Semi Bandit Dynamics in Congestion Games: Convergence to Nash Equilibrium and No-Regret Guarantees.

Ioannis Panageas *1 Stratis Skoulakis *2 Luca Viano *2 Xiao Wang 3 Volkan Cevher 2

Abstract

In this work, we introduce a new variant of online gradient descent, which provably converges to Nash Equilibria and simultaneously attains sublinear regret for the class of congestion games in the semi-bandit feedback setting. Our proposed method admits convergence rates depending only polynomially on the number of players and the number of facilities, but not on the size of the action set, which can be exponentially large in terms of the number of facilities. Moreover, the running time of our method has polynomial-time dependence on the implicit description of the game. As a result, our work answers an open question from (Cui et al., 2022).

1. Introduction

Congestion games is a class of multi-agent games at which $n$ selfish agents compete over a set of $m$ resources. Each agent selects a subset of the resources and her cost depends on the load of each of the selected resources (number of other agents using the resource). For example in Network Congestion Games, given a graph, each agent $i$ wants to travel from a starting vertex $s_i$ to a target position $t_i$ and thus needs to select set of edges forming an $(s_i, t_i)$-path. Due to their numerous applications congestion games have been extensively studied over the years (Koutsoupias & Papadimitriou, 1999; Roughgarden & Tardos, 2002; Christodoulou & Koutsoupias, 2005; Fotakis et al., 2005; de Keijzer et al., 2010; Roughgarden, 2009).

It is well-known that congestion games always admit a Nash Equilibrium (NE) which is a steady state at which no agent can unilaterally deviate without increasing her cost. At the same time, a long line of research studies the convergence properties to NE of game dynamics (e.g. no-regret, best response, fictitious play) at which the agents of a congestion game iteratively update their strategies based on the strategies of the other agents in their attempt to minimize their individual cost.

In most real-world scenarios, agents do not have access to the strategies of the other agents (full-information feedback) and are only informed on the loads/cost of the resources they selected at each round. For example a driver learns only the congestion on the highways that she selected and not the congestion of the alternatives that she did not select. This type of feedback is called semi-bandit feedback and has been extensively studied in the context of online learning.

Motivated by the above, (Cui et al., 2022) in their recent work investigate the following question:

**Question.** (Cui et al., 2022) Are there update rules under (semi)-bandit feedback, that once adopted by all agents of a congestion game, the overall system converges to a Nash Equilibrium with rate that is independent of the number of possible strategies?

We note that in congestion games the number of possible strategies can be exponentially large with respect to the game description. For example the number of $(s, t)$-paths can be of the order $O(2^{\Theta(m)})$ with respect to the number of edges $m$. (Cui et al., 2022) provide an update rule (based on the Frank-Wolfe method) that once adopted by all $n$ agents, the overall system requires $O(n^{12}m^9/\epsilon^6)$ time-steps (samples) to reach an $\epsilon$-approximate NE.

Despite its fast convergence properties, the update rule of (Cui et al., 2022) is not aligned with the selfish nature of the agents participating in a congestion game. This is because their method does not provide any kind of guarantees on the regret of the agents adopting it. As a result, (Cui et al., 2022) posed the following open question.

**Open Question.** (Cui et al., 2022) Are there update rules under (semi)-bandit feedback that

1. provide (adversarial) no-regret guarantees to any
agent that adopts them and

2. once adopted by all agents, the overall system converges to Nash Equilibrium with rate independent on the number of strategies?

The term no-regret refers to the fact that the time-average cost of any agent, adopting the update rule, is upper bounded by the time-average cost of the best fixed path in hindsight (no matter how the other agents select their strategies). Due to their remarkable guarantees, no-regret algorithms have been the standard choice for modeling selfish behavior in non-cooperative environments [Audibert et al., 2014].

Our Contribution. In this work, we provide a positive answer to the above open question, while we improve upon the results of [Cui et al., 2022] in two important aspects. Specifically, we propose a semi-bandit feedback online learning algorithm, called Semi-Bandit Gradient Descent with Carathéodory Exploration (SBGD – CE), with the following properties:

- **No-regret guarantees**: Any agent adopting SBGD-CE admits at most $O(n^2T^{3/5})$ regret no matter how the other agents select their strategies.

- **Convergence to NE**: If SBGD – CE is adopted by all agents, the overall system reaches an $\epsilon$-approximate NE within $O(n^{6.5}m^7/\epsilon^5)$ time steps improving the $O(n^{12}m^9/\epsilon^6)$ bound of [Cui et al., 2022].

- **Polynomial Update Rule**: The update rule of SBGD-CE runs in polynomial-time with respect to the implicit description of the strategy space (see Section 2.2). On the other hand, the update rule of [Cui et al., 2022] requires linear time and space complexity w.r.t the number of strategies. Thus, in many interesting settings such as network congestion games, the time and space complexity of Frank-Wolfe with Exploration II [Cui et al., 2022] is exponential w.r.t the number of edges.

**Remark 1.1** (Notion of convergence). As in [Cui et al., 2022; Leonardos et al., 2022; Ding et al., 2022; Anagnostides et al., 2022], the notion of convergence that we use is the so-called best-iterate convergence. Mathematically speaking, the time-averaged exploitability (defined as the sum of best deviation minus chosen strategy per agent) is bounded by $\epsilon$ (see Theorem 3.8). From a game-theoretic point of view, bset-iterate convergence implies that with high probability almost all iterates are $O(\epsilon)$-Mixed NE. From a learning point of view, best-iterate convergence implies that we can learn an approximate NE of an unknown congestion game by considering the strategy profile at a iterate $t \sim Unif(1, \ldots, T)$ (see Corollary 1). The term "best-iterate convergence" might not be the most descriptive for the above, however it is the one most commonly used in the literature.

Our Techniques and Related Work The fundamental difficulty in designing no-regret online learning algorithms under (semi)-bandit feedback is to guarantee that each strategy is sufficiently explored. Unfortunately, standard bandit algorithms such as EXP3 [Auer et al., 2002] result in exponentially large in $m$ regret bounds, e.g. $O(2^{\Omega(m)} \sqrt{T})$, as well as time and space complexity. A long line of research in the context of combinatorial bandits provides no-regret algorithms with polynomial dependence w.r.t to $m$ on the regret, while many of those algorithms can be efficiently implemented [Awerbuch & Kleinberg, 2004; Dani et al., 2007; Gyorgy et al., 2007; Bubeck et al., 2012; Cesa-Bianchi & Lugosi, 2012; Kalai & Vempala, 2005; Neu & Bartók, 2013; Audibert et al., 2014]. For example, (Bubeck et al., 2012) provide an online learning algorithm with regret $O(m\sqrt{T})$. However, in order to overcome the exploration problem, these algorithms use involved techniques (e.g. barycentric spanners or entropic projections [Awerbuch & Kleinberg, 2004; Bubeck et al., 2012]) which introduce major technical difficulties in their multi-agent analysis. To the best of our knowledge, none of these algorithms guarantees convergence to NE in congestion games once adopted by all agents.

**Remark 1.2**. We highly remark that the no-regret property does not imply convergence to Nash Equilibrium in congestion/potential games [Cohen et al., 2017; Babichenko & Rubinstein, 2020]. No-regret dynamics are guaranteed to converge in Coarse Correlated Equilibrium that is a strict superset of NE and that can even contain strictly dominated strategies [Viossat & Zapechelnyuk, 2013].

On the somehow opposite front, recent works studying the convergence properties of semi-bandit game dynamics in potential games use explicit exploration schemes at which each strategy is selected with a small probability [Leonardos et al., 2022; Ding et al., 2022]. However such exploration schemes lead to convergence rates that scale polynomially on the number of strategies (can be exponential w.r.t to $m$ in congestion games). (Cui et al., 2022) combine an explicit exploration scheme with the Frank-Wolfe method and establish that the resulting convergence rate (number of samples) to NE depends only polynomially in $m$. As mentioned above, the update rule of (Cui et al., 2022) does not guarantee the no-regret property in the adversarial case while its update rule is of exponential time and space. Table 1 concisely presents the above mentioned results.

In order to solve the exploration problem with schemes that are simple enough to analyze in the multi-agent case, we introduce the notion of Bounded-Away Description Polytope. These polytopes are subsets of description polytopes, the extreme points of which correspond to the available strategies and additionally impose lower bounds on the fractional selection of each resource. Our SBGD – CE method is based
Table 1. Comparison with related works. *See Remark 3.7 for regret bound $O(n^{3/2}T^{3/4})$ under different step size and exploration parameter choices. †By convergence to NE we mean best-iterate convergence as explained in Remark 1.1. We note that all the known results that provide rates use the same notion of convergence (as presented in the table).

<table>
<thead>
<tr>
<th>Method</th>
<th>Adversarial Regret</th>
<th>Convergence† to NE</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPPG</td>
<td>Not Established</td>
<td>$O\left(n^2\Theta(m)/\epsilon^5\right)$</td>
<td>Exp. in Available Resources</td>
</tr>
<tr>
<td>IPGA</td>
<td>Not Established</td>
<td>$O\left(n^3\Theta(m)/\epsilon^5\right)$</td>
<td>Exp. in Available Resources</td>
</tr>
<tr>
<td>FW with Exploration II</td>
<td>Not Established</td>
<td>$O\left(n^{12}m^9/\epsilon^6\right)$</td>
<td>Exp. in Available Resources</td>
</tr>
<tr>
<td>SBGD – CE (this work)</td>
<td>$O\left(n^2T^{4/5}\right)^*$</td>
<td>$O\left(n^{6.5}m^7/\epsilon^5\right)$</td>
<td>Poly. in Implicit Description</td>
</tr>
</tbody>
</table>


2. Preliminaries and Our Results

2.1. Congestion Games

A *congestion game* is composed by $n$ selfish agents and a set of resources $E$ with $|E| = m$. The strategy of each agent $i \in [n]$ is a subset of resources $p_i \in \mathcal{P}_i$ where $\mathcal{P}_i \subseteq 2^E$ is the strategy space of agent $i$. The set $\mathcal{P}$ denotes all joint strategy profiles, $\mathcal{P} := \mathcal{P}_1 \times \ldots \times \mathcal{P}_n$. Given a strategy profile $p \in \mathcal{P}$, we use the notation $p := (p_1, \ldots, p_n)$ where $p_i$ captures the strategy of agent $i$ and $p_{-i}$ denotes the strategies of all agents but $i$.

The *load* of a resource $e \in E$ under the strategy profile $p := (p_1, \ldots, p_n) \in \mathcal{P}$ is denoted by $\ell_e(p)$ and equals the number of agents using resource $e \in E$, i.e.,

$$\ell_e(p_1, \ldots, p_n) \equiv \sum_{i=1}^{n} \mathbb{1}[e \in p_i]. \quad (1)$$

Each resource $e$ admits a positive and non-decreasing cost function $c_e : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ where $c_e(\ell)$ denotes the congestion cost of $e \in E$ under load $\ell \in \mathbb{N}$. Additionally, we set $c_{\max} := \max_{e \in E} c_e(n)$.

Given a strategy profile $p := (p_1, \ldots, p_n) \in \mathcal{P}$, the cost of agent $i$ is defined as

$$C_i(p) \equiv \sum_{e \in P_i} c_e(\ell_e(p)).$$

**Definition 1 (Nash Equilibrium).** A path selection $p = (p_1, \ldots, p_n) \in \mathcal{P}$ is an $\epsilon$-approximate Pure Nash Equilibrium if and only if for each agent $i \in [n],$

$$C_i(p_i, p_{-i}) \leq C_i(p'_i, p_{-i}) + \epsilon \quad \text{for all} \quad p'_i \in \mathcal{P}_i.$$

A prob. distribution $\pi^* := (\pi^*_1, \ldots, \pi^*_n) \in \Delta(\mathcal{P}_1) \times \cdots \times \Delta(\mathcal{P}_n)$ is an $\epsilon$-approximate Mixed Nash Equilibrium (MNE) if and only if for each agent $i \in [n],$

$$\mathbb{E}_{\pi^*_1, \pi^*_{-i}}[C_i(p_i, p_{-i})] \leq \mathbb{E}_{\pi^*_1, \pi^*_{-i}}[C_i(p'_i, p_{-i})] + \epsilon.$$

for all $p'_i \in \Delta(\mathcal{P}_i)$. For convenience of notation, we will use the shorthand $c_i(\pi, \pi_{-i}) := \mathbb{E}_{\pi, \pi_{-i}}[C_i(p_i, p_{-i})]$.

**Theorem 2.1 (Folklore).** Congestion games always admit a Pure NE equilibrium $p^* \in \mathcal{P}$.

2.2. Implicit Description of the strategy Space

The strategy space $\mathcal{P}_i$ can be exponentially large w.r.t the number of resources $E$ ($|\mathcal{P}_i| \leq 2^{\Theta(m)}$). For example in
network congestion games, \( \mathcal{P}_i \) is the set of possible \((s_i, t_i)\)-paths in a given directed graph \( G(V, E) \) where \( s_i \in V \) is the starting node of agent \( i \) and \( t_i \in V \) is her destination. Typically the number of possible \((s_i, t_i)\)-paths is exponential in the number of edges.

Exponentially large strategy spaces can be described through the following implicit polytopal description (Kleer & Schäfer, 2021). Given a strategy \( p_i \in \mathcal{P}_i \), consider its equivalent description as \( \{0, 1\}^m \) vector

\[
x_{p_i} = \{ x \in \{0, 1\}^m : x_e = 1 \text{ iff } e \in p_i \}.
\]

Consider the set \( \tilde{\mathcal{P}}_i := \{ x_{p_i} \in \{0, 1\}^m : p_i \in \mathcal{P}_i \} \) and the polytope \( \mathcal{X}_i := \text{conv}(\tilde{\mathcal{P}}_i) \) denoting the convex hull of \( \tilde{\mathcal{P}}_i \). The polytope \( \mathcal{X}_i \) admits the alternative description \( \mathcal{X}_i := \{ x \in \{0, 1\}^m : A_i \cdot x \leq b_i \} \) where \( A_i \in \mathbb{R}^{r \times m} \) and \( b_i \in \mathbb{R}^r \). Thus \( \mathcal{P}_i \) can be implicitly described as the extreme points of a set \( \mathcal{X}_i := \{ x \in \{0, 1\}^m : A_i \cdot x \leq b_i \} \) that can be described with \( O(m^r) \) fractional numbers.

In many classes of congestion games, the implicit polytopal description of \( \mathcal{P}_i \) is polynomial in the number of resources \( r := \text{poly}(m) \) while \( |\mathcal{P}_i| = 2^\Theta(m) \) (Kleer & Schäfer, 2021). For example in directed acyclic graphs (DAGs) the number all possible \((s_i, t_i)\)-paths can be \( 2^\Theta(m) \)-while the set of \((s_i, t_i)\)-paths can be equivalently described as the extreme points of the following path polytope (see Appendix B).

\[
\mathcal{X}_i := \left\{ x \in \{0, 1\}^m : \sum_{e \in \text{Out}(s_i)} x_e = 1, \sum_{e \in \text{In}(t_i)} x_e = 1, \sum_{e \in \text{In}(v)} x_e = \sum_{e \in \text{Out}(v)} x_e \quad \forall v \in V \setminus \{s_i, t_i\} \right\}
\]

where \( \text{In}(v) \), \( \text{Out}(v) \subseteq E \) denote the incoming, outgoing edges respectively of the node \( v \in V \).

**Remark 2.2.** One can always compute an implicit polytopal description \( \mathcal{X}_i \) given an explicit description of \( \mathcal{P}_i \). In the remaining paper, we assume access to the implicit polytopal description \( \mathcal{X}_i := \{ x \in \{0, 1\}^m : A_i \cdot x \leq b_i \} \) where \( A_i \in \mathbb{R}^{r \times m} \) and \( b_i \in \mathbb{R}^r \). We note that the regret bounds and the convergence rates to NE of our proposed algorithm (Algorithm 2) only depend on the \( r \) and \( m \) and are totally independent of \( \max_{i \in [n]} r_i \). The exact same holds for the convergence guarantees of the update rule proposed by (Cu et al., 2022). On the other hand, the running time of Algorithm 2 is polynomial in \( r_i \) and \( m \) while the time and space complexity of algorithm of (Cu et al., 2022) scales linearly with \( \max_{i \in [n]} |\mathcal{P}_i| \) even if \( r_i = \text{poly}(m) \).

2.3. Semi-Bandit Learning Dynamics

In game dynamics with semi-bandit feedback, each agent iteratively updates her strategies based on the congestion cost of the previously selected resources so as to minimize her overall experienced cost. Semi-bandit dynamics in congestion are described in Algorithm 1.

**Protocol 1 Semi-Bandit Game Dynamics**

1: for each round \( t = 1, \ldots, T \) do
2: Each agent \( i \in [n] \) (randomly) selects a strategy \( p_i^t \in \mathcal{P}_i \) and suffers cost

\[
C_i(p_i^t, p_v^t) := \sum_{e \in p_i^t} c_e^t(p_i^t, p_v^t)
\]

3: Each agent \( i \in [n] \) learns only the congestion costs \( c_e(p_i^t, p_v^t) \) of her selected resources \( e \in p_i^t \).
4: end for

It is not clear how a selfish agent \( i \) should update her strategy to minimize her overall congestion cost since the loads depend on strategies of the other agents that can arbitrarily change over time. As a result, agent \( i \) tries to minimize her experienced cost under the worst-case assumption that the cost of the resources \( c_e \) are selected by a malicious adversary. We refer to the latter online learning setting as Online Resource Selection, described in Algorithm 2.

**Protocol 2 Online Resource Selection**

1: for each round \( t = 1, \ldots, T \) do
2: Agent \( i \) selects a prob. distribution \( \pi_i^t \in \Delta(\mathcal{P}_i) \).
3: An adversary selects a cost function \( c^t : E \rightarrow \mathbb{R}_{\geq 0} \).
4: Agent \( i \) samples a path \( p_i^t \sim \pi_i^t \) and suffers cost

\[
C_i(p_i^t, c^t) := \sum_{e \in p_i^t} c_e^t.
\]

4: Agent \( i \) learns the costs \( c_e^t \) for all resources \( e \in p_i^t \) and updates \( \pi_i^{t+1} \in \Delta(\mathcal{P}_i) \).
5: end for

A semi-bandit online learning algorithm \( \mathcal{A} \) for the Online Resource Selection selects a strategy \( p_i^t \in \mathcal{P}_i \) based on the observed costs of selected resources in rounds before \( t \). The quality of an algorithm \( \mathcal{A} \) is measured through the notion of regret capturing the overall cost of algorithm \( \mathcal{A} \) with respect to the overall cost of the best strategy in hindsight.

**Definition 2.3.** The regret of an online learning algorithm \( \mathcal{A} \) is defined as \( R_{\mathcal{A}}(T) := \max_{c_1, \ldots, c^T} \left[ \sum_{t=1}^T \mathbb{E}_{\pi_i^t} \left[ C_i(p_i^t, c^t) \right] \right] - \min_{p_i^t \in \mathcal{P}_i} \sum_{t=1}^T C_i(p_i^t, c^t) \]

In case \( R_{\mathcal{A}}(T) = o(T) \), i.e., it is sublinear in \( T \), the algorithm \( \mathcal{A} \) is called no-regret.
In the context of congestion games, if agent \( i \) adopts a no-regret algorithm \( \mathcal{A} \), her time-averaged cost approaches the cost of the optimal path with rate \( \mathcal{R}_\mathcal{A}(T)/T \to 0 \) no matter what the strategies of the other agents are.

We conclude the section with the main result of our work.

**Main Result** There exists a no-regret semi-bandit online learning algorithm that admits \( \mathcal{R}_\mathcal{A}(T) = O(n^3T^{4/5}) \) regret for Online Resource Selection (Algorithm 2). Moreover if Algorithm 2 is adopted by all agents, then the agents converge to an \( \epsilon \)-Mixed NE after \( \mathcal{O}(n^{6.5}m^7/\epsilon^3) \) time-steps.

In Section 3 we present our proposed semi-bandit online learning algorithm, called Semi-Bandit Gradient Descent with Caratheodory Exploration (Algorithm 2). In Section 3 we also present its regret guarantees (Theorem 3.6) and its convergence properties (Theorem 3.8 and Corollary 1). In Section 4 we provide the main steps for establishing the no-regret properties of Algorithm 2 while in Section 5 we present the main ideas of the convergence proof. Finally in Section 6 we experimentally evaluate our algorithm in network congestion games.

### 3. Semi-Bandit Gradient Descent with Caratheodory Exploration

In this section, we present our algorithm called Bandit Gradient Descent with Caratheodory Exploration for the Online Path Selection Problem.

#### 3.1. Exploring via Caratheodory Decomposition

To avoid the exponentially large strategy space, we re-parametrize the problem using a fractional selection \( x_i^t \) of the edges which represents the probability edge \( e \) is selected at round \( t \), i.e., \( x_i^t = \mathbb{P}[e \in p^t] \). The major challenge now is to ensure that each resource is sufficiently explored. We resolve the exploration problem by introducing the notion of Bounded-Away Description Polytope (Definition 3.2) which guarantees that the selection probability of each useful resource is greater than an exploration parameter \( \mu > 0 \). The only similar idea in the literature we are aware of comes from (Chen et al., 2021) that used it in the context of online predictions with experts advice.

We proceed with some necessary definitions and two important characterization lemmas.

**Definition 3.1.** The set of active resources for agent \( i \in [n] \) is the set \( E_i := \{ e \in E : e \in p_i \} \) for some \( p_i \in \mathcal{P}_i \).

**Lemma 1.** Given the implicit description \( X_i \) of the strategy space \( \mathcal{P}_i \), the set of active resources \( E_i \) can be computed in polynomial-time.

**Definition 3.2.** The \( \mu \)-Bounded-Away Description Polytope for an exploration parameter \( \mu > 0 \) is defined as,

\[
X_i^\mu := \left\{ x \in X_i : x_e \geq \mu \quad \forall e \in E_i \right\}
\]

**Lemma 2.** The set \( X_i^\mu \) is non-empty for all \( \mu \leq 1/|E_i| \).

Notice that if a point \( x_i \in X_i^\mu \) then \( x_i \in X_i \). Since \( X_i = \text{conv}(\mathcal{P}_i) \) then any point \( x_i \in X_i^\mu \) can be decomposed to a probability distribution over strategies \( p_i \in \mathcal{P}_i \).

**Theorem 3.3** (Carathéodory Decomposition). For any point \( x \in X_i \) there exists a probability distribution \( \pi_x \in \Delta(\mathcal{P}_i) \) with support at most \( m+1 \) strategies \( p_i \in \mathcal{P}_i \) such that for all edges \( e \in E \),

\[
x_e = \sum_{p_i \in \pi_x} \mathbb{P}_{p_i} \left[ p_i \text{ is selected} \right].
\]

Such a distribution (it may not be unique) \( \pi_x \) is called a Carathéodory decomposition of \( x \).

A Carathéodory decomposition of a point \( x \in X_i^\mu \) can be computed in polynomial-time through the decomposition algorithm described in (Grötschel et al., 1988).

**Theorem 3.4** (Grötschel et al., 1988) Let a polytope \( X := \{ x \in \mathbb{R}^m : A \cdot x \leq b \} \) with \( A \in \mathbb{R}^{r \times m} \) and \( b \in \mathbb{R}^r \). The Carathéodory Decomposition of any point \( x \in X \) can be computed in polynomial-time with respect to \( r \) and \( m \).

For the important special case of path polytopes described in Section 3 a point \( x \in X_i^\mu \) can be decomposed to a probability distribution over \( (s_i, t_i) \)-paths with a simple and efficient algorithm outlined in Algorithm 1.

**Algorithm 1** Efficient Computation of Carathéodory decomposition for Path Polytopes

1. **Input:** A point \( x \in X_i^\mu \)
2. \( \text{Res} \leftarrow \emptyset \)
3. **while** \( \sum_{e \in \text{Out}(s_i)} x_e > 0 \) **do**
4. \( \text{Let } A \leftarrow E \cap \{ e \in E : x_e > 0 \} \)
5. \( \text{Let } e_{\text{min}} \leftarrow \text{arg min}_{e \in A} x_e \) and \( x_{\text{min}} \leftarrow x_{e_{\text{min}}} \).
6. \( \hat{p}_i \leftarrow \text{An } (s_i, t_i) \)-path of \( G(V, A) \) with \( e_{\text{min}} \in \hat{p}_i \).
7. \( x_e \leftarrow x_e - x_{\text{min}} \) for all \( e \in \hat{p}_i \).
8. \( \text{Res} \leftarrow \text{Res} \cup \{ (\hat{p}_i, x_{\text{min}}) \} \).
9. **end while**
10. **return** \( \text{Res} \)

**Lemma 3.** Algorithm 1 requires \( \mathcal{O}(|V||E| + |E|^2) \) steps to give a Carathéodory Decomposition for path polytopes.

In our experimental evaluations for network congestion games, presented in Section 6, we use Algorithm 1 to more efficiently implement Online Gradient Descent with Caratheodory Exploration that we subsequently present.
3.2. Our Algorithm and Formal Guarantees

In this section we present our semi-bandit online learning algorithm called Online Gradient Descent with Caratheodory Exploration (Algorithm 2) for Online Resource Selection.

Algorithm 2 Online Gradient Descent with Caratheodory Exploration for Agent $i$

1: $\mu_1 \leftarrow 1/|E_i|$  
2: Agent $i$ selects an arbitrary $x_i^1 \in X_i^{\mu_1}$.  
3: for each round $t = 1, \ldots, T$ do  
4: Agent $i$ computes a Caratheodory decomposition $\pi_i^t \in \Delta(P_i)$ for $x_i^t \in X_i^{\mu_t}$.  
5: Agent $i$ samples a strategy $p_i^t \sim \pi_i^t$ and suffers cost,  
\[ C_i^t(p_i^t, c^t) := \sum_{e \in p_i^t} c_i^t. \]

/* $c_i^t$ is the cost of edge $e$ with load the number of agents that chose $e$ at time $t$. */

6: Agent $i$ sets $\hat{c}_i^t \leftarrow c_i^t \cdot 1[ e \in p_i^t ] / x_i^t$ for all $e \in E$.  
7: Agent $i$ updates $x_i^{t+1} \in X_i^{\mu_{t+1}}$ as,  
\[ x_i^{t+1} = \Pi X_i^{\mu_{t+1}} \left[ x_i^t - \gamma_t \cdot \hat{c}_i^t \right], \]

where $\gamma_t \leftarrow t^{-3/5}$ and $\mu_t \leftarrow \min(1/m_t, t^{-1/5})$.  
8: end for

At Step 4, OGD – CE performs a Caratheodory Decomposition to convert the fractional point $x_i^t \in X_i^{\mu_t}$ into a probability distribution $\pi_i^t$ over pure strategies $p_i \in P_i$. The latter guarantees that the experienced cost equals the fractional congestion cost, i.e.,
\[ \mathbb{E} \left[ \sum_{e \in p_i^t} c_i^t \right] = \langle c^t, x_i^t \rangle. \]  

At Step 7, OGD – CE runs a step of Online Gradient Descent to the time-expanding polytope $X_i^{\mu_t}$ that approaches $X_i$ as $t \to \infty$. The latter is crucial as it gives the following:

Lemma 4. The estimator $\hat{c}_i^t = c_i^t \cdot 1[ e \in p_i^t ] / x_i^t$ satisfies

1. $\mathbb{E}[\hat{c}_i^t] = c_i^t$ for $e \in E_i$ (Unbiasness).
2. $|\hat{c}_i^t| \leq c_{\text{max}}/\mu_t$ for $e \in E_i$ (Boundness).

Remark 3.5. Projecting to the time-expanding polytope $X_i^{\mu_t}$ and not to $X_i$ is crucial since in the latter case we cannot control the variance of the estimator, as $x_i^t$ can go to 0 arbitrarily fast. On the other hand, it is crucial to compute a Caratheodory Decomposition with respect to $X_i$ and not with respect to $X_i^{\mu_t}$ since the extreme points of $X_i^{\mu_t}$ do not correspond to pure strategies $p_i \in P_i$.

We conclude the section by presenting the formal guarantees of Algorithm 2. In Theorem 3.6 we establish the no-regret property of Algorithm 2. In Theorem 3.8 and Corollary 1 we present its convergence guarantees.

Theorem 3.6. Let $p_i^1, \ldots, p_i^T \in P_i$ the sequence of strategies produced by Algorithm 2 given as input the costs $c^1, \ldots, c^T$ with $\|c^t\|_\infty \leq c_{\text{max}}$. Then with probability $1 - \delta$,
\[ \sum_{t=1}^T \sum_{e \in p_i^t} c_i^t - \min_{p_i^t \in P_i} \sum_{e \in p_i^t} c_i^t \leq O \left( m c_{\text{max}}^2 T^{4/5} \log(1/\delta) \right). \]

Remark 3.7. Setting $\delta := (1/m T c_{\text{max}})$ directly implies that Algorithm 2 admits regret $O(m^2 c_{\text{max}}^2 T^{3/5})$ regret. The better $O(m^{3/2} T^{3/4})$ regret bound can be obtained by selecting the parameters $\gamma_t = c_{\text{max}}^{-1} m^{-1/2} t^{-3/4}$ and $\mu_t = (1/m_t, m_t^{-1/2} - 1/4)$. However such a parameter selection leads to $O(T^{-1/8})$ convergence rate to NE.

In Theorem 3.8 we establish that the agents converge to a NE if all agents adopt Algorithm 2.

Theorem 3.8. Let $\pi^1, \ldots, \pi^T$ the sequence of strategy profiles produced if all agents adopt Algorithm 2. Then for all $T \geq \Theta(m^{12.5} n^{7.5}/\epsilon^8)$,
\[ \frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^T \max_{\pi \in \Theta(P)} \left[ c_i(\pi^t, \pi^t_{\pi_i^t}) - \min_{\pi_i \in P_i} c_i(\pi_i, \pi_{\pi_i^t}) \right] \right] \leq \epsilon. \]

The same holds for $T \geq \Theta(n^{6.5} m^7/\epsilon^5)$ in case the agents know $n, m$ and select $\gamma_t := \Theta(m^{-4/5} n^{-8/5} c_{\text{max}}^{-1} t^{-3/5})$ and $\mu_t := \Theta(n^{-6/5} m^{-11/10} t^{-1/5})$.

We note that the exact same notion of best-iterate convergence (as in Theorem 3.8) is considered in [Cui et al., 2022;Leonardos et al., 2022;Ding et al., 2022;Anagnostides et al., 2022]. In Corollary 1 we present a more clear interpretation of Theorem 3.8.

Corollary 1. In case all agents adopt Algorithm 2 for $T \geq \Theta(n^{6.5} m^7/\epsilon^5)$ (resp. $\Theta(m^{12.5} n^{7.5}/\epsilon^8)$) then with probability $\geq 1 - \delta$,

- $(1 - \delta)T$ of the strategy profiles $\pi^1, \ldots, \pi^T$ are $\epsilon/\delta^2$-approximate Mixed NE.
- $\pi^t$ is an $\epsilon/\delta$-approximate Mixed NE once $t$ is sampled uniformly at random in $\{1, \ldots, T\}$.

4. Sketch of Proof of Theorem 3.6

In this section, we present the basic steps of the proof of Theorem 3.6. Due to Equation 2 established by the Caratheodory decomposition and the fact that $|\hat{c}_i^t| \leq c_{\text{max}}/\mu_t$ we establish the following concentration result for the quantity $\sum_{t=1}^T \mathbb{E} \left[ \sum_{e \in p_i^t} c_i^t - \langle c^t, x_i^t \rangle \right]$.
As a result, in the rest of the section we bound the term $X$ and separately bound each of the $A$, $C$ terms.

Up next, we decompose the right-hand term of Equation 4 as:

$$\sum_{t=1}^{T} \langle e^t, x^t \rangle = \sum_{t=1}^{T} \langle e^t, x^t - x^*_t \rangle + \sum_{t=1}^{T} \langle e^t, x^*_t - x^*_\mu \rangle + \sum_{t=1}^{T} \langle e^t - e^t, x^*_t - x^*_\mu \rangle.$$

The bound on term (A) is established in Lemma 6:

**Lemma 6.** Let the sequences $x^1, \ldots, x^T \in X_i$ and $\hat{p}^1, \ldots, \hat{p}^T \in P_i$ produced by Algorithm 2 Then,

$$\sum_{t=1}^{T} \langle e^t, x^t \rangle \leq \frac{2m}{\gamma T} + m c^2_{\max} \sum_{t=1}^{T} \mu_t.$$

The proof of Lemma 6 is based on extending the arguments of (Zinkevich, 2003) for Online Projected Gradient Descent. The basic technical difficulty comes from the fact that in Step 7, Algorithm 2 projects in the time-changing feasibility set $X^\mu_t$ while in the analysis of (Zinkevich, 2003) the feasibility set is invariant.

The term (B) quantifies the suboptimality of the projections of $x^*_t$ on the time expanding polytopes $X^\mu_t$. Notice that $x^*_t$ is possibly outside the $X^\mu_t$ where Algorithm 2 projects to.

**Lemma 7.** (Sub-optimality of Bounded Polytopes) Let a sequence of costs $c^1, \ldots, c^T$ with $\|c^t\|_\infty \leq c^\max$. Then,

$$\sum_{t=1}^{T} \langle e^t, x^*_t - x^*_\mu \rangle \leq m^2 c^\max \sum_{t=1}^{T} \mu_t.$$

where $x^*_t$ and $x^*_\mu$ are introduced in Definition 4.1.

Finally, we bound the term (C) that quantifies the concentration of the cost estimators built at Step 7 of Algorithm 2 and the realized costs. The latter is established in Lemma 5 and its proof lies on the Unbiasness and Boundness property of the estimator $c_t$.

**Lemma 8.** Let $\hat{c}^1, \ldots, \hat{c}^T$ the sequence produced by Algorithm 2 given as input the sequence of costs $c^1, \ldots, c^T$. Then with probability $1 - \delta/2$,

$$\sum_{t=1}^{T} \langle e^t - \hat{c}^t, x^*_t - x^*_\mu \rangle \leq \frac{mc^\max}{\mu T \log(2m/\delta)}.$$

**5. Sketch of Proof of Theorem 3.8**

In this section, we provide the main steps and ideas for proving Theorem 3.8 establishing that in case all agents adopt Algorithm 1 the overall system converges to a Mixed NE. We first introduce some important preliminary notions.

**Definition 2 (Fractional Potential Function).** Let the fractional potential function $\Phi : X_1 \times \ldots \times X_n \to \mathbb{R}$

$$\Phi(x) \equiv \sum_{e \in E} \sum_{S \subseteq [n]} \prod_{j \in S} \prod_{j \notin S} (1 - x_{j,e}) \sum_{i=0}^{|S|} c_e(i).$$

The potential function of Definition 2 is crucial in our analysis since we can recast the problem of converging to NE into the problem of converging to a stationary point of $\Phi(x)$. 

---

**Lemma 5.** Let the sequences $x^1, \ldots, x^T \in X_i$ and $p^1, \ldots, p^T \in P_i$ produced by Algorithm 2 Then, with probability $1 - \delta/2$,

$$\sum_{t=1}^{T} \sum_{e \in P^t_i} c^t_e = \sum_{t=1}^{T} \langle c^t, x^t \rangle + O\left(\sqrt{\max(m/\delta)\sqrt{T}}\right).$$

Let $p^*_t \in P_i$ denote the optimal strategy for the sequence of costs $c^1, \ldots, c^t$ and $x^*_t \in X_i$ the corresponding $\{0, 1\}^m$ extreme point of $X_i$. Then Lemma 5 implies that with probability $1 - \delta/2$,

$$\sum_{t=1}^{T} \left( \sum_{e \in P^t_i} c^t_e - \sum_{e \in P^*_t} c^t_e \right) = \sum_{t=1}^{T} \langle c^t, x^t - x^*_t \rangle + O\left(\sqrt{T}\right).$$

As a result, in the rest of the section we bound the term $\sum_{t=1}^{T} \langle c^t, x^t - x^*_t \rangle$.

Unfortunately, this term cannot be directly control because at any step $t$ the comparator point $x^*_t$ is not necessarily in $X^\mu_t$. To overcome the issue we construct a sequence of comparator points $\{x^*_\mu_t\}_{t=1}^T$ that are guaranteed to satisfy $x^*_\mu_t \in X^\mu_t$.

Formally, we have the following definition.

**Definition 4.1.** Let $p^*_t \in P_i$, the optimal strategy for the sequence $c^1, \ldots, c^t$ and $x^*_t \in X_i$ its corresponding extreme point in $X_i$. Moreover, consider constructing a collection of strategies $D = \{\hat{p}^t\}_{t=1}^m$ sampled as follows. For any active resource $e \in E_i$, add to $D$ a strategy $\hat{p} \in P_i$ such that $e \in \hat{p}$. Considering the collected strategies in a vector form, i.e. elements of $\{0, 1\}^m$, we define

$$x^*_\mu_t \equiv (1 - m \mu_t) x^*_t + \mu_t \sum_{t=1}^{m} \hat{p}^t.$$
Definition 3. A point \( x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \) is called an \((\epsilon, \mu)\)-stationary point of \( \Phi(x) \) if and only if
\[
\left\| x - \Pi_{\mathcal{X}^n} \left[ x - \frac{1}{2n^2c_{\max} \sqrt{m}} \nabla \Phi(x) \right] \right\| \leq \epsilon
\]
where \( \lambda^\mu \triangleq \mathcal{X}_1^\mu \times \ldots \times \mathcal{X}_n^\mu \).

In Lemma 9 we establish that the potential function is smooth and uniformly bounded over its domain.

Lemma 9. The potential function \( \Phi(\cdot) \) of Definition 2 is smooth. More precisely,
\[
\| \nabla \Phi(x) - \nabla \Phi(x') \|_2 \leq 2n^2c_{\max} \sqrt{m} \cdot \| x - x' \|_2
\]
for all \( x, x' \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \). Moreover, the potential function \( \Phi(x) \) is bounded by \( mnc_{\max} \). We also denote \( \lambda \triangleq (2n^2c_{\max} \sqrt{m})^{-1} \).

In Lemma 10, we formalize the link between approximate NE and approximate stationary points of the \( \Phi(x) \).

Lemma 10. Let \( \pi = (\pi_1, \ldots, \pi_n) \in \Delta(\mathcal{P}_1) \times \ldots \times \Delta(\mathcal{P}_n) \) and \( x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \) such that for all resources \( e \in E \),
\[
x_{i, e} = \mathbb{P}_{p_i \sim \pi_i} [ e \in p_i ],
\]
in case \( x \) is an \((\epsilon, \mu)\)-stationary point of \( \Phi(x) \) then \( \pi \) is a \((4n^2mcm + 2n^2mc_{\max} \mu)\)-approximate Mixed NE.

5.1. Convergence to stationary points

In this section, we show that in case all agents use Algorithm 1 to (randomly) select their strategies, the produced sequence \( x^t = (x_1^t, \ldots, x_n^t) \) converges to a stationary point of the potential function \( \Phi(x) \).

We first show that the updates generated by each agent’s individual implementation of Algorithm 2 can be equivalently described as the update performed by stochastic gradient descent on the potential function projected on the time-varying polytope \( \lambda^\mu \times \ldots \times \mathcal{X}_n^\mu \).

Theorem 5.1. If each agent \( i \) (randomly) selects its strategy according to Algorithm 1 then the produced sequence \( x^1, \ldots, x^T \) can be equivalently described as
\[
x^{t+1} = \Pi_{\lambda^\mu} \left[ x^t - \lambda_t \cdot \nabla \right]
\]
where the estimator \( \nabla \triangleq [c_1^t, \ldots, c_n^t] \) (\( c_i^t \) is the cost estimate generated by player \( i \) according to Step 1 in Algorithm 2) satisfies
1. \( \mathbb{E} [\nabla] = \nabla \Phi(x^t) \) and 2. \( \mathbb{E} [|| \nabla ||^2] \leq \frac{mnc_{\max} m}{\mu} \).

The main technical contribution of this section is to establish that the sequence \( x^1, \ldots, x^T \) produced by Equation 3 converges to an \((\epsilon, \mu)\)-stationary point of Definition 3. The major challenge in proving the latter comes from the fact that in Equation 3 the projection step is respect to the time-changing polytope \( \lambda^\mu \) while the projection in the definition of \((\epsilon, \mu)\)-stationary point is with respect to the polytope \( \lambda^\mu \).

Theorem 5.2. Let \( G(x) = \Pi_{\lambda^\mu} [ x - \lambda \nabla \Phi(x) ] \) and the sequence \( x^1, \ldots, x^T \) produced by Equation 3 then
\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [ || G(x^t) || ]
\]
is upper bounded by
\[
2 \left( \frac{X^2 m c_{\max} K}{2T \gamma_t} + \frac{\lambda c_{\max}}{2 \gamma_t} T \sum_{t=1}^{T} \frac{1}{\mu_t} + \frac{8 \sqrt{nm} T}{T} \sum_{t=1}^{T} \mu_t \right).
\]

6. Experiments

We aim at verifying our theoretical statement by providing experiments in network congestion games. We consider a multigraph with chain topology composed by a set of nodes.
The document contains a discussion on semi-bandit dynamics in congestion games. It introduces SBGD-CE, the first no-regret online learning algorithm with semi-bandit feedback that once adopted by all agents in a congestion game converges to NE in the best-iterate sense. The empirical evaluation inspires different future directions, in particular establishing last iterate convergence rates to NE as well as tightening the rates for best-iterate convergence.

Acknowledgements

This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043), by the Swiss National Science Foundation (SNSF) under grant number 200021_205011, by a PhD fellowship of the Swiss Data Science Center, a joint venture between EPFL and ETH Zurich, by Innosuisse under agreement number 100.960 IP-ICT. Ioannis Panageas would like to acknowledge UCI start-up grant. Xiao Wang acknowledges the Grant 202110458 from SUFE and support from the Shanghai Research Center for Data Science and Decision Technology.
References


A. Additional related work on congestion games

Congestion games, proposed in (Rosenthal [1973], are amongst the most well known and extensively studied class of games and have been successfully employed in myriad modeling problems. Congestion games have been proven to be isomorphic to potential games (Monderer & Shapley [1996], and as a result, they always admit a potential function and a pure Nash equilibrium. Moreover, (typical) due to existence of multiple Nash equilibria, Price of Anarchy has been proposed in (Koutsoupias & Papadimitriou [1999]) for the purpose of efficiency guarantees in congestion games and is arguably amongst the most developed areas within algorithmic game theory, e.g., (Koutsoupias & Papadimitriou [1999], Roughgarden & Tardos [2002], Christodoulou & Koutsoupias [2005], Fotakis et al. [2005], de Keijzer et al. [2010], Roughgarden [2009]). It is folklore knowledge that better-response dynamics in congestion games converge. In these dynamics, in every round, exactly one agent deviates to a better strategy. Convergence is guaranteed as the potential function always decreases along better response dynamics. Notwithstanding better dynamics converges, it has been shown that computing a pure Nash equilibrium is PLS-complete (Fabrikant et al. [2004]) and computing a (possible mixed) Nash equilibrium is CLS-complete (Babichenko & Rubinstein [2021]). Thus, it is unlikely to be able to provide an algorithm that computes (pure or mixed) Nash equilibrium in congestion games and runs in polynomial time (in the description of the game).

B. Proof for Section 2

B.1. Path Polytope and Directed Acyclic Graphs

Definition B.1. A directed graph \( G(V, E) \) is called acyclic in case there are no cycles in \( G(V, E) \).

Definition B.2. Let a directed acyclic graph \( G(V, E) \) and the vertices \( s_i, t_i \in V \). The \((s_i, t_i)\)-path polytope is defined as follows,

\[
\mathcal{X}_{si} \equiv \left\{ x \in \{0,1\}^m : \sum_{e \in \text{Out}(s_i)} x_e = 1, \sum_{e \in \text{In}(s_i)} x_e = \sum_{e \in \text{Out}(v) \setminus \{s_i, t_i\}} x_e \forall v \in V \setminus \{s_i, t_i\}, \sum_{e \in \text{In}(t_i)} x_e = 1 \right\}
\]

Lemma 11. The extreme points of the \((s_i, t_i)\)-path polytope \( \mathcal{X}_{si} \) correspond to \((s_i, t_i)\)-paths of \( G(V, E) \) and vice versa.

Proof. We first show that an \((s_i, t_i)\)-path \( p_i \in \mathcal{P}_i \) corresponds to an extreme point of \( \mathcal{X}_{si} \). Given an \((s_i, t_i)\)-path \( p \in \mathcal{P}_i \) consider the point \( x_{p_i} \) of the polytope with \( x_{e} = 1 \) for all \( e \in p_i \) and \( x_{e} = 0 \) otherwise. Let assume that \( x_{p_i} \) is not an extreme point of \( \mathcal{X}_{si} \). Notice that \( x_{p_i} \) satisfies all the constraints of \( \mathcal{P}_i \), and since is a \( \{0,1\} \)-vector it is an extreme point of \( \mathcal{X}_{si} \).

On the opposite direction we show that for any extreme point \( x \in \mathcal{X}_{si} \) the set of edges \( \{e \in E : x_e = 1\} \) is an \((s_i, t_i)\)-path of \( G(V, E) \). We first show that \( x \) is necessarily integral (\( x_e = 0 \) or \( x_e = 1 \)). Let assume that there exists an edge \( e \in E \) with \( x_e \in (0,1) \) and consider \( x_{min} \equiv \min_{x_e > 0} x_e \). Consider an \((s_i, t_i)\)-path containing edge \( e_{min} \equiv \arg \min_{x_e > 0} x_e \). Notice that for the point \( x \in \mathcal{X}_{si} \) the following holds,

\[
x = x_{min} \cdot p_{min} + (1 - x_{min}) \cdot \frac{x - p_{min}}{1 - x_{min}} y
\]

Notice that \( y \in \mathcal{X}_{si} \) and thus \( x \) can be written as convex combination of \( p_{min} \) and \( y \), meaning that \( x \) cannot be an extreme point of \( \mathcal{X}_{si} \). As a result, \( x \) is an \( \{0,1\}^m \)-vector. Now consider the set of edges \( p_{x} \equiv \{e \in E : x_e = 1\} \). Notice that node \( s_i \) admits exactly one edge \( e \in \text{Out}(s_i) \) belonging in \( p_{x} \). Similarly there exists exactly one edge \( e \in \text{In}(t_i) \) belonging in \( p_{x} \). Due to the fact that \( G(V, E) \) is acyclic, \( p_{x} \) is necessarily an \((s_i, t_i)\)-path.

\[\square\]

1Note that if two or more agents move at the same time then convergence is not guaranteed.
C. Proofs of Section 3

C.1. Proof of Lemma 1

Lemma 12. Given the implicit description $\mathcal{X}_i$ of the strategy space $\mathcal{P}_i$, the set of active resources $E_i$ can be computed in polynomial-time.

Proof. For each resource $e \in E$, consider the polytope $\mathcal{X}^e_i = \{ x \in \mathcal{X}_i : x_e = 1 \}$ and we check whether it is empty. Notice that $\mathcal{X}^e_i$ admits $r_i + 1$ linear constraints and thus checking for its feasibility is done in polynomial time with respect to $r_i$ and $m$.

The correctness of the above algorithm can be established with the following simple argument. Let an $x \in \mathcal{X}^e_i$ then $x \in \mathcal{X}_i$ and additionally $x_e^i = 1$. The latter implies that $x$ can be decomposed to $m + 1$ pure strategies $p_i \in \mathcal{P}_i$. Since $x_e^i = 1$ any strategy $p_i$ participating in the convex combination of $x$ must admit $e \in p_i$. Thus $e \in E_i$. On the opposite direction, $\mathcal{X}^e_i$ cannot be empty in case $e \in E_i$. Notice that in the latter direction there exits $p_i \in \mathcal{P}_i$ with $e \in \mathcal{P}_i$ and thus the corresponding $(0,1)$-vector $x_{p_i} \in \mathcal{X}^e_i$.

\[ \square \]

C.2. Proof of Lemma 2

Lemma 2. The set $\mathcal{X}^\mu_i$ is non-empty for all $\mu \leq 1/|E_i|$.

Proof. Initialize $y = (0, \ldots, 0) \in \{0,1\}^m$. For each resource $e \in E_i$ select a strategy $p^\mu_i \in \mathcal{P}_i$ and update $y$ as $y \leftarrow y + \mu \cdot x_{p^\mu_i}$. The consider $x := y/|E_i|$. Notice that $y \in \mathcal{X}_i$ as convex combination $x_{p^\mu_i} \in \mathcal{X}_i$. At the same time, $y_e \geq 1/|E_i|$ for each $e \in E_i$. Thus, the set $\mathcal{X}^\mu_i$ is not empty for any $\mu \leq 1/|E_i|$.

\[ \square \]


Proof. First, we notice that the algorithm always successful in DAGs because, in virtue of Lemma [15] we can always find a path in Step 6 of Algorithm [1]. At this point, in order to study the complexity of each iteration, we notice that Step 5 requires at most $|E|$ read operations and Step 6 can be performed in $O(|V| + |E|)$ operations. Therefore, every iteration of the outermost loop require $O(|V| + |E|)$ operations. Finally, the number iterations in bounded by $|E|$ because Step 7 makes one coordinate of the point $x$ equal to 0 at every iteration and this ensures that after at most $|E|$ iterations $\sum_{c\in \text{Out}(s)} x_e = 0$. This conclude the proof because the total operation complexity is $O(|E|(|E| + |V|))$ as stated in the main text.

\[ \square \]

C.3. Proof of Lemma 4

For proving that $\hat{c}_e^t$ is unbiased, we have to recall that $x_e^t = E[1_{e \in p^t}]$ from which it follows that

$$E[\hat{c}_e^t] = E\left[\frac{c^t_e}{x_e^t} 1_{e \in p^t}\right] = \frac{c^t_e}{x_e^t} E\left[1_{e \in p^t}\right] = c^t_e$$

The second part of the proof concerns bounding the absolute value of the cost estimate. More precisely, we show that $|\hat{c}_e^t| \leq c_{\text{max}}/\mu_t$ for all $e \in E_i$. Indeed,

$$|\hat{c}_e^t| = \left|\frac{c^t_e}{x_e^t} 1_{e \in p^t}\right| \leq \frac{c_{\text{max}}}{\mu}$$

D. Proofs of Section 4

D.1. Proof of Lemma 5

Lemma 5. Let the sequences $x_1^T, \ldots, x_i^T \in \mathcal{X}_i$ and $p_1^T, \ldots, p_i^T \in \mathcal{P}_i$ produced by Algorithm [2] Then, with probablity $1 - \delta/2$,

$$\sum_{t=1}^T \sum_{e \in p_i^t} c_e^t = \sum_{t=1}^T \langle c^t, x_1^T \rangle + O\left(\sqrt{c_{\text{max}} \log(m/\delta)} \sqrt{T}\right).$$
Proof. Since the objective is a linear function \( \min_{x \in X} \sum_{t=1}^{T} \langle c_t, x \rangle = \min_{p \in P} \sum_{t=1}^{T} \sum_{e \in p} c_{t,e} \), then we have that
\[
\sum_{t=1}^{T} X_t = \sum_{t=1}^{T} \sum_{e \in E} c_{t,e} \cdot (\mathbb{E}_t [\mathbb{I}[e \in p^t]] - \mathbb{I}[e \in p^t])
\]
where \( \mathbb{E}_t \) is an expectation over the choice of \( p^t \) conditioned on the filtration adapted to the process \( (p^1, c^1, \ldots, p^{t-1}, c^{t-1}) \). As \( c^t \) is conditionally independent on \( p^t \), we get that
\[
\sum_{t=1}^{T} X_t = \sum_{e \in E} \sum_{t=1}^{T} (\mathbb{E}_t [c_{t,e} \mathbb{I}[e \in p^t]] - c_{t,e} \mathbb{I}[e \in p^t])
\]
Hence, we recognize the martingale difference sequence \( ((\mathbb{E}_t [c_{t,e} \mathbb{I}[e \in p^t]] - c_{t,e} \mathbb{I}[e \in p^t]))^T \) that satisfies \( [\mathbb{E}_t [c_{t,e} \mathbb{I}[e \in p^t]] - c_{t,e} \mathbb{I}[e \in p^t]] \leq c_{\text{max}} \) therefore by Azuma-Hoeffding inequality we can conclude that with probability \( 1 - \delta_1 \) for every resource \( e \in E \)
\[
\sum_{t=1}^{T} (\mathbb{E}_t [c_{t,e} \mathbb{I}[e \in p^t]] - c_{t,e} \mathbb{I}[e \in p^t]) \leq \sqrt{\frac{1}{2} c_{\text{max}} \log \frac{m}{\delta_1} T}.
\]

Lemma 6. Let the sequences \( x^1_t, \ldots, x^T_t \in X_t \) and \( c^1, \ldots, c^T \) produced by Algorithm 2. Then,
\[
\sum_{t=1}^{T} \langle c_t, x^t_i - x^*_\mu_t \rangle \leq \frac{2m}{\gamma_T} + mc_{\text{max}}^2 \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2}.
\]

Proof.
\[
\langle c_t, x^t_i - x^*_\mu_t \rangle \leq \langle x^t_i - x^{t+1}_i, x^t_i - x^*_\mu_t \rangle \frac{\gamma_t}{\mu_t^2}
\]
\[
= \frac{1}{2 \gamma_t} \left( \|x^*_\mu_t - x^t_i\|^2 - \|x^*_\mu_t - x^{t+1}_i\|^2 + \|x^{t+1}_i - x^t_i\|^2 \right)
\]
\[
\leq \frac{1}{2 \gamma_t} \left( \|x^*_\mu_t - x^t_i\|^2 - \|x^*_\mu - x^{t+1}_i\|^2 \right) + \frac{\gamma_t}{2} \|c^t\|^2
\]
\[
\leq \frac{1}{2 \gamma_t} \left( \|x^*_\mu_t - x^t_i\|^2 - \|x^*_\mu - x^{t+1}_i\|^2 \right) + \frac{c_{\text{max}}^2 \gamma_t}{2} \mu_t^2,
\]
where in the first and third equality we use the contraction property of the projection, in the second equality we developed the square and in the last inequality we the bound on the norm of the estimator. Summing over \( t \) we obtain
\[
\sum_{t=1}^{T} \langle c_t, x^t_i - x^*_\mu_t \rangle \leq \sum_{t=1}^{T} \frac{1}{2 \gamma_t} \left( \|x^*_\mu_t - x^t_i\|^2 - \|x^*_\mu_t - x^{t+1}_i\|^2 \right) + \frac{c_{\text{max}}^2 m \gamma_t}{2} \mu_t^2
\]
\[
\leq \sum_{t=1}^{T} \left( \frac{1}{2 \gamma_t} \|x^*_\mu_t - x^t_i\|^2 - \frac{1}{2 \gamma_{t+1}} \|x^*_\mu_{t+1} - x^{t+1}_i\|^2 \right)
\]
\[
+ \sum_{t=1}^{T} \left( \frac{\|x^*_\mu_t - x^{t+1}_i\|^2}{2 \gamma_{t+1}} - \|x^*_\mu_t - x^{t+1}_i\|^2 \right) + \frac{c_{\text{max}}^2 m \gamma_t}{2} \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2}
\]
\[
\leq \frac{\|x^*_\mu_T - x^T_i\|^2}{2 \gamma_T} - \frac{\|x^*_\mu_0 - x^1_i\|^2}{2 \gamma_1} + \frac{3m \gamma_T}{2 \gamma_1} + \frac{c_{\text{max}}^2 m \gamma_T}{2} \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2}
\]
\[
\leq \frac{2m}{\gamma_T} + \frac{c_{\text{max}}^2 m}{2} \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2}.
\]
Semi Bandit Dynamics in Congestion Games

Where in the second last inequality we used

\[
\sum_{t=1}^{T} \left( \frac{\|x_{\mu_{t+1}}^{*} - x_{t+1}^{*}\|^{2}}{2\gamma_{t+1}} - \frac{\|x_{\mu_{t}}^{*} - x_{t}^{*}\|^{2}}{2\gamma_{t}} \right) = \sum_{t=1}^{T} \left( \frac{\|x_{\mu_{t+1}}^{*} - x_{t+1}^{*}\|^{2} - \|x_{\mu_{t}}^{*} - x_{t}^{*}\|^{2}}{2\gamma_{t}} \right) \\
+ \sum_{t=1}^{T} \left( \frac{\|x_{\mu_{t+1}}^{*} - x_{t+1}^{*}\|^{2} - \|x_{\mu_{t}}^{*} - x_{t}^{*}\|^{2}}{2\gamma_{t}} \right) \\
= \sum_{t=1}^{\infty} \left( \frac{\|x_{\mu_{t}}^{*} - x_{t}^{*}\|^{2}}{2\gamma_{t}} \left( \frac{1}{2\gamma_{t+1}} - \frac{1}{2\gamma_{t}} \right) \right) \\
+ \sum_{t=1}^{T} \left( \frac{\|x_{\mu_{t}}^{*} + x_{t}^{*} - 2x_{t}^{*} + x_{\mu_{t+1}}^{*} - x_{t}^{*}}{2\gamma_{t}} \right) \\
\leq m \sum_{t=1}^{T} \left( \frac{1}{2\gamma_{t+1}} - \frac{1}{2\gamma_{t}} \right) + \sum_{t=1}^{T} \left( \|x_{\mu_{t}}^{*} + x_{t}^{*} - 2x_{t}^{*} + x_{\mu_{t+1}}^{*} - x_{t}^{*}\| \right) \\
\leq \frac{m}{2\gamma T} + \sqrt{m} \sum_{t=1}^{T} \left( \frac{\|1 - m\mu_{t} x_{t}^{*} + m\mu_{t} s - (1 - m\mu_{t+1}) x_{t}^{*} - m\mu_{t+1} s\|}{\gamma_{t}} \right) \\
= \frac{m}{2\gamma T} + m^{3/2} \sum_{t=1}^{T} \left( \frac{\|x_{t}^{*} - s\|}{\gamma_{t}} \right) \\
\leq \frac{m}{2\gamma T} + m^{2} \sum_{t=1}^{T} \left( \frac{\mu_{t} - \mu_{t+1}}{\gamma_{t}} \right) \\
\leq \frac{m}{2\gamma T} + m^{2} \sum_{t=1}^{T} \left( \frac{\mu_{t} - \mu_{t+1}}{\gamma_{t}} \right) \\
\leq \frac{m}{2\gamma T} + m^{2} \left( \frac{\mu_{1}}{\gamma_{1}} \right) \\
\leq \frac{m}{2\gamma T} + m \left( \frac{\mu_{1}}{\gamma_{1}} \right) = \frac{3m}{2\gamma T}
\]

\[\square\]

Lemma 7. (Sub-optimality of Bounded Polytopes) Let a sequence of costs \( c_{t}, \ldots, c_{T} \) with \( \|c_{t}\|_{\infty} \leq c_{\text{max}} \). Then,

\[
\sum_{t=1}^{T} \langle c_{t}, x_{\mu_{t}}^{*} - x_{t}^{*} \rangle \leq m^{2} c_{\text{max}} \sum_{t=1}^{T} \mu_{t},
\]

where \( x_{t}^{*} \) and \( x_{\mu_{t}}^{*} \) are introduced in Definition 4.1.

Proof.

\[
\sum_{t=1}^{T} \langle c_{t}, x_{\mu_{t}}^{*} - x_{t}^{*} \rangle \leq \sum_{t=1}^{T} \|c_{t}\|_{\infty} \|x_{t}^{*} - x_{\mu_{t}}^{*}\|_{1} \\
\leq m^{2} c_{\text{max}} \sum_{t=1}^{T} \mu_{t}
\]

where we used the following bound on \( \|x_{t}^{*} - x_{\mu_{t}}^{*}\|_{1} \). Recall that \( s \) is the vector as constructed in Remark 4.2

\[
\|x_{t}^{*} - x_{\mu_{t}}^{*}\|_{1} = \|x_{t}^{*} - (1 - m\mu_{t}) x_{t}^{*} - m\mu_{t} s\|_{1}
\]

16
Lemma 8. Let $\hat{\epsilon}^1, \ldots, \hat{\epsilon}^T$ the sequence produced by Algorithm 2 given as input the sequence of costs $\epsilon^1, \ldots, \epsilon^T$. Then with probability $1 - \delta/2$,
\[
\sum_{t=1}^{T} \langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle \leq \frac{m c_{\max} \mu_T}{\mu} \sqrt{T \log(2m/\delta)}.
\]

Proof. We denote $\mathcal{F}^t$ a filtration adapted to the sigma algebra induced by the random variables $\{x^t\}_{t=1}^t$ and let $\mathbb{E}_t$ denote expectation conditioned on $\mathcal{F}^t$. We recognize that $\langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle$ is a bounded martingale difference sequence. Indeed,
\[
\mathbb{E}_t \left[ \langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle \right] = \mathbb{E}_t \left[ \langle \mathbb{E}_t \left[ \epsilon^t \right] - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle \right] = \langle \mathbb{E}_t \left[ \hat{\epsilon}^t \right] , x^t - x^*_\mu_t \rangle = 0
\]
where the last equality holds because $x^t$ is $\mathcal{F}^t$-measurable. In addition, $\langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle$ can be bounded as
\[
\left| \langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle \right| \leq \| \epsilon^t - \hat{\epsilon}^t \| \| x^t - x^*_\mu_t \| \leq 2 m c_{\max} (1 + 1/\mu_t)
\]
where we used $\| x^t - x^*_\mu_t \| \leq 2 m$ and $\| \epsilon^t - \hat{\epsilon}^t \| \leq c_{\max} (1 + 1/\mu_t)$. Therefore, by Azuma-Hoeffding, with probability $1 - \delta$,
\[
\sum_{t=1}^{T} \langle \epsilon^t - \hat{\epsilon}^t, x^t - x^*_\mu_t \rangle \leq 2 m c_{\max} \sqrt{2 \sum_{t=1}^{T} (1 + 1/\mu_t)^2 \log(1/\delta)} \leq 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)}
\]

At this point, we have all the elements for the proof of Theorem 3.6.

D.2. Proof of Theorem 3.6

Theorem 3.6. Let $p^1_1, \ldots, p^T_1 \in \mathcal{P}_1$ the sequence of strategies produced by Algorithm 2 given as input the costs $\epsilon^1, \ldots, \epsilon^T$ with $\| \epsilon^t \|_{\infty} \leq c_{\max}$. Then with probability $1 - \delta$,
\[
\sum_{t=1}^{T} \sum_{e \in P^t_1} \epsilon^t_e - \min_{p^t_1 \in P^t_1} \sum_{e \in P^t_1} \epsilon^t_e \leq O \left( m c_{\max}^2 T^{4/5} \log(1/\delta) \right)
\]

Proof. Combining the previous theorems we can obtained the following bounds
\[
\sum_{t=1}^{T} \sum_{e \in P^t_1} \epsilon^t_e - \min_{p^t_1 \in P^t_1} \sum_{e \in P^t_1} \epsilon^t_e \leq \sqrt{\frac{1}{2} c_{\max} \log \frac{m}{\delta_t}} \frac{m}{T} + \frac{c_{\max} m}{\mu_T} \sum_{t=1}^{T} \gamma_t + m c_{\max} \sum_{t=1}^{T} \mu_t
\]
\[
+ 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)}
\]
Now, replacing $\gamma_t = t^{-3/5}$ and $\mu_t = \min \{1/m, t^{-1/5}\}$, we obtain
\[
\sum_{t=1}^{T} \sum_{e \in P^t_1} \epsilon^t_e - \min_{p^t_1 \in P^t_1} \sum_{e \in P^t_1} \epsilon^t_e \leq \sqrt{\frac{1}{2} c_{\max} T \log \frac{m}{\delta_t}} + \frac{m}{T - 3/5} + \frac{c_{\max} m}{2} T^{2/5} \sum_{t=1}^{T} \frac{m}{t^{3/5}} + \frac{c_{\max} m}{2} T^{1/5} \sum_{t=1}^{T} \frac{m}{t^{1/5}}
\]
\[
+ m c_{\max} \frac{1}{m} + 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)}
\]
\[
= \sqrt{\frac{1}{2} c_{\max} T \log \frac{m}{\delta_t}} + m^2 T^{3/5} + \frac{6 c_{\max} m}{8} T^{4/5} + \frac{c_{\max} m}{2} T^{1/5} + m^{-1/3} + m c_{\max} \frac{5}{4} T^{4/5}
\]
which of the order stated in the main text.
E. Proof for Section 5

Lemma 9. The potential function $\Phi(\cdot)$ of Definition 2 is smooth. More precisely,

$$\|\nabla \Phi(x) - \nabla \Phi(x')\|_2 \leq 2n^2 \epsilon_{\max} \sqrt{m} \cdot \|x - x'\|_2$$

for all $x, x' \in X_1 \times \ldots \times X_n$. Moreover, the potential function $\Phi(x)$ is bounded by $nmc_{\max}$. We also denote $\lambda \triangleq \frac{1}{(2n^2 \epsilon_{\max} \sqrt{m})^{-1}}$

Proof. We start by taking a first partial derivative of the potential function for the fractional cost $\Phi(x)$, that is

$$\frac{\partial}{\partial x_{ie}} [\Phi(x)] = \sum_{S \subseteq [n], \bar{i} \in S, \bar{j} \notin S} \prod_{j \in S} x_{je} \prod_{j \notin S} (1 - x_{je}) \sum_{i=0}^{|S|} c_e(i) - \sum_{S \subseteq [n], \bar{i} \notin S, \bar{j} \notin S} \prod_{j \notin S} x_{je} \prod_{j \notin S} (1 - x_{je}) \sum_{i=0}^{|S|} c_e(i)$$

Taking a second a partial derivative with respect to $x_{je}$ we obtain

$$\frac{\partial^2}{\partial x_{je} \partial x_{ie}} [\Phi(x)] = \begin{cases} 
0 & \text{if } \bar{e} \neq e \\
0 & \text{if } \bar{i} = \bar{j} \\
\sum_{S \subseteq [n], \bar{i} \in S, \bar{j} \notin S} \prod_{j \in S} x_{je} \prod_{j \notin S} (1 - x_{je}) \sum_{i=0}^{|S|} c_e(i) - \sum_{S \subseteq [n], \bar{i} \notin S, \bar{j} \notin S} \prod_{j \notin S} x_{je} \prod_{j \notin S} (1 - x_{je}) \sum_{i=0}^{|S|} c_e(i) + 1 & \text{otherwise}
\end{cases}$$

Notice that the factor $\prod_{j \notin S} x_{je} \prod_{j \notin S} (1 - x_{je})$ is a multivariate probability distribution over the subsets of players in which the ones with index $\bar{j}$ and $\bar{i}$ select the edge $e$ therefore the sum $\sum_{S \subseteq [n], \bar{i} \in S, \bar{j} \notin S} \prod_{j \in S} x_{je} \prod_{j \notin S} (1 - x_{je}) = 1$. Similar observations hold for the other three terms in the nonzero partial derivatives. Then it follows that $-2n\epsilon_{\max} \leq \frac{\partial^2}{\partial x_{je} \partial x_{ie}} [\Phi(x)] \leq 2n\epsilon_{\max}$ where $c_{\max}$ is a uniform bound on $c_e(\cdot)$.

Furthermore, we observe that there are at most $mn(n - 1)$ nonzero elements of the Hessian and those satisfy $\left| \frac{\partial^2}{\partial x_{je} \partial x_{ie}} [\Phi(x)] \right|^2 \leq 4n^2 \epsilon_{\max}^2$. Therefore, we can bound the Hessian Frobenius norm as

$$\|\nabla^2 \Phi(x)\|_F = \sqrt{\sum_{e \in E} \sum_{\bar{e} \in E} \sum_{j=1}^N \sum_{i=1}^N \left| \frac{\partial^2}{\partial x_{je} \partial x_{ie}} [\Phi(x)] \right|^2} \leq \sqrt{mn(n - 1)4n^2 \epsilon_{\max}^2} \leq 2n^2 \epsilon_{\max} \sqrt{m}$$

Finally, since we have that for any matrix the Frobenius norm upper bounds the spectral norm, we can conclude that the maximum eigenvalue of $\nabla^2 \Phi(x)$ is at most $2n^2 \epsilon_{\max} \sqrt{m}$ and therefore the potential function is $2n^2 \epsilon_{\max} \sqrt{m}$-smooth.

E.1. Proof of Lemma 10

Before proving Lemma 10 we need an auxiliary lemma (Lemma 13) for which we need to introduce the notion of Caratheodory's decomposition image.

Definition E.1. We define as Caratheodory’s decomposition image of the set $X_i^\mu$, denoted as $\Delta(P_i^\mu)$, the set all all probability distributions $\pi_i \in P_i$ such that

$$\mathbb{P}_{\pi_i \sim \pi_i}[e \in p_i] \geq \mu \quad \text{for all } e \in E_i.$$
Lemma 13. Let \((\pi_i, \pi_{-i}) \in \Delta(P_1) \times \ldots \times \Delta(P_n)\) be a Caratheodory decomposition of \(x \in X_1 \times \ldots \times X_n\) and let \(x\) be a \((\epsilon, \mu)\)-stationary point according to Definition 2. Then for each agent \(i \in [n],\)

\[
c_i(\pi_i, \pi_{-i}) - \min_{\pi_i \in \Delta(P^n_i)} c_i(\bar{\pi}_i, \pi_{-i}) \leq 4n^2mc_{\max} \left\| x - \Pi X^n \left[ x - \frac{1}{2n^2c_{\max}\sqrt{m}} \nabla \Phi(x) \right] \right\|
\]

where \(\Delta(P^n_i) \subset \Delta(P_i)\) is the Caratheodory’s decomposition image of the set \(X^n_i\).

Proof. Let a probability distribution \(\bar{\pi}_i \in \Delta(P^n_i)\) and its marginalization \(\bar{x}_i\) i.e. \(\bar{x}_i = \mathbb{P}_{\pi_i \sim \bar{\pi}_i} [e \in p_i].\) Then Definition E.1 implies that \(\bar{x}_i \in X^n_i\). As a result, Lemma 16 implies

\[
c_i(\bar{\pi}_i, \pi_{-i}) - c_i(\bar{\pi}_i, \pi_{-i}) = \Phi(\bar{x}_i, x_{-i}) - \Phi(x_i, x_{-i})
\]

Let \(\lambda\) be the inverse of the smoothness parameter of \(\Phi\) that is \(\lambda = \frac{1}{2n^2c_{\max}\sqrt{m}}\). To simplify notation let \(\epsilon := \frac{1}{\lambda}\). Then Definition E.1 implies that \(x \in X^n_i\). As a result, Lemma 16 implies

\[
c_i(\bar{\pi}_i, \pi_{-i}) - c_i(\bar{\pi}_i, \pi_{-i}) = \Phi(\bar{x}_i, x_{-i}) - \Phi(x_i, x_{-i})
\]

Now plugging the value of \(\lambda\), setting \(\delta = x_i - \bar{x}_i\) on the left hand side, we get that for any player \(i \in [n]:\)

\[
(x_i - \bar{x}_i)\nabla g(\bar{x}_i) \leq 4n^2c_{\max} \epsilon
\]

The function \(g(\bar{x}_i)\) is a linear function (see Definition 2). By linearity of \(g(\bar{x}_i)\) it holds that

\[
\Phi(\bar{x}_i, x_{-i}) - \Phi(\bar{x}_i, x_{-i}) = g(\bar{x}_i) - g(x_i) = (x_i - \bar{x}_i)\nabla g(\bar{x}_i) \leq 4n^2c_{\max} \epsilon
\]

As a result for any \(\bar{\pi}_i \in \Delta(P^n_i),\)

\[
c_i(\bar{\pi}_i, \pi_{-i}) - c_i(\bar{\pi}_i, \pi_{-i}) \leq 4n^2c_{\max} \epsilon
\]

We conclude the section by presenting a slightly more general version of Lemma 10

Lemma 14. Let \(\pi = (\pi_1, \ldots, \pi_n) \in \Delta(P_1) \times \ldots \times \Delta(P_n)\) and \(x = (x_1, \ldots, x_n) \in X_1 \times \ldots \times X_n\) such that for all resources \(e \in E,\)

\[
x_{i,e} = \mathbb{P}_{\pi_i \sim \pi_i} [e \in p_i].
\]

Then the following holds,

\[
c_i(\pi_i, \pi_{-i}) - \min_{\pi_i \in \Delta(P^n_i)} c_i(\bar{\pi}_i, \pi_{-i}) \leq 4n^2mc_{\max} \cdot \left\| x - \Pi X^n \left[ x - \frac{1}{2n^2c_{\max}\sqrt{m}} \nabla \Phi(x) \right] \right\| + 2n^2c_{\max} \epsilon
\]

Proof. Lemma 13 implies that

\[
c_i(\pi_i, \pi_{-i}) - \min_{\pi_i \in \Delta(P^n_i)} c_i(\bar{\pi}_i, \pi_{-i}) \leq 4n^2mc_{\max} \cdot \left\| x - \Pi X^n \left[ x - \frac{1}{2n^2c_{\max}\sqrt{m}} \nabla \Phi(x) \right] \right\|
\]

As a result, we just need to bound

\[
\min_{\pi_i \in \Delta(P^n_i)} c_i(\bar{\pi}_i, \pi_{-i}) - \min_{\pi_i \in \Delta(P_i)} c_i(\bar{\pi}_i, \pi_{-i})
\]

19
Let $\pi^*_i \in \arg \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})$ and $\bar{x}_i^*$ its marginalization. Let also $\bar{x}_{i,\mu}^* := (1 - \mu)\bar{x}_i^* + \mu x_i$ where the vector $s$ is defined as in the proof of Lemma 7 and $\bar{\pi}_{i,\mu}$ its corresponding Caratheodory decomposition.

$$
\min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i}) - \min_{\pi_i \in \Delta(P_i)} c_i(\bar{\pi}_{i,\mu}, \pi_{-i}) \leq c_i(\bar{\pi}_{i,\mu}, \pi_{-i}) - c_i(\bar{\pi}_i^*, \pi_{-i})
$$

$$
= \Phi(\bar{x}_{i,\mu}^*, x_{i-}) - \Phi(\bar{x}_i^*, x_{i-})
$$

$$
= \sum_{e \in E} \sum_{S \subseteq \{i\} \cup S} \sum_{i \not\in S} (\bar{x}_{ie}^* - (1 - \mu)\bar{x}_{ie} - \mu x_i) \prod_{j \in S, j \neq i} x_{je} \prod_{j \not\in S} (1 - x_{je}) \sum_{l=0}^{\left\lvert S \right\rvert} c_e(l)
$$

$$
+ \sum_{e \in E} \sum_{S \subseteq \{i\} \cup S} (-\bar{x}_{ie} + (1 - \mu)\bar{x}_{ie} + \mu x_i) \prod_{j \in S} x_{je} \prod_{j \not\in S} (1 - x_{je}) \sum_{l=0}^{\left\lvert S \right\rvert} c_e(l)
$$

$$
\leq \sum_{e \in E} (\bar{x}_{ie}^* - (1 - \mu)\bar{x}_{ie} - \mu x_i) \prod_{j \in S} x_{je} \prod_{j \not\in S} (1 - x_{je}) + \sum_{e \in E} (-\bar{x}_{ie} + (1 - \mu)\bar{x}_{ie} + \mu x_i) \prod_{j \in S} x_{je} \prod_{j \not\in S} (1 - x_{je})
$$

$$
= 2m^2 n c_{\max} \mu
$$

\( \square \)

E.2. Proof of Theorem 5.1

**Theorem 5.1.** If each agent $i$ (randomly) selects its strategy according to Algorithm 7, then the produced sequence of vectors $x^1, \ldots, x^T$ can be equivalently described as

$$
x^{t+1} = \Pi_{X^{c_{t+1}}} [x^t - \gamma_t \cdot \nabla_t]
$$

(3)

where the estimator $\nabla_t \triangleq [c_1^t, \ldots, c_n^t]$ ($c_i^t$ is the cost estimate generated by player $i$ according to Step 7 in Algorithm 2) satisfies

1. $\mathbb{E}[\nabla_t] = \nabla \Phi(x^t)$ and 2. $\mathbb{E} \left\lVert \nabla_t \right\rVert^2 \leq \frac{n c_{\max} m}{\mu_t}$.

**Proof.** In the first part of the proof we show that the projection operator is separable. That is, for a generic set $X$

$$
\Pi_X(z) = \Pi_{X_1}[z_1], \ldots, \Pi_{X_n}[z_n]^T
$$

for any $z \in \mathbb{R}^{nm}$ in the form $z = [z_1, \ldots, z_n]$ with $z_j \in \mathbb{R}^m$ for all $j \in [n]$. To prove this, we proceed as follows

$$
\Pi_X[z] = \arg \min_{z' \in X} \left\lVert z - z' \right\rVert^2 = \arg \min_{z' \in X} \sum_{i=1}^{n} \left\lVert z_i - z'_i \right\rVert^2 = \sum_{i=1}^{n} \arg \min_{z_i' \in X_i} \left\lVert z_i - z_i' \right\rVert^2 = [\Pi_{X_1}[z_1], \ldots, \Pi_{X_n}[z_n]]
$$

Let $(\pi_1^*, \pi_{-1}^*)$ denote a Caratheodory Decomposition for $x^t := (x_1^t, x_{-1}^t)$. Then,

$$
\mathbb{E} \left\lVert \nabla_t \right\rVert^2 = \Pr_{\pi_{-1}^*} [\text{agent } i \text{ selects resource } e \text{ at round } t] \cdot \mathbb{E} \left[ c_i^t / x_{ie}^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right]
$$

$$
= x_{ie}^t \cdot \mathbb{E} \left[ c_i^t / x_{ie}^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right]
$$

$$
= \mathbb{E} \left[ c_i^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right]
$$

$$
= \sum_{S, e \subseteq \{1, \ldots, n\}} \left( \prod_{j \in S, j \neq i} \Pr_{\pi_j^*} [\text{agent } j \text{ selects } e \text{ at round } t] \cdot \prod_{j \not\in S, j \neq i} \Pr_{\pi_j^*} [\text{agent } j \text{ does not select } e \text{ at round } t] c_e(\left\lvert S \right\rvert + 1) \right)
$$

20
At the same time,

\[
\frac{\partial \Phi(x)}{\partial x_{ie}} = \sum_{S_{-i} \subseteq [n-1]} \prod_{j \in S_{-i}} x_{je} \prod_{j \notin S_{-i}} (1 - x_{je}) c_e (|S_{-i}| + 1)
\]

The second part of the proof concerns bounding the norm of the stochastic gradients of the potential function. More precisely, we show that \(\mathbb{E} \left[ \|\nabla \| \right] \leq \frac{nmc_{\max}}{\mu_t}\)

\[
\mathbb{E} \left[ \|\nabla \| \right]^2 = \sum_{i=1}^{n} \sum_{e \in E_t} \mathbb{P}_{r_{n_i}} [\text{agent } i \text{ selects resource } e \text{ at round } t] \cdot \mathbb{E} \left[ (c_{e} / x_{ie}^t)^2 \right] \cdot \text{agent } i \text{ selects resource } e \text{ at round } t
\]

\[
\leq \sum_{i=1}^{n} \sum_{e \in E_t} \mathbb{E} \left[ (c_{e} / x_{ie}^t)^2 \right] \cdot \text{agent } i \text{ selects resource } e
\]

\[
\leq \sum_{i=1}^{n} \sum_{e \in E_t} c_{\max} / \mu_t
\]

\[
\leq \sum_{i=1}^{n} \left| E_t \right| c_{\max}^2 / \mu_t
\]

\[
= \frac{nm c_{\max}^2}{\mu_t}
\]

\[
E.3. \text{Proof of Theorem 5.2}
\]

**Theorem 5.2.** Let \(G(x) = \Pi_{x^e} [x - \lambda \nabla \Phi(x)] - x\) and the sequence \(x^1, \ldots, x^T\) produced by Equation 3. Then \(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [\| G(x^t) \|_2] \) is upper bounded by

\[
2 \sqrt{\frac{\lambda^2 nm c_{\max}}{2T \gamma_T} + \frac{\lambda^2 c_{\max} nm \sum_{t=1}^{T} \frac{\gamma_{t}^2}{\mu_t}}{2T \gamma_T} + \frac{8 \sqrt{n} m n^3}{T} \sum_{t=1}^{T} \mu_t}
\]

**Proof.** We make use of the Moreau envelope function defined as follows,

\[
\phi_{\lambda, \gamma_{t+1}} (x) := \min_{y \in \mathcal{X}^{\gamma_{t+1}}} \left( \Phi(y) + \frac{1}{\lambda} \| x - y \|^2 \right)
\]

Let also \(y^{t+1} := \arg \min_{y \in \mathcal{X}^{\gamma_{t+1}}} \left( \Phi(y) + \frac{1}{\lambda} \| x^t - y \|^2 \right)\). It holds that

\[
\phi_{\lambda, \gamma_{t+1}} (x^{t+1}) \leq \Phi(y^{t+1}) + \frac{1}{\lambda} \| x^{t+1} - y^{t+1} \|^2
\]

\[
\leq \Phi(y^{t+1}) + \frac{1}{\lambda} \| x^t - \gamma_t \nabla_t - y^{t+1} \|^2
\]

\[
= \Phi(y^{t+1}) + \frac{1}{\lambda} \| x^t - y^{t+1} \|^2 + \frac{\gamma_t^2}{\lambda} \| \nabla_t \|^2 - \frac{2 \gamma_t}{\lambda} (x^t - y^{t+1})^T \nabla_t
\]

\[
\leq \Phi(y^{t+1}) + \frac{1}{\lambda} \| x^t - y^{t+1} \|^2 + \frac{\gamma_t^2}{\lambda} \| \nabla_t \|^2 - \frac{2 \gamma_t}{\lambda} (x^t - y^{t+1})^T \nabla_t
\]

21
where the first inequality comes from the definition of Moreau envelope, the second inequality comes from projection property on convex sets and last by the definition $y^{t+1} = \arg \min_{y \in \mathcal{X}^{t+1}} \left( \Phi(y) + \frac{1}{\lambda} \| x^t - y \|^2 \right)$.

Then, taking total expectation on both sides and using the monotonicity property of expectation, we have

$$
\mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^{t+1}) \right] \leq \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^t) \right] + \frac{\gamma_t^2}{\lambda} \mathbb{E} \left[ \| \nabla \| \right] - \frac{2\gamma_t}{\lambda} \mathbb{E} \left[ (x^t - y^{t+1})^T \nabla \phi(x^t) \right].
$$

At this point using the bound on the expected squared norm (see Theorem 5.1) of the cost estimator and using that the cost estimator is unbiased we obtain

$$
\mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^{t+1}) \right] \leq \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^t) \right] + \frac{\gamma_t^2}{\lambda} \mathbb{E} \left[ c_{\text{max}}^2 n m + \frac{2\gamma_t}{\lambda} \mathbb{E} \left[ (x^t - y^{t+1})^T \nabla \phi(x^t) \right] \right].
$$

At this point, using the fact that $\Phi(\cdot)$ is $\frac{1}{\lambda}$-smooth we get that

$$(y^{t+1} - x^t)^T \nabla \phi(x^t) \leq \Phi(y^{t+1}) - \Phi(x^t) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2$$

which implies that

$$
\mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^{t+1}) \right] \leq \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^t) \right] + \frac{\gamma_t^2}{\lambda} \mathbb{E} \left[ \frac{c_{\text{max}}^2 n m}{\mu_t} \right] + \frac{2\gamma_t}{\lambda} \mathbb{E} \left[ \Phi(y^{t+1}) - \Phi(x^t) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 \right].
$$

Due to the fact that $\mathcal{X}^{\mu_t} \subseteq \mathcal{X}^{t+1}$ we get that $\phi_{\lambda, \mathcal{X}^{t+1}}(x^t) \leq \phi_{\lambda, \mathcal{X}^{t+1}}(x^t)$ and thus

$$
\mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^{t+1}) \right] \leq \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x^t) \right] + \frac{\gamma_t^2}{\lambda} \mathbb{E} \left[ \frac{c_{\text{max}}^2 n m}{\mu_t} \right] + \frac{2\gamma_t}{\lambda} \mathbb{E} \left[ \Phi(y^{t+1}) - \Phi(x^t) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 \right].
$$

By reordering the terms and summing from $t = 1$ to $T$ we get that

$$
- \frac{2}{\lambda T} \sum_{t=1}^{T} \gamma_t \mathbb{E} \left[ \Phi(y^{t+1}) - \Phi(x^t) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 \right] \leq \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x_1) \right] - \mathbb{E} \left[ \phi_{\lambda, \mathcal{X}^{t+1}}(x_T) \right] + \frac{c_{\text{max}}^2 n m}{\lambda T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t} \leq \frac{\Phi_{\text{max}}}{T} + \frac{c_{\text{max}}^2 n m}{\lambda T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t},
$$

(4)

since $\phi_{\lambda, \mathcal{X}^{t+1}}(x_1) \leq \Phi(x_1) \leq nmc_{\text{max}}$.

Since $\Phi(x)$ is $\frac{1}{\lambda}$-smooth the function $H(x) = \Phi(x) + \frac{1}{\lambda} \| x - x^t \|^2$ is convex. Now the definition $y^{t+1} = \arg \min_{y \in \mathcal{X}^{t+1}} H(y)$ that implies $\langle \nabla H(y^{t+1}), x - y^{t+1} \rangle \geq 0$ for all $x \in \mathcal{X}^{t+1}$. The latter holds also for all $x \in \mathcal{X}^{t+1}$ since $\mathcal{X}^{t+1} \subseteq \mathcal{X}^{t+1}$. At this point we get that,

$$
\Phi(x^t) - \Phi(y^{t+1}) - \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 = H(x^t) - H(y^{t+1}) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 \geq \nabla H(y^{t+1})^T (x^t - y^{t+1}) + \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2 \geq \frac{1}{2\lambda} \| y^{t+1} - x^t \|^2
$$

Therefore, plugging into (4), it holds that

$$
\frac{1}{T\lambda} \sum_{t=1}^{T} \gamma_t \mathbb{E} \left[ \| y^{t+1} - x^t \|^2 \right] \leq \frac{nmc_{\text{max}}}{T} + \frac{c_{\text{max}}^2 n m}{\lambda T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t}.
$$

22
Then, using the lower bound $\frac{1}{T} \sum_{t=1}^{T} \gamma_t E \| y^{t+1} - x^t \|^2 \geq \frac{2\gamma}{T} \sum_{t=1}^{T} E \| y^{t+1} - x^t \|^2$ on the left hand side, using Jensen’s inequality and taking square root on both sides we obtain

$$\frac{1}{T} \sum_{t=1}^{T} E \| y^{t+1} - x^t \| \leq \sqrt{\frac{\lambda^2 nmc_{\max}}{T \gamma_T} + \frac{\lambda c_{\max}^2 nm}{T \gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t}}. \tag{5}$$

Before concluding the proof we need to bound the difference between the elements in the sequence $y^t$ and the iterates of proximal point projected always on the final set $\mathcal{X}^{\mu_t}$. To this end, we introduce the sequence $\tilde{y}^{t+1} = \Pi_{\mathcal{X}^{\mu_T}} \left[ x^t - \frac{\lambda}{2} \nabla \Phi(\tilde{y}^{t+1}) \right]$ and we notice that

$$\| y^{t+1} - \tilde{y}^{t+1} \| = \| \Pi_{\mathcal{X}^{\mu_t}} \left[ x^t - \frac{\lambda}{2} \nabla \Phi(y^{t+1}) \right] - \Pi_{\mathcal{X}^{\mu_T}} \left[ x^t - \frac{\lambda}{2} \nabla \Phi(\tilde{y}^{t+1}) \right] \|

At this point by defining $w^t := x^t - \frac{\lambda}{2} \nabla \Phi(y^{t+1})$ and $\tilde{w}^t := x^t - \frac{\lambda}{2} \nabla \Phi(\tilde{y}^{t+1})$ we get that

$$\| y^{t+1} - \tilde{y}^{t+1} \| \leq \| \Pi_{\mathcal{X}^{\mu_t}} [ w^t ] - \Pi_{\mathcal{X}^{\mu_t}} [ \tilde{w}^t ] \| + \| \Pi_{\mathcal{X}^{\mu_t}} [ w^t ] - \Pi_{\mathcal{X}^{\mu_T}} [ \tilde{w}^t ] \| + \| \Pi_{\mathcal{X}^{\mu_t}} [ w^t ] - \Pi_{\mathcal{X}^{\mu_t}} [ \tilde{w}^t ] \| \leq \sqrt{nm^3 \mu_t + \sqrt{nm^3 \mu_T} + \| w^t - \tilde{w}^t \|} \leq \sqrt{nm^3 \mu_t + \sqrt{nm^3 \mu_T} + \frac{\lambda}{2} \| \nabla \Phi(y^{t+1}) - \nabla \Phi(\tilde{y}^{t+1}) \|} \leq \sqrt{nm^3 \mu_t + \sqrt{nm^3 \mu_T} + \frac{1}{2} \| y^{t+1} - \tilde{y}^{t+1} \|}$$

where in the first inequality, we used the bound distance between a point in $\mathcal{X}$ and its projection in $\mathcal{X}^{\mu_t}$ used in the proof of Lemma 7 and in the last inequality we used the Lipschitz continuity of the gradients of the potential function. The above estimation implies

$$\| y^{t+1} - \tilde{y}^{t+1} \| \leq 2\sqrt{nm^3 (\mu_t + \mu_T)}. \tag{6}$$

Moreover, by a simple application of triangular inequality and the bounds in Equation (5) and in Equation (6), we obtain

$$\frac{1}{T} \sum_{t=1}^{T} E \| x^t - \tilde{y}^{t+1} \| \leq \frac{1}{T} \sum_{t=1}^{T} E \| y^{t+1} - \tilde{y}^{t+1} \| + \frac{1}{T} \sum_{t=1}^{T} E \| y^{t+1} - x^t \| \leq \frac{1}{T} \sum_{t=1}^{T} 2\sqrt{nm^3 (\mu_t + \mu_T)} + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T \gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T \gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t}} \leq \frac{4\sqrt{nm^3}}{T} \sum_{t=1}^{T} \mu_t + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T \gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T \gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t}} \leq \frac{\lambda}{2T} \sum_{t=1}^{T} E \| \nabla \Phi(x^t) - \nabla \Phi(\tilde{y}^{t+1}) \| + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T \gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T \gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t} + \frac{4\sqrt{nm^3}}{T} \sum_{t=1}^{T} \mu_t} \leq \frac{\lambda}{2T} \sum_{t=1}^{T} E \| \nabla \Phi(x^t) - \nabla \Phi(\tilde{y}^{t+1}) \| + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T \gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T \gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\mu_t} + \frac{4\sqrt{nm^3}}{T} \sum_{t=1}^{T} \mu_t}$$
Semi Bandit Dynamics in Congestion Games

\[
\begin{align*}
&\leq \lambda \frac{1}{2T} \sum_{t=1}^{T} \mathbb{E}\left[\|x^t - \hat{y}^{t+1}\|_2^2 + \sqrt{\frac{\lambda^2 nm_{\text{max}}}{2T\gamma_T} + \frac{\lambda c_{\text{max}}^2 nm}{2T\gamma_T}} \sum_{t=1}^{T} \gamma_t^2 \mu_t + \frac{4\sqrt{nm}}{T} \sum_{t=1}^{T} \mu_t}
\end{align*}
\]

that concludes the proof. \[\square\]

**E.4. Proof of Theorem 3.8**

**Theorem 3.8.** Let \(\pi^1, \ldots, \pi^T\) the sequence of strategy profiles produced if all agents adopt Algorithm 2. Then for all \(T \geq \Theta\left(m^{12.5}n^{7.5}/e^5\right)\),

\[
\frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [n]} \left[c_i(\pi_i^t, \pi_{-i}^t) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})\right]\right] \leq \epsilon.
\]

The same holds for \(T \geq \Theta(n^{6.5}m^7/e^5)\) in case the agents know \(n, m\) and select \(\gamma_t := \Theta(m^{-4/5}n^{-8/5}c_{\text{max}}^{-1}t^{-3/5})\) and \(\mu_t := \Theta(n^{-6/5}m^{-11/10}t^{-1/5})\).

**Proof.** Let \(x^t\) denote the marginalization of \(\pi^t\) then by applying Lemma 14 for \(\mu := \mu_T\) we get that

\[
\max_{i \in [n]} \left[c_i(\pi_i^t, \pi_{-i}^t) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})\right] \leq 4n^2mc_{\text{max}} \|G(x^t)\| + 2m^2nc_{\text{max}}\mu_T
\]

As a result,

\[
\frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [n]} \left[c_i(\pi_i^t, \pi_{-i}^t) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})\right]\right] \leq 4n^2mc_{\text{max}} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|G(x^t)\|\right] + 2m^2nc_{\text{max}}\mu_T
\]

where \(G(x) = \Pi_{\chi^{\mu_T}} [x - \lambda \nabla \Phi(x)] - x\). Then by Theorem 5.2 and the fact that \(\lambda = (2n^2c_{\text{max}}\sqrt{m})^{-1}\) we get

\[
\frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [n]} \left[c_i(\pi_i^t, \pi_{-i}^t) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})\right]\right] \leq 4\sqrt{m} \sqrt{\frac{nm_{\text{max}}}{2T\gamma_T} + \frac{c_{\text{max}}^2 nm}{2T\gamma_T}} \sum_{t=1}^{T} \gamma_t^2 \mu_t + \frac{16\sqrt{nm}}{T} \sum_{t=1}^{T} \mu_t + 2m^2nc_{\text{max}}\mu_T
\]

To simplify notation let

\[
(A) := \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [n]} \left[c_i(\pi_i^t, \pi_{-i}^t) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi_{-i})\right]\right]
\]

At this point by choosing the sequence \(\gamma_t = C_{\gamma} t^{-3/5}\) and \(\mu_t = C_{\mu} \min \left\{1/m, t^{-1/5}\right\}\) we have that

\[
(A) \leq 4\sqrt{m} \sqrt{\frac{nm_{\text{max}}}{2T^{2/5}C_{\gamma}} + \frac{c_{\text{max}}^2 nmC_{\gamma}}{2T^{2/5}C_{\mu}}} \sum_{t=1}^{T} \frac{1}{mt^{6/5}} + \frac{c_{\text{max}}^2 nmC_{\gamma}}{2T^{2/5}C_{\mu}} \sum_{t=1}^{T} \frac{t^{1/5}}{t^{6/5}}
\]

\[
+ 20\sqrt{\pi^2m^2C_{\mu}} + 2m^2nc_{\text{max}}C_{\mu} T^{-1/5}
\]

\[
\leq 4\sqrt{m} \sqrt{\frac{nm_{\text{max}}}{2T^{2/5}C_{\gamma}} + \frac{c_{\text{max}}^2 nmC_{\gamma}}{2T^{2/5}C_{\mu}}} \frac{\log T + 1}{T^{1/5}} + 20\sqrt{\pi^2m^2C_{\mu}} + 2m^2nc_{\text{max}}C_{\mu} T^{-1/5}
\]

\[
= \left(4\sqrt{m} \sqrt{\frac{nm_{\text{max}}}{2T^{2/5}C_{\gamma}} + \frac{c_{\text{max}}^2 nmC_{\gamma}}{2T^{2/5}C_{\mu}}} \frac{\log T + 1}{T^{1/5}} + 20\sqrt{\pi^2m^2C_{\mu}} + 2m^2nc_{\text{max}}C_{\mu}\right) T^{-1/5}
\]

24
By replacing the values of $\lambda = (2n^2c_{\max}\sqrt{m})^{-1}$, we obtain

\[
(A) \leq \left(4\sqrt{m} \frac{mnc_{\max}}{2C_\gamma} + \frac{c^3n^{3}m^{3/2}C_\gamma (\log T + 1)}{C_\mu} + 80m^{5/2}n^{5/2}c_{\max}C_\mu \right) T^{-1/5}
\]

By neglecting non-dominant terms and choosing $C_\gamma = (m^{4/5}n^{8/5}c_{\max})^{-1}$ and $C_\mu = (n^{6/5}m^{11/10})^{-1}$, we obtain:

\[
(A) \leq \left(4\sqrt{m} \frac{c^2 \max n^{13/5}m^{9/5}}{2} + \frac{2 \max n^{13/5}m^{9/5}(\log T + 1) + 80m^{14/10}n^{13/10}c_{\max}}{T^{1/5}} \right) T^{-1/5}
\]

The latter implies that if $T = O \left( m^{7n^{6.5}/\epsilon^5} \right)$ then $(A) \leq \epsilon$.

By the choosing $C_\gamma = C_\mu = 1$ we obtain

\[
(A) \leq O \left( \frac{\sqrt{m^{5}n^{3}}c_{\max}^{3}(\log T + 1)}{T^{1/5}} \right)
\]

which implies that if $T = O \left( m^{12.5n^{7.5}/\epsilon^5} \right)$ then $(A) \leq \epsilon$.

**Corollary 1.** In case all agents adopt Algorithm 2 for $T \geq \Theta \left( n^{6.5}m^{7}/\epsilon^{5} \right)$ (resp. $T \geq \Theta \left( m^{12.5n^{7.5}/\epsilon^{5}} \right)$) then with probability $\geq 1 - \delta$,

- $(1 - \delta)T$ of the strategy profiles $\pi^1, \ldots, \pi^T$ are $\epsilon/\delta^2$-approximate Mixed NE.
- $\pi^t$ is an $\epsilon/\delta$-approximate Mixed NE once $t$ is sampled uniformly at random in $\{1, \ldots, T\}$.

**Proof.** Let the random variable $E_t := \max_{i \in [n]} \left[ c_i(\pi^t, \pi^t_{-i}) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi^t_{-i}) \right]$. Consider the random variable $E$ taking the value of the random variable $E_t$ with $t$ being sampled uniformly at random. Notice that

\[
\mathbb{E}[E] = \frac{1}{T} \mathbb{E} \left[ \max_{i \in [n]} \left[ c_i(\pi^t_{-i}, \pi^t_{-i}) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi^t_{-i}) \right] \right] \leq \epsilon.
\]

which by Markov inequality implies that with probability $\geq 1 - \delta$, $E \leq \epsilon/\delta$. Thus with probability $\geq 1 - \delta$, $\max_{i \in [n]} \left[ c_i(\pi^t_{-i}, \pi^t_{-i}) - \min_{\pi_i \in \Delta(P_i)} c_i(\pi_i, \pi^t_{-i}) \right] \leq \epsilon/\delta$ meaning that $\pi^t$ is an $\epsilon/\delta$-Mixed NE. This establishes the second item of Corollary 1. Now consider the set of time steps $B := \{ t \in \{1, \ldots, T\} : E_t > \epsilon/\delta^2 \}$. With probability $1 - \delta$, $\sum_{t=1}^{T} E_t \leq \frac{\epsilon}{\delta^2}$ we directly get that we probability $1 - \delta$, $|B| \leq \delta T$. As a result, with probability $\geq 1 - \delta$, $(1 - \delta)$ fraction of the profiles $\pi^1, \ldots, \pi^T$ are $\epsilon/\delta^2$-Mixed NE.

**F. Auxiliary Lemmas**

**Lemma 15.** Let the set $A$ be defined as in Algorithm 1 then $\forall e \in A$, given $x \in X^n$ such that $x_e > 0$, there exists a simple path $\hat{p}$ such that

- (i) $e \in \hat{p}$,
- (ii) $x_e > 0 \ \forall e \in \hat{p}$.

Therefore, Step 6 in Algorithm 1 can always be implemented.

**Proof.** By the Carathéodory’s theorem, there exists a collection of simple paths $\{\hat{p}_1, \ldots, \hat{p}_{m+1}\}$ and scalars $\lambda_1, \ldots, \lambda_{m+1}$ such that $\lambda_i \geq 0 \ \forall i \in \{1, m + 1\}$ and $\sum_{i=1}^{m+1} \lambda_i = 1$ that allows to write $x = \sum_{j=1}^{m+1} \lambda_j \hat{p}_j$. At this point, assume by contradiction that $\hat{p}_{j_e} = 0$ for all $j \in \{1, m + 1\}$. This implies that $x_e = 0$ which is a clear contradiction. That means that there exist $j^* \in \{1, m + 1\}$ such that $e \in \hat{p}_{j^*}$ proving part (i). In addition it must be true that the weight in the convex combination is positive, i.e. $\lambda_{j^*} > 0$. This implies that $x_e \geq \lambda_{j^*} \hat{p}_{j^*} e$. Therefore, for the edges $e$ s.t. $\hat{p}_{j^*} e = 1$, it holds that $x_e \geq \lambda_{j^*} > 0$. 

\[\square\]
Lemma 16. Let $x, \bar{x} \in \mathcal{X}$ and $\pi, \bar{\pi}$ be their respective Caratheodory decompositions. Then,

$$c_i (\bar{\pi}_i, \pi_{-i}) - c_i (\pi_i, \pi_{-i}) = \Phi(\bar{x}_i, x_{-i}) - \Phi(x_i, x_{-i})$$

Proof. We start by manipulating the cost difference

$$c_i (\bar{\pi}_i, \pi_{-i}) - c_i (\pi_i, \pi_{-i}) = \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathbb{P} \left( [p_{-i}] \right) \left( \mathbb{P}_{\pi_i} \left[ p_i \right] C_i(p_i, p_{-i}) - \mathbb{P}_{\bar{\pi}_i} \left[ p_i \right] C_i(p_i, p_{-i}) \right)$$

$$= \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathbb{P} \left( [p_{-i}] \right) \mathbb{P}_{\pi_i} \left[ p_i \right] \mathbb{P}_{\pi_i} \left[ p_i \right] \left( C_i(p_i, p_{-i}) - C_i(p_i, p_{-i}) \right)$$

$$= \sum_{e \in \mathcal{E}_{p_i \in \mathcal{P}_i, p_{-i} \in \mathcal{P}_{-i}}} \mathbb{P} \left( [p_{-i}] \right) \mathbb{P}_{\pi_i} \left[ p_i \right] \left( \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) - \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) \right)$$

At this point we can consider the two terms of the last expression can be written as the potential function in Definition 2

$$\sum_{e \in \mathcal{E}_{p_i \in \mathcal{P}_i, p_{-i} \in \mathcal{P}_{-i}}} \mathbb{P} \left( [p_{-i}] \right) \mathbb{P}_{\pi_i} \left[ p_i \right] \left( \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) - \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) \right)$$

$$= \sum_{e \in \mathcal{E}_{p_i \in \mathcal{P}_i, p_{-i} \in \mathcal{P}_{-i}}} \mathbb{P} \left( [p_{-i}] \right) \mathbb{P}_{\pi_i} \left[ p_i \right] \left( \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) \right)$$

$$- \sum_{e \in \mathcal{E}_{p_i \in \mathcal{P}_i, p_{-i} \in \mathcal{P}_{-i}}} \mathbb{P} \left( [p_{-i}] \right) \mathbb{P}_{\pi_i} \left[ p_i \right] \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i)$$

G. On the difference with ε-greedy exploration

We present two examples to highlight the difference between ε-greedy and Exploration with Bounded-Away Polytopes for the special case of simplex. Exploration with Bounded-Away Polytopes takes a mixed strategy $x \in \Delta_n$ and transforms it to the strategy

$$x' := \Pi_{\Delta_n^\mu}(x)$$

where $\Delta_n^\mu = \Delta_n \cap \{ x : x_i \geq \mu \}$. On the other hand ε-greedy exploration transforms $x$ to $x'$ as follows,

$$x' := (1 - \epsilon)x + \frac{\epsilon}{n} (1, \ldots, 1)$$

These are two different transformation that do not coincide. We will provide two specific examples: one example for $x \in \Delta_n^\mu$ and one for $x \notin \Delta_n^\mu$. 

\[26\]
Example for $x \in \Delta_n^\mu$: Consider $\epsilon = \mu$ and $x = (2/3, 1/3)$. In this case $\epsilon$-greedy exploration selects the strategy $x' = ((1 - \mu)2/3 + \mu/2, (1 - \mu)1/3 + \mu/2)$ while Exploration with Bounded-Away Polytopes selects the strategy $x' = (2/3, 1/3)$ because $x \in \Delta_n^\mu$ implies that $x = x'$.

Example for $x \notin \Delta_n^\mu$: Consider $\epsilon = \mu$ and $x = (8/10, 2/10, 0)$. In this case $\epsilon$-greedy exploration selects the strategy $x' = ((1 - \mu)8/10 + \mu/3, (1 - \mu)2/10 + \mu/3, \mu/3)$. For Exploration with Bounded-Away Polytopes with $\mu \leq 0.4$ we can use the KKT conditions of the problem to derive that $x' = (x'_1, x'_2, x'_3)$ must satisfy the following system

\[
\begin{align*}
2(x'_1 - 0.8) + \lambda &= 0 \\
2(x'_2 - 0.2) + \lambda &= 0 \\
2(x'_3 - 0.0) + \lambda &\geq 0 \\
-x'_1 + \mu &< 0 \\
-x'_2 + \mu &< 0 \\
-x'_3 + \mu &= 0
\end{align*}
\]  

which admits the solution $x' = (0.8 - \mu/2, 0.2 - \mu/2, \mu)$ which is a different transformation than the one obtained with $\epsilon$-greedy. To see this, take for example $\epsilon = \mu = 0.12$. $\epsilon$ greedy exploration gives $((1 - \mu)8/10 + \mu/3, (1 - \mu)2/10 + \mu/3, \mu/3)) = (0.744, 0.216, 0.04)$ while Exploration with Bounded-Away Polytopes gives $(0.8 - \mu/2, 0.2 - \mu/2, \mu) = (0.74, 0.14, 0.12)$.