Convergence to Equilibrium of No-regret Dynamics in Congestion Games

Volkan Cevher[1], Wei Chen[2], Leello Dadi[1], Jing Dong[3], Ioannis Panageas[4], Stratis Skoulakis[1], Luca Viano[1], Baoxiang Wang[3], Siwei Wang[2], and Jingyu Wu[5]

[1] École polytechnique fédérale de Lausanne (EPFL) first.last@epfl.ch
[2] Microsoft Research Asia {weic,siweiwang}@microsoft.com
[3] The Chinese University of Hong Kong, Shenzhen jingdong@link.cuhk.edu.cn
bxiangwang@cuhk.edu.cn
[4] Computer Science, University of California, Irvine ipanagea@ics.uci.edu
[5] University of Science and Technology of China
wujingyu@mail.ustc.edu.cn

Abstract. The congestion game is a powerful model that encompasses a range of engineering systems such as traffic networks and resource allocation. It describes the behavior of a group of agents who share a common set of F facilities and take actions as subsets of k facilities. In this work, we study the online formulation of congestion games, where agents participate in the game repeatedly and observe feedback with randomness. We note that this paper is the result of the merging of [24] arXiv:2306.13673 and [19] arXiv:2401.09628. In [24], we propose CongestEXP, a decentralized algorithm that is based on the classic exponential weights method. By maintaining weights on the facility level, the regret bound of CongestEXP avoids the exponential dependence on the size of possible facility sets, i.e., $\binom{F}{k} \approx F^k$, and scales only linearly with F. Specifically, we show that CongestEXP attains a regret upper bound of $O(kF\sqrt{T})$ for every individual player, where T is the time horizon. If a strict Nash equilibrium exists, we show that CongestEXP can converge to the strict Nash policy almost exponentially fast in $O(F\exp(-t^{1-\alpha}))$, where t is the number of iterations and $\alpha \in (1/2, 1)$. In [19], we present an online learning algorithm in the bandit feedback model that, once adopted by all agents of a congestion game, results in game-dynamics that converge to an ϵ-approximate Nash Equilibrium in a polynomial number of rounds with respect to $1/\epsilon$, the number of players and the number of available resources. The proposed algorithm also guarantees sublinear regret to any agent adopting it. As a result, our work answers an open question from [17] and extends the recent results of [37] to the bandit feedback model.

Keywords: Congestion Game · Convergence to Nash equilibrium · Online Learning.

1 Introduction

Congestion games are a class of general-sum games that can be used to describe the behavior of agents who share a common set of facilities (resources) [6]. In these games, each player chooses a combination of facilities, and popular facilities will become congested, yielding a lower utility for the players who choose them. Thus, players are incentivized to avoid congestion by choosing combinations that are less popular among the other players. A range of real-world scenarios can be captured by the congestion game model, such as traffic flow, data routing, and wireless communication networks [44,14,45].

In the model of the congestion game, the Nash equilibrium is a popular concept to describe the behavior of selfish players and the dynamics induced by decentralized algorithms. It describes a stable state of the game where no player can improve their utility by unilaterally changing their choice of actions. When a unique Nash equilibrium exists in the congestion game, it can be a reference point for players to coordinate to avoid suboptimal outcomes. Beyond the Nash equilibrium, social welfare is a significant metric, capturing the overall utility or well-being of all players involved. It serves as a crucial benchmark, enabling the evaluation of the efficiency loss incurred when transitioning from centralized to decentralized algorithms.

In the classic one-shot congestion game setting, the Nash equilibrium and the loss of efficiency due to decentralized dynamics have been well studied [42,40]. However, these results do not provide insights into how players arrive at the equilibrium. This motivates the study of congestion games in an online learning framework, where players participate in the game repeatedly at every time step. This framework better models many realistic scenarios, such as the traffic congestion problem in urban areas. In this repeated congestion game setting, players such as drivers in a congested city must choose between different routes to reach their destinations every day. As more drivers use a particular route, the congestion on that route increases, leading to higher travel times and lower utility. In this scenario, players can update their desired route every day to optimize their utility, but the observed utility by each player may be subject to randomness due to uncertainty in the actual congestion situation (e.g., the influence of the weather). All these make it suitable to model the congestion game in an online learning framework.

This paper is the merged version of [19] and [24]. The rest of the paper is summarized as follows. We will first summarize the setting and the contribution of each paper, then we will focus mainly on the results presented in [24]. The full results and proofs of [19] can be found at arXiv:2401.09628.

1.1 Results and Techniques in [19]

Despite the long interest in bandit online learning algorithms for congestion games, the convergence to Nash Equilibrium of such bandit no-regret learning algorithms is not as well explored. The broad question under consideration here is whether or not the uncoordinated selfish behavior of agents can converge to

equilibrium. To the best of our knowledge [17] were the first to provide an update rule (performing under bandit feedback) that once adopted by all agents of an atomic congestion game, the resulting strategies converge to an ϵ-approximate Nash Equilibrium with rate polynomial in n, m and $1/\epsilon$. However, their method does not guarantee the no-regret property. As a result, [17] asked the following question:

Is there a no-regret algorithm, in the bandit feedback model, that once adopted by all agents, results in strategies that converge to an ϵ-approximate Nash Equilibrium in $\mathrm{poly}(n, m, 1/\epsilon)$ *rounds?*

The main contribution of our work consists in providing a positive answer to the open question of [17]. More precisely, we provide an online learning algorithm, called *Online Gradient Descent with Caratheodory Exploration* (BGD − CE), that simultaneously provides both regret guarantees and convergence to Nash Equilibrium.

Informal Theorem *There exists an online learning algorithm (*BGD − CE*) that performs under bandit feedback and guarantees $\mathcal{O}(m^{2.8}T^{4/5})$ regret to any agent that adopts it. Moreover if all agent adopt* BGD − CE*, then the resulting strategies converge to an ϵ-Nash Equilibrium after $\mathcal{O}(n^{13.5}m^{13}/\epsilon^5)$ steps.*

Our proposed online learning algorithm additionally improves the convergence rate of the algorithm of [17].

Informal Theorem *For Network Congestion games in Acyclic Directed Graphs (DAGs), Online Gradient Descent with Caratheodory Exploration, can be implemented in polynomial time.*

The above result follows from strategy spaces admitting polynomial size descriptions in this setting. We further exploit the specific structure of DAGs to compute exact 1-barycentric-spanners, which as noted in [5,9] are not trivial to obtain for DAGs. We underline that exact spanners are not necessary, and the approximate method of [5] is perfectly suitable. However, our approach is simple, more efficient, and can be of independent interest.

Our Techniques The fundamental difficulty in designing no-regret online learning algorithms under bandit feedback is to guarantee that each resource is sufficiently explored. Unfortunately, standard bandit algorithms such as EXP3 [3] result in regret bounds of the form $\mathcal{O}(2^{m/2}\sqrt{T})$, that scale exponentially with respect to m. However, a long line of research in combinatorial bandits has produced algorithms with a regret polynomially dependent on m [5,20,27,7,9,30,34,2]. These algorithms, in order to overcome the exploration problem, use various techniques that can roughly be categorized two camps, simultaneous exploration versus alternating explore-exploit, as described by [1]. However, to the best of our knowledge, none of these algorithms have been shown to converge to NE in congestion games once adopted by all agents.

Our online learning algorithm, guaranteeing both no-regret and convergence to equilibrium, is based on combining Online Gradient Descent [46] with a novel exploration scheme, much like [25]. Our exploration strategy is based on coupling

the notion of barycentric spanners [5] with the notion of Bounded-Away Polytopes proposed by [37] for the semi-bandit case. More precisely, [37] introduced the notion of μ-Bounded Away Polytope which corresponds to the description polytope of the strategy space (convex hull of all pure strategies) with the additional constraint that each resource is selected with probability at least $\mu > 0$. Projecting on this polytope ensures that the variance of the unobserved cost estimators remains bounded. In order to capture bandit estimators, we extend the notion of μ-Bounded Away Polytope to denote the subset of the description polytope for which each point admits a decomposition with least μ weight on a pre-selected barycentric spanner $\mathcal{B}$.

This technique of projecting on μ-Bounded polytopes closely ressembles the *mixing* strategies employed in online learning schemes that have alternating explore-exploit rounds. In those strategies, a fixed measure is added to bias the algorithm's chosen strategy. The projection on μ-Bounded polytopes, however, scales the point before adding a bias, and, in some rounds, does not alter the point. It is therefore a mix of simultaneous and alternating exploration, depending on the round.

Finally, in order to provide a polynomial-time implementation of OGD − CE for Network Congestion Games on Directed Acyclic Graphs we need exploit its well disposed combinatorial structure. In [19], we propose a novel construction of barycentric spanners for DAGs that outputs a 1-barycentric spanner in polynomial time and yields an efficient selfish routing scheme that converges to an equilibrium.

1.2 Results and Techniques in [24]

While there have been various decentralized algorithms that can attain the Nash equilibria efficiently for general online games, they can suffer from a linear dependency on the number of actions when directly applied to the congestion game [21,43,13,29,22,26], which is exponential with k, F. On the other hand, algorithms designed specifically for congestion games either only converge to Nash equilibria asymptotically [31,32,36], on average [18], on best-iterate [38] or require additional assumptions on the structure of the game [10,11]. Moreover, to the best of our knowledge, there is no algorithm that can simultaneously guarantee both low regret and fast convergence to Nash equilibrium for each player. While some online learning algorithms, such as exponential weights, have been shown to converge faster than others due to the specific choice of regularization [26], previous regret results indicate that their guarantees still rely on the exponentially large number of actions, due to their specific form of updates (exponential weighting) [15].

In this paper, we study the online congestion game with semi-bandit and full information feedback. We propose a decentralized algorithm that modifies the celebrated exponential weights algorithm, which can be utilized by each player without additional information about other players' utility. From the individual player's perspective, we show that the algorithm guarantees sublinear individual regret, with respect to the best action in hindsight when holding the other

Table 1: Summary of previous results. Here F denotes the number of facilities, T denotes the time horizon, and k is the number of facilities in an action. The semi-bandit feedback means that the learner can only observe the reward associated with the facility chosen, and full information refers to the case where the rewards associated with all facilities are revealed. Best iterate convergence means that the algorithm guarantees the existence of an iterate that meets the criteria of a Nash equilibrium. On the other hand, the last iterate convergence indicates that the final output of the algorithm constitutes a Nash equilibrium.

	Nash convergence type	Convergence to Nash	Regret
[10]	Last iterate	Asymptotic Semi-bandit	None
[18]	Best iterate	$O(F^{1.5}T^{-1/6})$ Semi-bandit	None
[38]	Best iterate	$O\left(F^{1.4}T^{-1/5}\right)$ Semi-bandit	$O\left(F^2T^{4/5}\right)$ Semi-bandit
Ours	Last iterate	$O\left(F\exp(-T^{1-\alpha})\right)$, $\alpha \in (1/2, 1)$ Full-information	$O\left(kF\sqrt{T}\right)$ Semi-bandit

player's strategy fixed. We remark that the regret is also only linear with respect to the number of facilities. As a result of this, we show that the optimal social welfare can be efficiently approximated, up to an error that is only linear with respect to the number of facilities. When a strict Nash equilibrium exists for the congestion game, we also prove that our algorithm is capable of converging to the strict Nash equilibrium fast, with an almost exponentially fast rate that is only linear with respect to the number of facilities.

2 Related works

2.1 Learning in games

Online learning has a long history that is closely tied to the development of game theory. The earliest literature can be traced back to Brown's proposal on using fictitious play to solve two-player zero-sum games [6]. It is now understood that fictitious play can converge very slowly to Nash equilibrium [23]. On the other side, it has been shown that if each player of a general-sum, multi-player game experiences regret that is at most $f(T)$, the empirical distribution of the joint policy converges to a coarse correlated equilibrium of the game with a rate of $O(f(T)/T)$ [8]. This implies that a variety of online learning algorithms such as Hedge and Follow-The-Regularized-Leader algorithms can converge to coarse correlated equilibria at a rate of $O(1/\sqrt{T})$.

While the standard no-regret learning dynamic can guarantee convergence to equilibria, it has been shown that more specialized no-regret learning protocols can do better [21,43,13,29,22]. It has also been shown that when strict pure

Nash equilibria are present, algorithms that are based on entropic regularization (e.g. exponential weights) can converge fast to the equilibria [15,26]. Moreover, such convergence rate holds for a variety of different feedback models, from full information to bandit feedback.

Though all of the above-mentioned methods are applicable to congestion games, the results usually involve a linear dependency on the number of actions. As each action is a combination of the different facilities (resources) in the congestion games, the results lead to undesirable exponential dependency on the number of facilities.

2.2 Learning in online congestion games

Congestion games were first introduced in the seminal work [39] as a class of games with pure-strategy Nash equilibria. It has then been extensively studied, where its Nash equilibria have been characterized in [42] and a comprehensive introduction has been given in [40].

In the online setting, many works use no-regret learning to develop learning dynamics in this class of games for efficient convergence. [31] are the first to study no-regret learning for congestion games. They showed that the well-known multiplicative weights learning algorithm results in convergence to pure equilibria. Furthermore, they identified a set of mixed Nash equilibria that are weakly stable and showed that the distribution of play converges to this set. Followup works [32] showed that multiplicative weights algorithms converge to the set of Nash equilibria in the sense of Cesàro means, and [36] investigated the effect of learning rate on convergence to Nash equilibria.

With an additional assumption of convex potential functions, [10,11] established a non-asymptotic convergence rate. However, their rate has an exponential dependency on the number of facilities. [18] gave the first non-asymptotic convergence rate under semi-bandit feedback and without an exponential dependency on the number of facilities. However, the convergence is with respect to the averaged-over-time policy and with a $O(T^{-1/6})$ convergence rate. This result is later improved by concurrent work [38], who proposed an online stochastic gradient descent algorithm that converges to an ϵ-approximate Nash equilibrium in $O(\epsilon^{-5})$ time while each individual player enjoys a regret of $O(T^{4/5})$. However, their convergence is only best iterate convergence and only in terms of potential function values (see a detailed comparison in Table 1).

2.3 Combinatorial bandits and shortest path

Combinatorial bandits offer an extension of the classic multi-armed bandit problem where the player must select an action that involves a combination of various resources [9,12,33]. In a special case, the shortest path problem can be viewed as a combinatorial bandit problem where the resources are edges on a graph and the action is a path [28]. Efficient algorithms have been proposed for these problems, and it has been shown that the sublinear regret only linearly depends on the number of resources. However, it is important to note these algorithms

are designed for a single player, and as a result, they may not converge to a Nash equilibrium when applied directly to congestion games by allowing each player to execute the algorithm.

3 Problem Formulation

3.1 Congestion game

A congestion game with n players is defined by $\mathcal{G} = (\mathcal{F}, \{\mathcal{A}_i\}_{i=1}^n, \ \{r^f\}_{f\in\mathcal{F}})$, where i) $\mathcal{F}$ is the facility set that contains F facilities; ii) $\mathcal{A}_i$ is the action space for player i and contains A actions (we assume that the action space for each player is the same), where each action $a_i \in \mathcal{A}_i$ is a combination of k facilities in $\mathcal{F}$; and iii) $r^f : (\mathcal{A}_1 \times \cdots \times \mathcal{A}_n) \to [0,1]$ is the reward function for facility $f \in \mathcal{F}$, which only depends on the number of players choosing this facility, i.e., $\sum_{i=1}^n \mathbb{I}\{f \in a_i\}$.We denote $a = (a_i, a_{-i})$ as a joint action, where a_{-i} is the actions of all other players except player i. The total reward collected by player i with joint action $a = (a_i, a_{-i})$ is $r_i(a_i, a_{-i}) = \sum_{f\in a_i} r^f(a_i, a_{-i})$. Without loss of generality, we assume that $r^f(a) \in [0,1]$.

Deterministically playing actions $a = (a_i, a_{-i})$ is referred to as a pure strategy. The player can also play a mixture of pure strategy, $\omega_i \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ denotes the probability simplex of action space $\mathcal{A}_i$. Similarly, we use $\omega = (\omega_i, \omega_{-i})$ to denote a joint randomized policy.

3.2 Nash equilibrium

One of the commonly used solution concepts in congestion games is Nash equilibrium (NE), and the policies that lead to a Nash equilibrium are referred to as Nash policies. The players are said to be in a Nash equilibrium when no player has an incentive from deviating from its current policy (as described in the definition below).

Definition 1 (Nash equilibrium). *A policy $\omega^* = (\omega_1^*, \ldots, \omega_n^*)$ is called a **Nash equilibirum** if for all $i \in [n]$, $r_i(\omega_i^*, \omega_{-i}^*) \geq r_i(\omega_i, \omega_{-i}^*), \forall \omega_i \in \Delta(\mathcal{A}_i)$. When ω^* is pure, the equilibrium is called a **pure Nash equilibrium**. In addition, when the strategy is pure and the inequality is a strict inequality, the equilibrium is called a **strict Nash equilibrium**.*

Fact 1 ([39]). *There exists a pure Nash equilibrium in any congestion game.*

3.3 Social welfare and price of anarchy

Except for Nash equilibrium, another commonly used metric to measure the efficiency of the dynamics between the players is through social welfare. For a given joint action $a = \{a_i\}_{i=1}^n$, the social welfare is defined to be the sum of players' rewards, i.e., $W(a) = \sum_{i=1}^n r_i(a)$, and the optimal social welfare of the game is defined as $\mathrm{OPT} = \max_{a\in\mathcal{A}_1\times\cdots\times\mathcal{A}_n} W(a)$. This optimality is under the

case where a central coordinator could dictate each player's strategy, and each player's individual incentives are not considered.

Based on the definition of OPT, We can define smooth games as follows.

Definition 2 (Smooth game [41]). *A game is (λ, μ)-smooth if there exists a joint action a^* such that for any joint action a, $\sum_{i \in n} r_i\left(a_i^*, a_{-i}\right) \geq \lambda \mathrm{OPT} - \mu W(a)$.*

The results in [35] show that congestion games are smooth when the reward function are affine, that is, when $r^f(a)$ is an affine function on the scalar variable $\sum_{i=1}^n \mathbb{I}\{f \in a_i\}$. This property enables certain decentralized no-regret learning dynamics to efficiently approximate the optimal welfare [43].

3.4 Online congestion game

In this paper, we study the congestion game in an online setting with a finite time horizon T, where the underlying reward function is unknown. At each time step $t \in [T]$, each player chooses (randomized) policy ω_i^t, from which it forms a joint policy $\omega^t = (\omega_1^t, \ldots, \omega_n^t)$. Then each player i draws a random action $a_i^t \sim \omega_i^t$, plays this action (denote a^t the joint action), and receives overall reward of $\sum_{f \in a_i^t} R^f(a^t)$, where $R^f(a^t)$'s are random variables that satisfy the following assumption.

Assumption 1. *For any facility $f \in \mathcal{F}$, any joint action a^t and any player $i \in [n]$, let $\mathcal{H}_t$ be the history up to time step $t-1$. Then, 1) $R^f(a) \in [0,1]$, and 2) $\mathbb{E}\left[R^f(a^t) \mid \mathcal{H}_t\right] = r^f(a^t)$.*

The assumption implies that the mean of $R^f(a^t)$ is always $r^f(a^t)$. Hence the Nash equilibrium and expected social welfare of the online congestion game is the same as those of the offline congestion game.

We consider two types of feedback rules in this paper: *semi-bandit feedback*, and *full information feedback*. In the semi-bandit feedback, player i observes all the $R^f(a^t)$'s for any $f \in a_i$ (only the facilities he played); and in full information feedback, player i observes all possible information $R^f(a_i, a_{-i})$, for every $a_i \in \mathcal{A}_i$, $\forall f \in a$.

The efficiency of a sequence of policy $\{\omega_i^t\}_{t=1}^T$ can be measured by the individual regret of all the players (which is defined as follows).

Definition 3 (Individual regret). *The individual regret of player i playing policy $\{\omega_i^t\}_{t=1}^T$ is defined as the cumulative difference between the received rewards and the rewards incurred by a best-in-hindsight policy, that is* $\mathrm{Regret}_i(T) = \max_{\omega_i \in \mathcal{A}_i, \{\omega_{-i}^t\}_{t=1}^T} \sum_{t=1}^T r_i(\omega_i, \omega_{-i}^t) - r_i(\omega_i^t, \omega_{-i}^t)$.

4 Algorithm

In this section, we introduce CongestEXP, a decentralized algorithm for online congestion games (Algorithm 1).

Algorithm 1 CongestEXP

Input: learning rate η.
For all $f \in \mathcal{F}$, initialize $\tilde{y}_i^0(f) = 0$, $\omega_i^0(a)$ initialized according to Equation (2).
for $t = 1, \ldots, T$ **do**
    Players play strategy $a^t \sim \omega^t$, $\omega^t = (\omega_1^t, \ldots, \omega_n^t)$.
    Each player i observe $R^f(a^t) \sim r^f(a^t)$, for each $f \in a_i^t$.
    Each player $i \in [N]$ computes $\tilde{y}_i^t(f) = 1 - \frac{\mathbb{I}\{f \in a_i^t\}(1 - R^f(a_i^t, a_{-i}^t))}{q_i^t(f)}$, where $q_i^t(f) = \sum_{a_i \in \mathcal{A}_i, f \in a_i} \omega_i^t(a_i)$.
    Each player $i \in [N]$ updates $\omega_i^{t+1}(a) = \frac{\prod_{f \in a} \tilde{\omega}_i^t(f)}{\sum_{a_i \in \mathcal{A}_i} \prod_{f' \in a_i} \tilde{\omega}_i^t(f')}, \forall a \in \mathcal{A}_i$, where $\tilde{\omega}_i^t(f) = \tilde{\omega}_i^{t-1}(f) \exp\left(\eta \tilde{y}_i^t(f)\right)$.
end for

The algorithm uses the combinatorial nature of the action space. Each player maintains a sampling distribution ω_i^t and a facility-level reward estimator $\tilde{y}_i^t$. At each time step, they first draw a random action $a_i^t \sim \omega_i^t$ and play this action. Then they use their received information to update $\tilde{y}_i^t(f)$'s (for all $f \in \mathcal{F}$) as follows

$$\tilde{y}_i^t(f) = 1 - \frac{\mathbb{I}\{f \in a_i^t\}(1 - R^f(a^t))}{q_i^t(f)}, \quad q_i^t(f) = \sum_{a_i \in \mathcal{A}_i, f \in a_i} \omega_i^t(a_i), \tag{1}$$

where $q_i^t(f)$ is the probability that player i selects facility f at time t based on its current policy ω_i^t. One can easily check that $\tilde{y}_i^t(f)$ is an unbiased estimator for $r^f(a^t)$, and with these facility-level reward estimators, the players then update ω_i^{t+1} as follows (exponential weighting), and then proceed to the next time step.

$$\omega_i^{t+1}(a) = \frac{\prod_{f \in a} \tilde{\omega}_i^t(f)}{\sum_{a_i \in \mathcal{A}_i} \prod_{f' \in a_i} \tilde{\omega}_i^t(f')}, \forall a \in \mathcal{A}_i, \quad \tilde{\omega}_i^t(f) = \exp\left(\eta \sum_{j=1}^{t} \tilde{y}_i^j(f)\right). \tag{2}$$

On the one hand, in the semi-bandit setting, our algorithm leverages this kind of feedback and estimates rewards at the facility level. We note that this algorithm has also been previously utilized to tackle online shortest path problems and combinatorial bandit problems, as documented in the literature [28,9,12,16]. This enables us to achieve a low individual regret (Theorem 1) and guarantee a lower bound for the overall social welfare (Corollary 1). On the other hand, our algorithm constructs exponential weights based on the reward estimation at the action level. This makes sure that the joint policy ω^t can converge to a Nash equilibrium fast when it is nearby (Theorem 2 and 3).

In summary, our results indicate that adopting Algorithm 1 in a congestion game leads to favorable outcomes. Each player enjoys favorable cumulative individual rewards, without compromising the overall social welfare. Moreover, when the joint policy is close to the Nash equilibrium, players can quickly converge to a stable equilibrium state, avoiding inefficient and chaotic dynamics.

5 Theoretical Results

In this section, we present our main theoretical results.

5.1 Sublinear individual regret with linear dependency on F

Our first theorem shows that each individual player enjoys a sublinear individual regret.

Theorem 1. *Under semi-bandit feedback, Algorithm 1 with* $\eta = \frac{1}{\sqrt{T}}$ *satisfies that for all* $i \in [n]$, $\mathbb{E}\left[Regret_i(T)\right] = O\left(kF\sqrt{T}\right)$.

Remark 1. *By Markov's inequality, the above stated result also holds with high probability.*

Compared with naively applying exponential weights on the congestion game (with a regret of $\tilde{O}\left(\sqrt{A_i T}\right)$ [4]), we can see that Theorem 1 reduces the factor $\sqrt{A_i}$ to kF. This is a significant improvement since $A \approx F^k$ is exponentially larger than kF.

Though there exist some works to achieve a similar regret upper bound [22], we emphasize that these algorithms only work in the full-information setting, but not the semi-bandit setting. Besides, our algorithm can converge to a strict Nash equilibrium fast, while existing ones can only guarantee to converge to a coarse correlated equilibrium (please see details in Section 5.2). We include the proof in the arXiv version of the paper, and we summarize the technical highlights here.

Technical highlight of Theorem 1 In classical proofs of exponential weights algorithms, the regret is closely linked to the quadratic term of the reward estimator, i.e., $\mathbb{E}_t[\sum_{a_i \in \mathcal{A}_i} \omega_i^t(a_i) \left(\tilde{y}_i^t(a_i)\right)^2]$[4,33], where $\tilde{y}_i^t(a_i)$ is the estimated reward of action a_i at time step t, and $\mathbb{E}_t[\cdot]$ denotes the conditional expectation over all history up to time t. If we can upper bound this term by a constant polynomial with k and F, then we can remove the exponential factor in the individual regret upper bound.

Specifically, one would get a regret decomposition as follows, $\mathbb{E}\left[\text{Regret}_i(T)\right] \leq \mathbb{E}\left[\frac{kF}{\eta} + \eta \sum_{t=1}^{T} \sum_{a_i \in \mathcal{A}_i} \omega_i^t(a_i) \left(\sum_{f \in a_i} \tilde{y}_i^t(f)\right)^2\right]$.

With our facility level estimator (Eq. (1)), $\tilde{y}_i^t(a_i) = \sum_{f \in a_i} \tilde{y}_i^t(f)$, the second term could be upper bounded as $\sum_{a_i \in \mathcal{A}_i} \omega_i^t(a_i) \left(\sum_{f \in a_i} \tilde{y}_i^t(f)\right)^2 \leq k + k \sum_{f \in \mathcal{F}} \left(\frac{\mathbb{I}\{f \in a_i^t\}(1 - R^f(a_i^t, a_{-i}^t))}{q_i^t(f)}\right)^2 q_i^t(f)$. Let $\mathbb{E}_t[\cdot]$ denote conditional expectation over all history up to time t. Notice that our estimator $\tilde{y}_i^t(f)$ is unbiased and

$(1 - R_i^f(a^t))^2$ is upper bounded by 1. Then, taking expectations on both sides yield

$$\mathbb{E}\left[\sum_{a_i\in\mathcal{A}_i}\omega_i^t(a_i)\left(\sum_{f\in a_i}\tilde{y}_i^t(f)\right)^2\right] \le k + k\mathbb{E}\left[\sum_{f\in\mathcal{F}}\left(\frac{\mathbb{I}\{f\in a_i^t\}\left(1-R_i^f\left(a_i^t,a_{-i}^t\right)\right)}{q_i^t(f)}\right)^2 q_i^t(f)\right]$$
$$\le k + k\mathbb{E}\left[\mathbb{E}_{t-1}\left[\sum_{f\in\mathcal{F}}\frac{\mathbb{I}\{f\in a_i^t\}}{q_i^t(f)}\right]\right] \le k + kF\,.$$

From the above explanation, one can see the necessity of estimating the rewards at the facility level. If the reward estimator is constructed at the action level, that is, an estimator of the form $\tilde{y}_i^t(a_i) = k - \frac{\mathbb{I}\{a_i=a_i^t\}\left(k-\sum_{f\in a_i^t}R_i^f\left(a_i^t,a_{-i}^t\right)\right)}{\omega_i^t(a_i)}$. Consider the case that $R_i^f\left(a_i^t,a_{-i}^t\right)$ is always 0 and at the beginning $\omega_i^t(a_i) = 1/|\mathcal{A}_i|$, then this quadratic term $\mathbb{E}_t[\sum_{a_i\in\mathcal{A}_i}\omega_i^t(a_i)\left(\tilde{y}_i^t(a)\right)^2]$ is approximately, $\mathbb{E}_t\left[\sum_{a_i\in\mathcal{A}_i}\omega_i^t(a_i)\left(\tilde{y}_i^t(a)\right)^2\right] \approx \sum_{a_i\in\mathcal{A}_i}\omega_i^t(a_i)\cdot\left(\omega_i^t(a_i)\left(\frac{k}{\omega_i^t(a_i)}\right)^2\right) = k^2|\mathcal{A}_i|$, which scales with the number of actions and is thus always exponentially large.

5.2 Tight approximation to optimal welfare

One immediate consequence of Theorem 1 is that our proposed algorithm can achieve a tight approximation to the optimal social welfare.

Corollary 1. *Under semi-bandit feedback, if the congestion game is (λ,μ)-smooth, then Algorithm 1 with $\eta = \frac{1}{\sqrt{T}}$ satisfies $\frac{1}{T}\sum_{t=1}^T W(\omega^t) \ge \frac{\lambda}{1+\mu}\mathrm{OPT} - O\left(\frac{nkF}{\sqrt{T}(1+\mu)}\right)$.*

We remark that $\frac{\lambda}{1+\mu}\mathrm{OPT}$ is shown to be a tight approximation of optimal social welfare possible by offline algorithms that attain Nash equilibrium in congestion games [41]. Therefore, the above result shows that our algorithm is as efficient as any offline Nash policy asymptotically.

5.3 Fast convergence to strict Nash equilibrium

Beyond the low individual regret, we also show that our algorithm can produce a set of policies $\{\omega_i^t\}_{t=1}^T$ that converges fast to a strict Nash equilibrium ω_i^* in the full-information setting.

We first consider a simple case, where each player observes the expected rewards directly (which also take expectation on the randomness of $a_{-i}^t \sim \omega_{-i}^t$, i.e. $\mathbb{E}_{a_{-i}^t\sim\omega_{-i}^t}[r^f(a_i,a_{-i}^t)]$ for any a_i). In addition, we maintain the following assumption on the action.

Assumption 2. *We assume that any k facilities form a valid action.*

We note that the above assumption is only used to simplify our analysis. When the game does not have such a property, we can include dummy actions to meet this requirement, and the analysis in this section almost remains the same.

Theorem 2. *Under assumption 2, consider the case where each player receives* $\mathbb{E}_{a^t_{-i}\sim\omega^t_{-i}}[r^f(a^t_i, a^t_{-i})], \forall a_i \in \mathcal{A}_i, \forall f \in a_i$ *in a game that permits a strict Nash equilibrium* $\omega^* = (\omega^*_1, \cdots, \omega^*_n)$*, and let* $\tilde{y}^t_i(f) = \mathbb{E}_{a^t_{-i}\sim\omega^t_{-i}}[r^f(a^t_i, a^t_{-i})]$ *in Line 6 of Algorithm 1. Suppose* $\tilde{y}^0_i(f), \forall i \in [n]$ *is initialized such that* $\omega^0 \in U_M$*, where* U_M *is a neighborhood of the strict Nash equilibrium (as defined in Eq. (3)), then for any* $i \in [n]$ *and any* t*, and under Assumption 2, we have* $\|\omega^t_i - \omega^*_i\|_1 \leq 2(kF\exp(-M - \eta\epsilon t))$*, where* $M \geq \left|\log\left(\frac{\epsilon}{2kF}\right)\right|$*, and* ϵ *is a constant that is game-dependent only.*

Remark 2. *We note that the convergence rate of the algorithm can be improved by increasing the step size* η*. This is because when each player receives expected rewards, the player can take greedy steps toward the equilibrium strategy. This agrees with greedy strategies that are previously employed to reach strict Nash equilibrium [15]. However, such greedy policies would not work in the presence of reward uncertainty, as we will discuss in Theorem 3.*

It is worth mentioning that the convergence rate of our algorithm does not rely on the number of actions A_i, but rather solely on the number of facilities F. This is an improvement over the previous findings for exponential weights algorithms with non-combinatorial action spaces in the context of a congestion game, where the rate is linearly dependent on the number of actions [15]. The reason for this is the utilization of our facility-level reward estimation technique once again.

Previous studies on the convergence rate of congestion games [10,11] have established a linear convergence rate when the game possesses a convex potential function and the algorithm is given an appropriate initial starting point. The potential function provides a means to capture the incentives of all players to modify their actions and can be used to characterize the dynamics in policy updates. Assuming the convexity of the potential function implies optimization on a simpler policy optimization landscape. In contrast, our algorithm achieves a much faster rate of convergence, and this convergence rate still holds even in the absence of a convex potential function. This is because that we adopt a different approach where we directly argue through the algorithm update rule that the updated policy will always fall within a neighborhood around the Nash equilibrium. This bypasses the need for a smoothness potential function and demonstrates the effectiveness of our approach.

In addition, we remark that though some variants of Mirror Descent (MD) or Follow-the-Regularized-Leader (FTRL) algorithms are also proven to enjoy sublinear regret with logarithmic dependency on action space in the full information setting [22], these results only imply convergence to an approximate coarse correlated equilibrium and do not enjoy convergence to Nash equilibrium. In

comparison, Nash equilibrium is much more stable, as the dynamic will remain there unless external factors change, while coarse correlated equilibrium may be more sensitive to small changes in the correlation method, which can lead to deviation from the equilibrium [35].

[18] and [38] have also investigated the convergence to Nash equilibrium for congestion game, under the notion of best-iterate convergence of the rewards value. Their algorithm ensures that with high probability, there exists a $t \in [T]$ such that $r_i(\omega_i^t)$ is close enough to the rewards attained by Nash equilibrium. [38] also showed their algorithm can attain sublinear individual regret simultaneously. However, these results do not directly guarantee the convergence of the actual action sequence $\{\omega_i^t\}_{t=1}^T$ (as what we do in this paper).

We include the proof in the arXiv version of the paper, and we summarize the technical highlights here. To prove Theorem 2, we first identify that there exists a neighborhood around the strict Nash equilibrium, such that for any player i, his action in the strict Nash equilibrium is the only optimal choice.

Lemma 1. *If there exists a strict Nash equilibrium $a^* = (a_1^*, \ldots, a_n^*)$, then there exists $\epsilon > 0$ and a neighborhood U_ϵ of a^*, such that for all $\tilde{\omega} = (\tilde{\omega}_i, \tilde{\omega}_{-i}) \in U_\epsilon$, $r_i(a_i^*, \tilde{\omega}_{-i}) - r_i(a_i, \tilde{\omega}_{-i}) \geq \epsilon\,, \quad \forall i \in [n]\,, a_i \in \mathcal{A}_i\,, a_i \neq a_i^*$, where $r_i(a_i, \omega_{-i})$ is defined as $\mathbb{E}_{a_{-i} \sim \omega_{-i}}[r_i(a_i, a_{-i})]$.*

Moreover, if the difference in reward estimator $\tilde{z}_i^t(a_i) = \sum_{j=0}^t \Big(\sum_{f \in a_i} \tilde{y}_i^j(f) - \sum_{f' \in a_i^*} \tilde{y}_i^j(f') \Big)$ is upper bounded by some small enough constant, then the induced policy of Algorithm 1 falls into the neighborhood set U_ϵ.

Lemma 2. *Let $\tilde{z}_i^t(a_i) = \sum_{j=0}^t \Big(\sum_{f \in a_i} \tilde{y}_i^j(f) - \sum_{f' \in a_i^*} \tilde{y}_i^j(f') \Big)$, and define*

$$U_M = \left\{ \omega^t \text{ computed by Algorithm 1} \mid \tilde{z}_i^t(a_i) \leq -M\,, \forall a_i \neq a_i^*, \forall i \in [n] \right\}. \quad (3)$$

For sufficiently large M, $U_M \subseteq U_\epsilon$. Moreover, following the updates of Algorithm 1, and under Assumption 2,, if $\omega^t \in U_M$, then $\omega^{t+1} \in U_M$.

Thus, if ω^0 is in the neighborhood $U_M \subseteq U_\epsilon$, then the reward estimator $\tilde{z}_i^t(a_i)$ can only decrease (by Lemma 1), and hence the algorithm will give an updated policy ω^t that is also within the neighborhood set U_ϵ.

Also, note that ω_i^* is a strict Nash equilibrium, which implies that $|\omega_i^t - \omega_i^*|_1 = 2(1 - \omega_i^t(a_i^*))$. Hence, to establish the convergence rate, we need to lower bound $\omega_i^t(a_i^*) = \frac{\prod_{f \in a_i^*} \tilde{\omega}_i^t(f)}{\sum_{a' \in \mathcal{A}} \prod_{f' \in a'} \tilde{\omega}_i^t(f')} = \frac{\prod_{f \in a_i^*} \exp\left(\eta \sum_{j=0}^t \tilde{y}_i^j(f)\right)}{\sum_{a' \in \mathcal{A}} \prod_{f' \in a'} \exp\left(\eta \sum_{j=0}^t \tilde{y}_i^j(f')\right)}$.

Technical challenge We remark that if we directly apply Lemma 2, we can get

$$\begin{aligned}\omega_i^t(a_i^*) = \frac{\prod_{f \in a_i^*} \tilde{\omega}_i^t(f)}{\sum_{a' \in \mathcal{A}} \prod_{f' \in a'} \tilde{\omega}_i^t(f')} &\geq \frac{1}{1 + \sum_{a_i \in \mathcal{A}_i, a_i \neq a_i^*} \Big(\prod_{f \in a_i} \tilde{\omega}_i^t(f) - \prod_{f' \in a_i^*} \tilde{\omega}_i^t(f') \Big)} \\ &\geq 1 - \sum_{a_i \in \mathcal{A}_i, a_i \neq a_i^*} \Big(\prod_{f \in a_i} \tilde{\omega}_i^t(f) - \prod_{f' \in a_i^*} \tilde{\omega}_i^t(f') \Big).\end{aligned}$$

Suppose one can upper bound $\sum_{a_i \in \mathcal{A}_i, a_i \neq a_i^*} \left(\prod_{f \in a_i} \tilde{\omega}_i^t(f) - \prod_{f' \in a_i^*} \tilde{\omega}_i^t(f') \right) \leq \exp(-t)$, then this gives $1 - \sum_{a_i \in \mathcal{A}_i, a_i \neq a_i^*} \exp(-t)$, which yields a convergence rate of $(|\mathcal{A}_i| - 1)\exp(-t)$ as $\|\omega_i^t - \omega_i^*\|_1 = 2(1 - \omega_i^t(a_i^*))$. However, this approach implies that the convergence rate scales linearly with the number of actions (thus exponentially with the number of facilities), and is what we wanted to avoid in the analysis.

To overcome this exponential dependency, we utilize the fact that any k-facility combination is an action, which means that we can order the facility from $f_1, \ldots, f_F$ in decreasing order of $\tilde{\omega}_i^t(f)$ and $f_1, \ldots, f_k$ form the optimal pure Nash action a_i^*.

Then we consider the case that each player observes only $R^f(a_i^t, a_{-i}^t)$, instead of the expected rewards.

Theorem 3. *Under Assumption 2, consider the case where each player receives a stochastic reward under the full information setting and under Assumption 2. Assume the game permits a strict Nash equilibrium $\omega^* = (\omega_1^*, \cdots, \omega_n^*)$. Let $\tilde{y}_i^t(f) = R^f(a_i^t, a_{-i}^t)$ in Line 6 of Algorithm 1, and set the learning rate to be time-dependent such that $\sum_{t=0}^{\infty} \eta_t^2 \leq \frac{\delta \cdot M^2}{8kn(F-1)} \leq \sum_{t=0}^{\infty} \eta_t = \infty$. Suppose $\tilde{y}_i^0(f), \forall i \in [n]$ is initalized such that $\omega^0 \in U_{2M} \subseteq U_\epsilon$, then for any $i \in [n]$ and any t, we have $\|\omega_i^t - \omega_i^*\|_1 \leq 2kF \exp\left(-M - \epsilon \sum_{j=0}^{t} \eta_j\right)$, with probability at least $1 - \delta$, where $M \geq \left|\log\left(\frac{\epsilon}{2kF}\right)\right|$, and ϵ is a constant that is game-dependent only.*

We remark that in the case of stochastic rewards, the convergence rate of our algorithm cannot be arbitrarily large, as the learning rate η cannot be taken to be arbitrarily large. If we take $\eta_t = \beta t^{-\alpha}$, with β being a small positive constant and $\alpha \in (1/2, 1)$. Then our convergence rate is $\|\omega_i^t - \omega_i^*\|_1 \leq O\left(\exp\left(-\frac{\beta}{1-\alpha} t^{1-\alpha}\right)\right)$, which is close to exponentially fast convergence. When the reward function is smooth, we remark that this can imply each player only experiences constant regret. We include the proof to Theorem 3 in the arXiv version of the paper.

6 Conclusion

We studied the congestion game under semi-bandit feedback and presented a modified version of the well-known exponential weights algorithm. The algorithm ensures sublinear regret for every player, with the regret being linearly dependent on the number of facilities. Additionally, the proposed algorithm can learn a policy that rapidly converges to the pure Nash policy, with the convergence rate also being linearly dependent on the number of facilities. To our best knowledge, these are the first results on congestion games for sublinear individual regret and geometric Nash convergence rate, without an exponential dependency on the number of facilities.

There are several possible directions to further study the online congestion game. First, as our work only considers the semi-bandit feedback model for individual regret, the regret and convergence rate under the full-bandit feedback model remains unclear. For the Nash convergence result, our algorithm only enjoys theoretical guarantees in the full-information setting. It remains future work to extend this result to semi-bandit and full-bandit feedback models. Moreover, it also remains in question whether the results of this work can be extended to the online Markov congestion game proposed by [18].

Acknowledgement

Baoxiang Wang and Jing Dong are partially supported by the National Natural Science Foundation of China (62106213, 72394361) and an extended support project from the Shenzhen Science and Technology Program. Volkan Cevher was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data), and the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048. Leello Dadi acknowledges support from Google. Luca Viano is funded (in part) through a PhD fellowship of the Swiss Data Science Center, a joint venture between EPFL and ETH Zurich.

References

1. Abernethy, J.D., Hazan, E., Rakhlin, A.: Competing in the dark: An efficient algorithm for bandit linear optimization (2009)
2. Audibert, J.Y., Bubeck, S., Lugosi, G.: Regret in online combinatorial optimization. Math. Oper. Res. **39**(1), 31–45 (02 2014)
3. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM J. Comput. **32**(1), 48–77 (2002)
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM journal on computing **32**(1), 48–77 (2002)
5. Awerbuch, B., Kleinberg, R.D.: Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing. pp. 45–53 (2004)
6. Brown, G.W.: Some notes on computation of games solutions. Tech. rep., RAND CORP SANTA MONICA CA (1949)
7. Bubeck, S., Cesa-Bianchi, N., Kakade, S.M.: Towards minimax policies for online linear optimization with bandit feedback. In: Mannor, S., Srebro, N., Williamson, R.C. (eds.) COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland. JMLR Proceedings, vol. 23, pp. 41.1–41.14. JMLR.org (2012)
8. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge university press (2006)
9. Cesa-Bianchi, N., Lugosi, G.: Combinatorial bandits. Journal of Computer and System Sciences **78**(5), 1404–1422 (2012)

10. Chen, P.A., Lu, C.J.: Playing congestion games with bandit feedbacks. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (2015)
11. Chen, P.A., Lu, C.J.: Generalized mirror descents in congestion games. Artificial Intelligence **241**, 217–243 (2016)
12. Chen, W., Wang, Y., Yuan, Y.: Combinatorial multi-armed bandit: General framework and applications. In: International conference on machine learning (2013)
13. Chen, X., Peng, B.: Hedging in games: Faster convergence of external and swap regrets. Advances in Neural Information Processing Systems (2020)
14. Cheung, M.H., Southwell, R., Huang, J.: Congestion-aware network selection and data offloading. In: 2014 48th Annual Conference on Information Sciences and Systems (CISS). pp. 1–6. IEEE (2014)
15. Cohen, J., Héliou, A., Mertikopoulos, P.: Exponentially fast convergence to (strict) equilibrium via hedging. arXiv preprint arXiv:1607.08863 (2016)
16. Combes, R., Talebi Mazraeh Shahi, M.S., Proutiere, A., et al.: Combinatorial bandits revisited. Advances in neural information processing systems (2015)
17. Cui, Q., Xiong, Z., Fazel, M., Du, S.S.: Learning in congestion games with bandit feedback (2022)
18. Cui, Q., Xiong, Z., Fazel, M., Du, S.S.: Learning in congestion games with bandit feedback. In: Advances in Neural Information Processing Systems (2022)
19. Dadi, L., Panageas, I., Skoulakis, S., Viano, L., Cevher, V.: Polynomial convergence of bandit no-regret dynamics in congestion games. arXiv preprint arXiv:2401.09628 (2024)
20. Dani, V., Hayes, T.P., Kakade, S.M.: The price of bandit information for online optimization. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. p. 345–352. NIPS'07, Curran Associates Inc., Red Hook, NY, USA (2007)
21. Daskalakis, C., Deckelbaum, A., Kim, A.: Near-optimal no-regret algorithms for zero-sum games. In: Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms (2011)
22. Daskalakis, C., Fishelson, M., Golowich, N.: Near-optimal no-regret learning in general games. Advances in Neural Information Processing Systems (2021)
23. Daskalakis, C., Pan, Q.: A counter-example to karlin's strong conjecture for fictitious play. In: IEEE 55th Annual Symposium on Foundations of Computer Science (2014)
24. Dong, J., Wu, J., Wang, S., Wang, B., Chen, W.: Taming the exponential action set: Sublinear regret and fast convergence to nash equilibrium in online congestion games. arXiv preprint arXiv:2306.13673 (2023)
25. Flaxman, A.D., Kalai, A.T., McMahan, H.B.: Online convex optimization in the bandit setting: gradient descent without a gradient. arXiv preprint cs/0408007 (2004)
26. Giannou, A., Vlatakis-Gkaragkounis, E.V., Mertikopoulos, P.: On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. Advances in Neural Information Processing Systems (2021)
27. György, A., Linder, T., Lugosi, G., Ottucsák, G.: The on-line shortest path problem under partial monitoring. J. Mach. Learn. Res. **8**, 2369–2403 (2007)
28. György, A., Linder, T., Lugosi, G., Ottucsák, G.: The on-line shortest path problem under partial monitoring. Journal of Machine Learning Research **8**(10) (2007)
29. Hsieh, Y.G., Antonakopoulos, K., Mertikopoulos, P.: Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In: Conference on Learning Theory (2021)

30. Kalai, A., Vempala, S.: Efficient algorithms for online decision problems. Journal of Computer and System Sciences **71**(3), 291–307 (2005), learning Theory 2003
31. Kleinberg, R., Piliouras, G., Tardos, É.: Multiplicative updates outperform generic no-regret learning in congestion games. In: Proceedings of the forty-first annual ACM symposium on Theory of computing (2009)
32. Krichene, W., Drighès, B., Bayen, A.M.: Online learning of nash equilibria in congestion games. SIAM Journal on Control and Optimization **53**(2), 1056–1081 (2015)
33. Lattimore, T., Szepesvári, C.: Bandit algorithms. Cambridge University Press (2020)
34. Neu, G., Bartók, G.: An efficient algorithm for learning with semi-bandit feedback. In: Jain, S., Munos, R., Stephan, F., Zeugmann, T. (eds.) Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings. Lecture Notes in Computer Science, vol. 8139, pp. 234–248. Springer (2013)
35. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: Algorithmic game theory. Cambridge university press (2007)
36. Palaiopanos, G., Panageas, I., Piliouras, G.: Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. Advances in Neural Information Processing Systems **30** (2017)
37. Panageas, I., Skoulakis, S., Viano, L., Wang, X., Cevher, V.: Semi bandit dynamics in congestion games: Convergence to nash equilibrium and no-regret guarantees. (2023)
38. Panageas, I., Skoulakis, S., Viano, L., Wang, X., Cevher, V.: Semi bandit dynamicsin congestion games: Convergence to nash equilibrium and no-regret guarantees. International Conference on Machine Learning (2023)
39. Rosenthal, R.W.: A class of games possessing pure-strategy nash equilibria. International Journal of Game Theory **2**, 65–67 (1973)
40. Roughgarden, T.: Routing games. Algorithmic game theory **18**, 459–484 (2007)
41. Roughgarden, T.: Intrinsic robustness of the price of anarchy. In: Proceedings of the forty-first annual ACM symposium on Theory of computing. pp. 513–522 (2009)
42. Roughgarden, T., Tardos, É.: How bad is selfish routing? Journal of the ACM (JACM) **49**(2), 236–259 (2002)
43. Syrgkanis, V., Agarwal, A., Luo, H., Schapire, R.E.: Fast convergence of regularized learning in games. Advances in Neural Information Processing Systems **28** (2015)
44. Tekin, C., Liu, M., Southwell, R., Huang, J., Ahmad, S.H.A.: Atomic congestion games on graphs and their applications in networking. IEEE/ACM Transactions on Networking **20**(5), 1541–1552 (2012)
45. Zhang, F., Wang, M.M.: Stochastic congestion game for load balancing in mobile-edge computing. IEEE Internet of Things Journal **8**(2), 778–790 (2020)
46. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: Fawcett, T., Mishra, N. (eds.) Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA. pp. 928–936. AAAI Press (2003)