^k(s,\boldsymbol{a}) \ge Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) - (H-h)\epsilon, \quad \underline{Q}_{h,i}^k(s,\boldsymbol{a}) \le Q_{h,i}^{\pi^k}(s,\boldsymbol{a}), \tag{20}$$

$$\overline{V}_{h,i}^k(s) \ge V_{h,i}^{\dagger,\pi_{-i}^k}(s) - (H-h+1)\epsilon, \quad \underline{V}_{h,i}^k(s) \le V_{h,i}^{\pi^k}(s.), \tag{21}$$

where $\underline{Q}_{h,k}^k$ and $\underline{V}_{h,i}^k$ are defined in (16).

Proof. The proof is adapted from Liu et al. [2021] and goes by induction from $h=H+1$ to $h=1$. We can see that inequalities (21) obviously hold when $h=H+1$ since by definition we have $\overline{V}_{H+1,i}^k(s) = \underline{V}_{H+1,i}^k(s) = 0$ for any (k,i,s). Now, suppose inequalities (21) hold for $h+1$. Then, if we have $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) = HF$, it holds trivially that $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) \ge Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a})$. Otherwise, by Bellman equations (15) and update rule in Algorithm 4, we have

$$\begin{aligned}&\overline{Q}_{h,i}^k(s,\boldsymbol{a}) - Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) \\ =&(\hat{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a}) + (\widehat{\mathbb{P}}_h^k\overline{V}_{h+1,i}^k)(s,\boldsymbol{a}) - (\mathbb{P}_h V_{h+1,i}^{\dagger,\pi_{-i}^k})(s,\boldsymbol{a}) + b_h^k(s,\boldsymbol{a}) \\ =&\underbrace{(\hat{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})}_{\text{(A)}} + \underbrace{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^k - V_{h+1,i}^{\dagger,\pi_{-i}^k})(s,\boldsymbol{a})}_{\text{(B)}} + \underbrace{((\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1,i}^{\dagger,\pi_{-i}^k})(s,\boldsymbol{a})}_{\text{(C)}} + b_h^k(s,\boldsymbol{a}).\end{aligned}$$

Now, recall that $b_h^k(s,\boldsymbol{a}) = b_h^{k,\mathrm{pv}}(s,\boldsymbol{a}) + b_h^{k,\mathrm{r}}(s,\boldsymbol{a})$. By reward definition in congestion game, we have

$$(\hat{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a}) = \sum_{f\in a_i}(\hat{r}_{h,i}^{k,f}(s^f,n^f(\boldsymbol{a})) - r_{h,i}^f(s^f,n^f(\boldsymbol{a}))).$$

Thus, by using standard Hoefding's inequality and union bound, we can immediately have $|(\mathrm{A})| \leq b_h^{k,\mathrm{r}}(s,\boldsymbol{a})$. Then, since $V_{h,i}^{\dagger,\pi_{-i}^k} \in [0, HF]^{\mathcal{S}}$, by Lemma 16, we have $|(\mathrm{C})| \leq b_h^{k,\mathrm{pv}}(s,\boldsymbol{a})$. That is, we have $(\mathrm{A}) + (\mathrm{C}) + b_h^k(s,\boldsymbol{a}) \geq 0$.

Then, by inductive hypothesis, we know that $\overline{V}_{h+1,i}^k \geq V_{h+1,i}^{\dagger,\pi_{-i}^k} - (H-h)\epsilon$, which implies $(\mathrm{B}) \geq 0$. Therefore, we have $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) - Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) \geq -(H-h)\epsilon$.

For $\overline{V}_{h,i}^k$ and $V_{h,i}^{\dagger,\pi_{-i}^k}$, we notice that in Algorithm 4, π^k is computed as the ϵ-approximate Nash equilibrium of $(\overline{Q}_{h,1}^k, \ldots, \overline{Q}_{h,m}^k)$. Therefore, it holds that

$$\overline{V}_{h,i}^k(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi_h^k(\cdot|s)}\left[\overline{Q}_{h,i}^k(s,\boldsymbol{a})\right] \geq \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}'\sim(\nu,\pi_{h,-i}^k)(\cdot|s)}\left[\overline{Q}_{h,i}^k(s,\boldsymbol{a}')\right] - \epsilon.$$

By Bellman equations (15), we also have

$$V_{h,i}^{\dagger,\pi_{-i}^k}(s) = \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}'\sim(\nu,\pi_{h,-i}^k)(\cdot|s)}\left[Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}')\right].$$

Since $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) - Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) \geq -(H-h)\epsilon$, we immediately have $\overline{V}_{h,i}^k(s) - V_{h,i}^{\dagger,\pi_{-i}^k}(s) \geq -(H-h+1)\epsilon$. Thus, by induction, we have that $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) \geq Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) - (H-h)\epsilon$ and $\overline{V}_{h,i}^k(s) \geq V_{h,i}^{\dagger,\pi_{-i}^k}(s) - (H-h+1)\epsilon$ for all $h \in [H]$.

The inequalities for $\underline{V}_{h,i}^k$ and $\underline{Q}_{h,i}^k$ can be proved similarly. □

F.4 Additional Lemmas for Bandit Feedback

The following lemma shows that the reward estimation error can be bounded by the reward bonus term.

Lemma 18. *With probability at least $1-\delta$, simultaneously for all $(i,k,h,s,\boldsymbol{a})$, it holds that $|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| \leq \tilde{b}_h^{k,\mathrm{r}}(s,\boldsymbol{a})$, where $\tilde{r}_{h,i}^k$ and $\tilde{b}_h^{k,\mathrm{r}}$ are defined in (11) and (12).*

Proof. The proof is extremely similar to Lemma 3. By construction, we have

$$\begin{aligned}
|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| &= \left|\left\langle A_i(s,\boldsymbol{a}), \widehat{\theta}_h - \theta_h \right\rangle\right| \\
&\leq \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \left\|\widehat{\theta}_h - \theta_h\right\|_{V_h^k} \\
&\leq \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \left(\|\theta_h\|_2 + \sqrt{\log\left(\det(V_h^k)\right) + \iota}\right).
\end{aligned}$$

(By Theorem 20.5 in Lattimore and Szepesvári [2020].)

Since each element in θ_h is bounded in $[0,1]$ by construction, we have $\|\theta_h\|_2 \leq \sqrt{d}$.

Then, by Lemma 4, we have $\det\left(V_h^k\right) \leq \left(1 + \frac{mkF}{d}\right)^d$ since by construction $\|A_i(s,\boldsymbol{a})\|_2^2 \leq F$.

Finally, to make this bound valid for all player $i \in [m]$, we only need to take maximization over $i \in [m]$. Therefore, with probability at least $1-\delta$, we have

$$|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| \leq \max_{i\in[m]} \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \sqrt{\beta_k} = \tilde{b}_h^{k,\mathrm{r}}(s,\boldsymbol{a}),$$

where $\sqrt{\beta_k} = \sqrt{d} + \sqrt{d\log\left(1 + \frac{mkF}{d}\right) + \iota}$. □

The follow lemma bound the sum of reward bonus under bandit feedback.

Lemma 19. *For any $h \in [H]$, it holds that*

$$\sum_{k=1}^K \max_{i\in[m]} \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_{(V_h^k)^{-1}} \leq \widetilde{\mathcal{O}}\left(\sqrt{dFK}\right),$$

where $d = m\sum_{f\in\mathcal{F}} S^f$.

Proof. First, since $V_h^k = I + \sum_{k'=1}^{k-1}\sum_{i=1}^m A_i(s_h^{k'}, \boldsymbol{a}_h^{k'})A_i(s_h^{k'}, \boldsymbol{a}_h^{k'})^\top$, we have $V_h^k \succeq I$ and thus $\left(V_h^k\right)^{-1} \preceq I$. Therefore, we have

$$\left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_{\left(V_h^k\right)^{-1}} \le \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_I = \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_2 \le \sqrt{F}.$$

For simplicity, let $A_{h,i}^k = A_i(s_h^k, \boldsymbol{a}_h^k)$. Then, as a result, we have

$$\begin{aligned}
\sum_{k=1}^K \max_{i\in[m]} \left\|A_{h,i}^k\right\|_{\left(V_h^k\right)^{-1}} &= \sum_{k=1}^K \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|_{\left(V_h^k\right)^{-1}}, \sqrt{F}\right\} \\
&\le \sqrt{K\sum_{k=1}^K \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|^2_{\left(V_h^k\right)^{-1}}, F\right\}} \\
&\le \sqrt{FK\sum_{k=1}^K \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|^2_{\left(V_h^k\right)^{-1}}, 1\right\}} \\
&\le \sqrt{2FKd\log\left(1+\frac{mKF}{d}\right)} \qquad \text{(By Lemma 4.)} \\
&= \widetilde{\mathcal{O}}\left(\sqrt{dFK}\right).
\end{aligned}$$

□

F.5 Technical Lemmas

Lemma 20. *For any $f \in \mathcal{F}$, it holds that*

$$\sum_{k=1}^K\sum_{h=1}^H \sqrt{\frac{1}{N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k)) \vee 1}} \le \widetilde{\mathcal{O}}\left(\sqrt{mHS^fT}\right).$$

Proof. Here, we have

$$\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \sqrt{\frac{1}{N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k)) \vee 1}} &= \sum_{h=1}^H \sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^m \sum_{\ell=1}^{N_h^{K,f}(s^f,n)} \sqrt{\frac{1}{\ell}} \\
&\le 2\sum_{h=1}^H \sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^m \sqrt{N_h^{K,f}(s^f,n)} \qquad \text{(By standard technique)} \\
&\le 2\sqrt{(m+1)HS^f\sum_{h=1}^H \sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^m N_h^{K,f}(s^f,n)} \\
&= \widetilde{\mathcal{O}}\left(\sqrt{mHS^fT}\right).
\end{aligned}$$

The last line above holds because $\sum_{h=1}^H \sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^m N_h^{K,f}(s^f,n) = T$. This is based on a pigeonhole principle argument. In particular, whenever the players take one more action, for any $f \in \mathcal{F}$, the count for some tuple (h, s^f, n) will increase exactly by 1. □

Lemma 21 (Chen et al. [2020])**.** *For any $f, f' \in \mathcal{F}$ and $f \neq f'$, it holds that*

$$\sum_{k=1}^K\sum_{h=1}^H \sqrt{\frac{1}{\left(N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'}, n^{f'}(\boldsymbol{a}_h^{k,f'}))\right) \vee 1}} \le \widetilde{\mathcal{O}}\left(m^2HS^fS^{f'}\right).$$

Proof. We define the joint empirical counter

$$N_h^{k,f,f'}(s^f, s^{f'}, n, n') = \sum_{k'=1}^k \mathbb{1}\left\{(s_h^{k',f}, s_h^{k',f'}, n^f(\boldsymbol{a}_h^{k'}), n^{f'}(\boldsymbol{a}_n^{k'})) = (s^f, s^{f'}, n, n')\right\}.$$

Obviously, we have $N_h^{f,f'}(s^f, s^{f'}, n, n') \leq \min\left\{N_h^{k,f}(s^f, n), N_h^{k,f'}(s^{f'}, n')\right\}$, which implies

$$N_h^{k,f,f'}(s, s^{f'}, n, n') \leq \sqrt{N_h^{k,f}(s^f, n) N_h^{k,f'}(s^{f'}, n')}.$$

Therefore, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{\left(N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'}, n^{f'}(\boldsymbol{a}_h^{k,f'}))\right)\vee 1}} \\
\leq &\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_h^{k,f,f'}(s_h^{k,f}, s_h^{k,f'}, n^f(\boldsymbol{a}_h^k), n^{f'}(\boldsymbol{a}_h^k))\vee 1} \\
= &\sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{s^{f'}\in\mathcal{S}^{f'}}\sum_{n=0}^{m}\sum_{n'=0}^{m}\sum_{\ell=1}^{N_h^{K,f,f'}(s^f,s^{f'},n,n')}\frac{1}{\ell} \\
= &\widetilde{\mathcal{O}}\left(m^2 H S^f S^{f'}\right).
\end{aligned}$$

□