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 $\mathcal{O}\left(2^{\Theta(m)}\right)$ 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 $\mathcal{O}(n^{12}m^9/\epsilon^6)$ time-steps (samples) to reach an ϵ-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*

[*]Equal contribution [1]Donald Bren School of Information and Computer Science, University of California Irvine, Irvine, California, USA [2]Laboratory for Information and Inference Systems, IEM STI, École Polytechnique Fédérale de Lausanne, Lausanne, Vaud, Switzerland [3]Istitute of Theoretical Computer Science, Shanghai University of Finance and Economics, Shanghai, China. Correspondence to: Stratis Skoulakis <efstratios.skoulakis@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

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 (Even-Dar et al., 2009).

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 Caratheodory Exploration ($\mathrm{SBGD} - \mathrm{CE}$), with the following properties:

- ***No-regret guarantees***: Any agent adopting SBGD-CE admits at most $\tilde{\mathcal{O}}(m^2 T^{4/5})$ regret no matter how the other agents select their strategies.
- ***Convergence to NE***: If $\mathrm{SBGD} - \mathrm{CE}$ is adopted by all agents, the overall system reaches an ϵ-approximate NE within $\mathcal{O}(n^{6.5} m^7/\epsilon^5)$ time steps improving the $\mathcal{O}(n^{12} m^9/\epsilon^6)$ 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 the is 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 ϵ (see Theorem 3.8). From a *game-theoretic* point of view, bset-iterate convergence implies that with high probability *almost all iterates* are $\mathcal{O}(\epsilon)$-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 $t \sim \mathrm{Unif}(1, \ldots, T)$ (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. $\mathcal{O}\left(2^{\Omega(m)}\sqrt{T}\right)$, 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; György 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 $\mathcal{O}(m\sqrt{T})$. 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 present 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 $\mathrm{SBGD} - \mathrm{CE}$ method is based

Table 1. Comparison with previous related works. *See Remark 3.7 for regret bound $\mathcal{O}(m^{3/2}T^{3/4})$ 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).

Method	**Adversarial Regret**	**Convergence† to NE**	**Running Time**
IPPG (Leonardos et al., 2022)	Not Established	$\mathcal{O}\left(n2^{\Theta(m)}m/\epsilon^6\right)$	Exp. in Available Resources
IPGA (Ding et al., 2022)	Not Established	$\mathcal{O}\left(n^3 2^{\Theta(m)}m^5/\epsilon^5\right)$	Exp. in Available Resources
FW with Exploration II (Cui et al., 2022)	Not Established	$\mathcal{O}\left(n^{12}m^9/\epsilon^6\right)$	Exp. in Available Resources
$\mathrm{SBGD}-\mathrm{CE}$ (this work)	$\mathcal{O}(m^2T^{4/5})^\star$	$\mathcal{O}\left(n^{6.5}m^7/\epsilon^5\right)$	Poly. in Implicit Description

into running Online Gradient Descent (Zinkevich, 2003) while projecting into a *time-expanding* Bounded-Away Description Polytope. At each round, $\mathrm{SBGD}-\mathrm{CE}$ also uses the Caratheodory Decomposition to (randomly) select a valid set of resources. By extending the analysis of Online Gradient Descent as well as of Stochastic Gradient Descent constrained to *time-varying feasibility sets*, we establish no-regret guarantees as well as fast converge to NE.

Further Related Work (Anagnostides et al., 2022) establish best-iterate convergence rates to NE in congestion games, though *full-information feedback* is assumed and the rates depends on the strategy space of each agent. (Cominetti et al., 2010) and (Palaiopanos et al., 2017; Héliou et al., 2017) prove *asymptotic last-iterate convergence* of no-regret dynamics under *full-information feedback* and *bandit-feedback* respectively in potential/congestion games. To the best of our knowledge there do not exist *last-iterate convergence rates* for congestion games (even with exponential dependence on the number of resources) unless the initial condition is close enough to an equilibrium. Even the well-studied *full-information better-response dynamics* is only known to converge in the *best-iterate sense* (Chien & Sinclair, 2007). (Cui et al., 2022) also provide provide convergence guarantees for congestion games under *bandit feedback* with slightly worse rates that the one presented in Table 1. (Vu et al., 2021) study accelerated methods to converging to *Wardrop Equilibrium* in network congestion games. Other works studying no-regret dynamics beyond congestion games include (Piliouras & Shamma, 2014; Mertikopoulos & Staudigl, 2017; Cohen et al., 2017; Mertikopoulos et al., 2018; Bravo et al., 2018; Mertikopoulos & Zhou, 2019; Vlatakis-Gkaragkounis et al., 2020; Giannou et al., 2021).

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 $p := (p_i, p_{-i})$ 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.,

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

Each resource e admits a positive and non-decreasing cost function $c_e : \mathbb{N} \mapsto \mathbb{R}_{\geq 0}$ 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

$$C_i(p) \triangleq \sum_{e \in p_i} c_e\left(\ell_e(p)\right).$$

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

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

A prob. distribution $\pi^\star := (\pi_1^\star, \ldots, \pi_n^\star) \in \Delta(\mathcal{P}_1) \times \cdots \times \Delta(\mathcal{P}_n)$ is an ϵ-approximate *Mixed Nash Equilibrium* (MNE) if and only if for each agent $i \in [n]$,

$$\mathbb{E}_{\pi_i^\star, \pi_{-i}^\star}\left[C_i(p_i, p_{-i})\right] \leq \mathbb{E}_{\pi_i', \pi_{-i}^\star}\left[C_i(p_i, p_{-i})\right] + \epsilon.$$

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

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

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

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

Consider the set $\hat{\mathcal{P}}_i := \{x_{p_i} \in \{0,1\}^m \ : \ \text{for } p_i \in \mathcal{P}_i\}$ and the polytope $\mathcal{X}_i := \text{conv}(\hat{\mathcal{P}}_i)$ denoting the convex hull of $\hat{\mathcal{P}}_i$. 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 *implictly* 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 $\mathcal{O}(rm)$ 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),

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

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 $A_i \in \mathbb{R}^{r_i \times m}$ and $b_i \in \mathbb{R}^{r_i}$. We note that the regret bounds and the convergence rates to NE of our proposed algorithm (Algorithm 2) only depend on the n 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 (Cui 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 (Cui 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

```
for each round $t = 1, \dots, T$ do
    Each agent $i \in [n]$ (randomly) selects a strategy $p_i^t \in \mathcal{P}_i$ and suffers cost
        $C_i(p_i^t, p_{-i}^t) := \sum_{e \in p_i^t} c_e^t(p_i^t, p_{-i}^t)$
    Each agent $i \in [n]$ learns only the congestion costs $c_e(p_i^t, p_{-i}^t)$ of her selected resources $e \in p_i^t$.
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^t 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

```
for each round $t = 1, \dots, T$ do
    Agent $i$ selects a prob. distribution $\pi_i^t \in \Delta(\mathcal{P}_i)$.
    An adversary selects a cost function $c^t : E \mapsto \mathbb{R}_{\geq 0}$.
    Agent $i$ samples a path $p_i^t \sim \pi_i^t$ and suffers cost
        $C_i(p_i^t, c^t) := \sum_{e \in p_i^t} c_e^t.$
    Agent $i$ learns the costs $c_e^t$ for all resources $e \in p_i^t$ and updates $\pi_i^{t+1} \in \Delta(\mathcal{P}_i)$.
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 $\mathcal{R}_{\mathcal{A}}(T) :=$

$$\max_{c^1, \dots, c^T} \left[\sum_{t=1}^{T} \mathbb{E}_{\pi_i^t} \left[C_i(p_i^t, c^t) \right] - \min_{p_i^\star \in \mathcal{P}} \sum_{t=1}^{T} C_i(p_i^\star, c^t) \right]$$

In case $\mathcal{R}_{\mathcal{A}}(T) = o(T)$, 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 $\mathcal{R}_{\mathcal{A}}(T) = \mathcal{O}(m^2 T^{4/5})$ regret for Online Resource Selection (Algorithm 2). Moreover if Algorithm 2 is adopted by all agents, then the agents converge to an ϵ-Mixed NE after $\mathcal{O}(n^{6.5} m^7/\epsilon^5)$ 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_e^t of the edges which represents the probability edge e is selected at round t, i.e., $x_e^t = \mathbb{P}\left[e \in p^t\right]$. 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 $E_i := \{e \in E : \ e \in p_i \text{ for some } p_i \in \mathcal{P}_i\}$.

Lemma 1. *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.*

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

$$\mathcal{X}_i^{\mu} \triangleq \left\{ x \in \mathcal{X}_i : \ x_e \geq \mu \quad \forall e \in E_i \right\}$$

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

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

Theorem 3.3 (Carathéodory Decomposition)**.** *For any point $x \in \mathcal{X}_i$ 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 $e \in E$,*

$$x_e = \sum_{p_i : e \in p_i} \mathrm{Pr}_{\pi_x}\left[p_i \textit{ is selected}\right].$$

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

A Carathéodory decomposition of a point $x \in \mathcal{X}_i^{\mu}$ 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 $\mathcal{X} := \{x \in \mathbb{R}^m : \ A \cdot x \leq b\}$ with $A \in \mathbb{R}^{r \times m}$ and $b \in \mathbb{R}^r$. The Carathéodory Decomposition of any point $x \in \mathcal{X}$ can be computed in polynomial-time with respect to r and m.*

For the important special case of *path polytopes* described in Section 2, a point $x \in \mathcal{X}_i^{\mu}$ 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 Charatheodory decomposition for Path Polytopes

```
**Input:** A point $x \in \mathcal{X}_i^{\mu}$
Res $\leftarrow \varnothing$
**while** $\sum_{e \in \mathrm{Out}(s_i)} x_e > 0$ **do**
    Let $A \leftarrow E \cap \{e \in E : \ x_e > 0\}$
    Let $e_{\min} \leftarrow \arg\min_{e \in A} x_e$ and $x_{\min} \leftarrow x_{e_{\min}}$.
    $\hat{p}_i \leftarrow$ An $(s_i, t_i)$-path of $G(V, A)$ with $e_{\min} \in \hat{p}_i$.
    $x_e \leftarrow x_e - x_{\min}$ for all $e \in \hat{p}_i$.
    Res $\leftarrow$ Res $\cup \{(\hat{p}_i, x_{\min})\}$.
**end while**
**return** Res
```

Lemma 3. *Algorithm 1 requires $\mathcal{O}\left(|V||E| + |E|^2\right)$ steps to give a Caratheodory 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 $x_i^1 \in \mathcal{X}_i^{\mu_1}$.
3: for each round $t = 1, \dots, T$ do
4:     Agent $i$ computes a Caratheodory decomposition $\pi_i^t \in \Delta(\mathcal{P}_i)$ for $x_i^t \in \mathcal{X}_i^{\mu_t}$.
5:     Agent $i$ samples a strategy $p_i^t \sim \pi_i^t$ and suffers cost,
           $C_i^t(p_i^t, c^t) := \sum_{e \in p_i^t} c_e^t.$
       /* $c_e^t$ is the cost of edge $e$ with load the number of agents that chose $e$ at time $t$. */
6:     Agent $i$ sets $\hat{c}_e^t \leftarrow c_e^t \cdot \mathbb{1}\left[e \in p_i^t\right] / x_e^t$ for all $e \in E$.
7:     Agent $i$ updates $x_i^{t+1} \in \mathcal{X}_i^{\mu}$ as,
           $x_i^{t+1} = \Pi_{\mathcal{X}_i^{\mu_{t+1}}}\left[x_i^t - \gamma_t \cdot \hat{c}^t\right],$
       where $\gamma_t \leftarrow t^{-3/5}$ and $\mu_t \leftarrow \min(1/m_i, t^{-1/5})$.
8: end for
```

At Step 4, OGD $-$ CE performs a Caratheodory Decomposition to convert the fractional point $x_i^t \in \mathcal{X}_i^{\mu_t}$ into a probability distribution π_i^t over pure strategies $p_i \in \mathcal{P}_i$. The latter guarantees that the experienced cost equals the fractional congestion cost, i.e.,

$$\mathbb{E}\left[\sum_{e \in p^t} c_e^t \mid x_i^t\right] = \left\langle c^t, x_i^t \right\rangle. \tag{2}$$

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

Lemma 4. *The estimator $\hat{c}_e^t = c_e^t \cdot \mathbb{1}\left[e \in p^t\right] / x_e^t$ satisfies*

1. $\mathbb{E}\left[\hat{c}_e^t\right] = c_e^t$ *for $e \in E_i$ (Unbiasness).*
2. $|\hat{c}_e^t| \leq c_{max}/\mu_t$ *for $e \in E_i$ (Boundness).*

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

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 $p_i^1, \dots, p_i^T \in \mathcal{P}_i$ the sequence of strategies produced by Algorithm 2 given as input the costs $c^1, \dots, c^T$ with $\|c^t\|_\infty \leq c_{\max}$. Then with probability $1 - \delta$,*

$$\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t - \min_{p_i^\star \in \mathcal{P}_i} \sum_{e \in p_i^\star} c_e^t \leq \mathcal{O}\left(m c_{\max}^2 T^{4/5} \log(1/\delta)\right)$$

Remark 3.7. Setting $\delta := \mathcal{O}\left(1/mTc_{\max}\right)$ directly implies that Algorithm 2 admits regret $\tilde{\mathcal{O}}\left(m^2 c_{\max}^2 T^{4/5}\right)$ regret. The better $\mathcal{O}(m^{3/2}T^{3/4})$ regret bound can be obtained by selecting the parameters $\gamma_t = c_{\max}^{-1} m^{-1/2} t^{-3/4}$ and $\mu_t = \min(1/m_i, m^{-1/2}t^{-1/4})$. However such a parameter selection leads to $\mathcal{O}(T^{-1/8})$ 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 $\pi^1, \dots, \pi^T$ the sequence of strategy profiles produced if all agents adopt Algorithm 2. Then for all $T \geq \Theta\left(m^{12.5} n^{7.5}/\epsilon^5\right)$,*

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

The same holds for $T \geq \Theta(n^{6.5}m^7/\epsilon^5)$ in case the agents know n, m and select $\gamma_t := \Theta(m^{-4/5}n^{-8/5}c_{\max}^{-1}t^{-3/5})$ 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. $\Theta\left(m^{12.5}n^{7.5}/\epsilon^5\right)$) then with probability $\geq 1 - \delta$,*

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

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 $|\hat{c}_e^t| \leq c_{\max}/\mu_t$ we establish the following concentration result for the quantity $\sum_{t=1}^{T}\left[\sum_{e \in p^t} c_e^t - \langle c^t, x_i^t \rangle\right]$.

Lemma 5. *Let the sequences $x_i^1, \dots, x_i^T \in \mathcal{X}_i$ and $p_i^1, \dots, p_i^T \in \mathcal{P}_i$ produced by Algorithm 2. Then, with probabilty $1 - \delta/2$,*

$$\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t = \sum_{t=1}^{T} \left\langle c^t, x_i^t \right\rangle + \mathcal{O}\left(\sqrt{c_{\max}} \log(m/\delta)\sqrt{T}\right).$$

Let $p_i^* \in \mathcal{P}_i$ denote the optimal strategy for the sequence of costs $c^1, \dots, c^T$ and $x_i^\star \in \mathcal{X}_i$ the corresponding $\{0,1\}^m$ extreme point of $\mathcal{X}_i$. Then Lemma 5 implies that with probability $1 - \delta/2$,

$$\sum_{t=1}^{T} \left(\sum_{e \in p_i^t} c_e^t - \sum_{e \in p_i^\star} c_e^t \right) = \sum_{t=1}^{T} \left\langle c^t, x_i^t - x_i^\star \right\rangle + \tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

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

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

Formally, we have the following definition.

Definition 4.1. Let $p_i^\star \in \mathcal{P}_i$ the optimal strategy for the sequence $c^1, \dots, c^T$ and $x_i^\star \in \mathcal{X}_i$ its corresponding extreme point in $\mathcal{X}_i$. Moreover, consider constructing a collection of strategies $\mathcal{D} = \{\tilde{p}^\ell\}_{\ell=1}^m$ sampled as follows. For any active resource $e \in E_i$, add to $\mathcal{D}$ a strategy $\tilde{p} \in \mathcal{P}_i$ such that $e \in \tilde{p}$. Considering the collected strategies in a vector form, i.e. elements of $\{0,1\}^m$, we define

$$x_{\mu_t}^\star \triangleq (1 - m\mu_t)x_i^\star + \mu_t \sum_{\ell=1}^{m} \tilde{p}^\ell.$$

Remark 4.2. To see that $x_{\mu_t}^\star \in \mathcal{X}^{\mu_t}$, denote as s the vector $s = \frac{1}{m}\sum_{i=1}^m \tilde{p}^\ell$. Since by construction, for any $e \in E_i$ there exists $\ell \in [m]$ such that $e \in \tilde{p}^\ell$, we have that $s \geq 1/m$. Moreover, $s \in \mathcal{X}$ which implies $s \in \mathcal{X}^{1/m}$. At this point, we can write $x_{\mu_t}^\star = (1 - m\mu_t)x_i^\star + m\mu_t s$. Then, it is evident that $s \geq 1/m$ implies $x_{\mu_t}^\star \geq \mu_t$ and that $x_{\mu_t}^\star$ is a convex combination of $x_i^\star$ and s because $\mu_t \leq \frac{1}{m}$ and therefore that $x^\star \in \mathcal{X}$. These two facts allow to conclude that $x_{\mu_t}^\star \in \mathcal{X}^{\mu_t}$.

Up next, we decompose the right-hand term of Equation 4 and separately bound each of the $(\mathrm{A} - \mathrm{C})$ terms.

$$\sum_{t=1}^{T} \left\langle c^t, x_i^t - x^\star \right\rangle = \underbrace{\sum_{t=1}^{T} \left\langle \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle}_{(\mathrm{A})} + \underbrace{\sum_{t=1}^{T} \left\langle c^t, x_{\mu_t}^\star - x^\star \right\rangle}_{(\mathrm{B})}$$
$$+ \underbrace{\sum_{t=1}^{T} \left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle}_{(\mathrm{C})}.$$

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

Lemma 6. *Let the sequences $x_i^1, \dots, x_i^T \in \mathcal{X}_i$ and $\hat{c}^1, \dots, \hat{c}^T$ produced by Algorithm 2. Then,*

$$\sum_{t=1}^{T} \left\langle \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \leq \frac{2m}{\gamma_T} + m c_{\max}^2 \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2}.$$

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 $\mathcal{X}_i^{\mu_t}$ while in the analysis of (Zinkevich, 2003) the feasibility set is invariant.

The term (B) quantifies the suboptimality of the projections of $x_i^\star$ on the time expanding polytopes $\mathcal{X}_i^{\mu_t}$. Notice that $x_i^\star$ is possibly outside the $\mathcal{X}_i^{\mu_t}$ where Algorithm 2 projects to.

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

$$\sum_{t=1}^{T} \left\langle c^t, x_{\mu_t}^\star - x_i^\star \right\rangle \leq m^2 c_{\max} \sum_{t=1}^{T} \mu_t.$$

where $x_i^\star$ and $x_{\mu_t}^\star$ 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 $\hat{c}_t$.

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

$$\sum_{t=1}^{T} \left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \leq \frac{m c_{\max}}{\mu_T} \sqrt{T \log(2m/\delta)}.$$

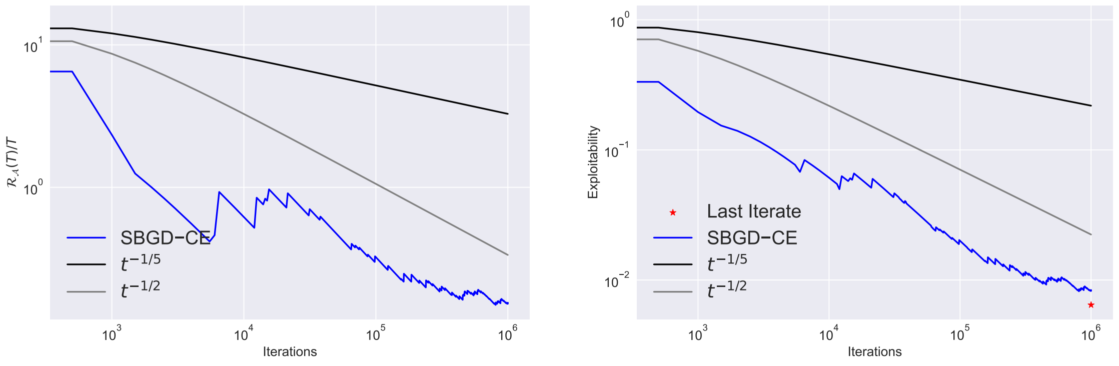
5. Sketch of Proof of Theorem 3.8

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 $\Phi : \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \to \mathbb{R}$

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

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 $\Phi(x)$.



(a) $\frac{\mathcal{R}_\mathcal{A}(T)}{T}$

(b) Exploitability

Figure 1. Regret and Exploitability on network games with 2 agents.

Definition 3. A point $x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ is called an (ϵ, μ)-*stationary point* $\Phi(x)$ if and only if

$$\left\| x - \Pi_{\mathcal{X}^\mu}\left[x - \frac{1}{2n^2 c_{\max}\sqrt{m}} \nabla\Phi(x) \right] \right\| \leq \epsilon$$

where $\mathcal{X}^\mu \triangleq \mathcal{X}_1^\mu \times \cdots \times \mathcal{X}_n^\mu$.

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,*

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

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 $\lambda \triangleq (2n^2 c_{\max}\sqrt{m})^{-1}$

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 $e \in E$,*

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

In case x is an (ϵ, μ)-stationary point of $\Phi(x)$ then π is a $(4n^2 m c_{\max}\epsilon + 2m^2 n c_{\max}\mu)$-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* $\mathcal{X}^{\mu_t} := \mathcal{X}_1^{\mu_t} \times \ldots \times \mathcal{X}_n^{\mu_t}$.

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

$$x^{t+1} = \Pi_{\mathcal{X}^{\mu_{t+1}}}\left[x^t - \gamma_t \cdot \nabla_t\right] \tag{3}$$

where the estimator $\nabla_t \triangleq [\hat{c}_1^t, \ldots, \hat{c}_n^t]$ ($\hat{c}_i^t$ 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.* $\mathbb{E}\left[\|\nabla_t\|^2\right] \leq \frac{nc_{\max}^2 m}{\mu_t}$.

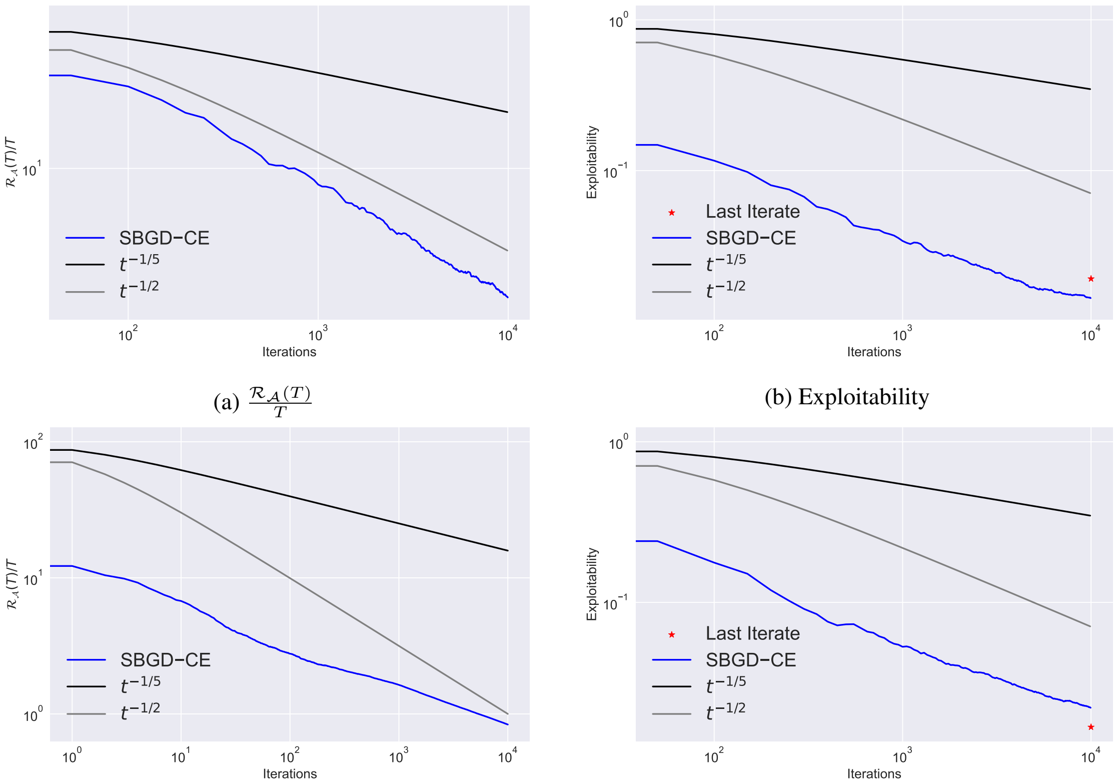
The main technical contribution of this section, is to establish that the sequence $x^1, \ldots, x^T$ produced by Equation 3 converges to an (ϵ, μ)-stationary point of Definition 3. The major challenge in proving the latter comes from the fact that in Equation 3 the projection step is respect to the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

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

$$2\sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T} + \frac{\lambda c_{\max}^2 nm \sum_{t=1}^T \frac{\gamma_t^2}{\mu_t}}{2T\gamma_T} + \frac{8\sqrt{nm^3}}{T}\sum_{t=1}^T \mu_t}.$$

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



(a) $\frac{\mathcal{R}_\mathcal{A}(T)}{T}$

(b) Exploitability

(c) $\frac{\mathcal{R}_\mathcal{A}(T)}{T}$

(d) Exploitability

Figure 2. Experiments on network games with 20 nodes for 20 (Figure 2a and 2b) and 5 agents (Figure 2c and 2d). Curves averaged over 10 seeds for the 20 agents case and 50 seeds for 5 agents.

$\{v_i\}_i^{|V|}$ (with $|V| = 19$) where every node v_i is connected only to v_{i+1} by 2 edges. Under this setting, Frank-Wolfe with Exploration (Cui et al., 2022) can not be implemented efficiently since there are 2^{19} possible paths. The same holds for (Leonardos et al., 2022; Ding et al., 2022). In order to verify empirically verify the convergence to NE, we monitor the exploitability of a strategy profile (π_i, π_{-i}) defined as $\max_{i\in[n]} \frac{c_i(\pi_i,\pi_{-i}) - \min_{\pi_i'} c_i(\pi_i',\pi_{-i})}{c_i(\pi_i',\pi_{-i})}$ which is 0 for any NE. Figure 1b shows the exploitability of the average path chosen by SBGD-CE. We notice it decreases at a rate $\approx t^{-1/2}$ which is better the theoretical bound provided in Theorem 3.8. Furthermore, the red star in Figure 1b represents the exploitability of the last iterate produced by Algorithm 2, it can be seen that it also achieves a small value of exploitability. We also verify the no-regret property of the algorithm in Figure 1a. Experiments with 5 and 20 agents are provided in Figure 2. The code is available at `https://github.com/lviano/SBGD-CE`.

7. Conclusion

This work introduces SBGD-CE which is 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*. As a result, our work answers an open question of (Cui et al., 2022) and improves upon their rates and complexity. 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

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2019. URL `https://arxiv.org/abs/1908.00261`.

Anagnostides, I., Panageas, I., Farina, G., and Sandholm, T. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 536–581. PMLR, 2022.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1): 31–45, feb 2014.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Awerbuch, B. and 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*, STOC '04, pp. 45–53, 2004.

Babichenko, Y. and Rubinstein, A. Communication complexity of nash equilibrium in potential games (extended abstract). In Irani, S. (ed.), *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pp. 1439–1445. IEEE, 2020.

Babichenko, Y. and Rubinstein, A. Settling the complexity of nash equilibrium in congestion games. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1426–1437, 2021.

Bravo, M., Leslie, D. S., and Mertikopoulos, P. Bandit learning in concave n-person games. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5666–5676, 2018.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In Mannor, S., Srebro, N., and Williamson, R. C. (eds.), *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, volume 23 of *JMLR Proceedings*, pp. 41.1–41.14. JMLR.org, 2012.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *J. Comput. Syst. Sci.*, 78(5):1404–1422, 2012.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pp. 1216–1259. PMLR, 2021.

Chien, S. and Sinclair, A. Convergence to approximate nash equilibria in congestion games. In Bansal, N., Pruhs, K., and Stein, C. (eds.), *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pp. 169–178. SIAM, 2007.

Christodoulou, G. and Koutsoupias, E. The price of anarchy of finite congestion games. *STOC*, pp. 67–73, 2005.

Cohen, J., Héliou, A., and Mertikopoulos, P. Hedging under uncertainty: Regret minimization meets exponentially fast convergence. In Bilò, V. and Flammini, M. (eds.), *Algorithmic Game Theory - 10th International Symposium, SAGT 2017, L'Aquila, Italy, September 12-14, 2017, Proceedings*, volume 10504 of *Lecture Notes in Computer Science*, pp. 252–263. Springer, 2017.

Cominetti, R., Melo, E., and Sorin, S. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83, 2010. Special Issue In Honor of Ehud Kalai.

Cui, Q., Xiong, Z., Fazel, M., and Du, S. S. Learning in congestion games with bandit feedback, 2022.

Dani, V., Hayes, T. P., and Kakade, S. M. The price of bandit information for online optimization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pp. 345–352, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

de Keijzer, B., Schäfer, G., and Telelis, O. A. On the inefficiency of equilibria in linear bottleneck congestion games. In Kontogiannis, S., Koutsoupias, E., and Spirakis, P. (eds.), *Algorithmic Game Theory*, volume 6386 of *Lecture Notes in Computer Science*, pp. 335–346. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-16169-8. doi: 10.1007/978-3-642-16170-4_29. URL `http://dx.doi.org/10.1007/978-3-642-16170-4_29`.

Ding, D., Wei, C., Zhang, K., and Jovanovic, M. R. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML*

2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, pp. 5166–5220. PMLR, 2022.

Even-Dar, E., Mansour, Y., and Nadav, U. On the convergence of regret minimization dynamics in concave games. In Mitzenmacher, M. (ed.), *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pp. 523–532. ACM, 2009.

Fabrikant, A., Papadimitriou, C., and Talwar, K. The complexity of pure Nash equilibria. In *ACM Symposium on Theory of Computing (STOC)*, pp. 604–612. ACM, 2004.

Fotakis, D., Kontogiannis, S., and Spirakis, P. Selfish unsplittable flows. *Theoretical Computer Science*, 348(2–3):226–239, 2005. ISSN 0304-3975. doi: http://dx.doi.org/10.1016/j.tcs.2005.09.024. URL `http://www.sciencedirect.com/science/article/pii/S0304397505005347`. Automata, Languages and Programming: Algorithms and Complexity (ICALP-A 2004)Automata, Languages and Programming: Algorithms and Complexity 2004.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

Giannou, A., Vlatakis-Gkaragkounis, E., and Mertikopoulos, P. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22655–22666, 2021.

Grötschel, M., Lovász, L., and Schrijver, A. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, 1988.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403, 2007.

Héliou, A., Cohen, J., and Mertikopoulos, P. Learning with bandit feedback in potential games. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6369–6378, 2017.

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2004.10.016. URL `https://www.sciencedirect.com/science/article/pii/S0022000004001394`. Learning Theory 2003.

Kleer, P. and Schäfer, G. Computation and efficiency of potential function minimizers of combinatorial congestion games. *Math. Program.*, 190(1):523–560, 2021.

Koutsoupias, E. and Papadimitriou, C. H. Worst-case equilibria. In *STACS*, pp. 404–413, 1999.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=gfwON7rAm4`.

Mertikopoulos, P. and Staudigl, M. Convergence to nash equilibrium in continuous games with noisy first-order feedback. In *56th IEEE Annual Conference on Decision and Control, CDC 2017, Melbourne, Australia, December 12-15, 2017*, pp. 5609–5614. IEEE, 2017.

Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Math. Program.*, 173(1-2):465–507, 2019.

Mertikopoulos, P., Papadimitriou, C. H., and Piliouras, G. Cycles in adversarial regularized learning. In Czumaj, A. (ed.), *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pp. 2703–2717. SIAM, 2018.

Monderer, D. and Shapley, L. S. Potential games. *Games and Economic Behavior*, pp. 124–143, 1996.

Neu, G. and Bartók, G. An efficient algorithm for learning with semi-bandit feedback. In Jain, S., Munos, R., Stephan, F., and Zeugmann, T. (eds.), *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, volume 8139 of *Lecture Notes in Computer Science*, pp. 234–248. Springer, 2013.

Palaiopanos, G., Panageas, I., and Piliouras, G. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5872–5882, 2017.

Piliouras, G. and Shamma, J. S. Optimization despite chaos: Convex relaxations to complex limit sets via poincaré recurrence. In Chekuri, C. (ed.), *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pp. 861–873. SIAM, 2014.

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

Roughgarden, T. Intrinsic robustness of the price of anarchy. In *Proc. of STOC*, pp. 513–522, 2009.

Roughgarden, T. and Tardos, É. How bad is selfish routing? *Journal of the ACM (JACM)*, 49(2):236–259, 2002.

Viossat, Y. and Zapechelnyuk, A. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013. ISSN 0022-0531.

Vlatakis-Gkaragkounis, E., Flokas, L., Lianeas, T., Mertikopoulos, P., and Piliouras, G. No-regret learning and mixed nash equilibria: They do not mix. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Vu, D. Q., Antonakopoulos, K., and Mertikopoulos, P. Fast routing under uncertainty: Adaptive learning in congestion games via exponential weights. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14708–14720, 2021.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In Fawcett, T. and 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.

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, (typically) 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[1]. 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), i.e., 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 $(s_, t_i)$-*path polytope* is defined as follows,

$$\mathcal{X}_i \triangleq \Bigg\{ x \in \{0,1\}^m : \sum_{e \in \mathrm{Out}(s_i)} x_e = 1 \\ \sum_{e \in \mathrm{In}(v)} x_e = \sum_{e \in \mathrm{Out}(v)} x_e \quad \forall v \in V \setminus \{s_i, t_i\} \\ \sum_{e \in \mathrm{In}(t_i)} x_e = 1 \Bigg\}$$

Lemma 11. *The extreme points of the (s_i, t_i)-path polytope $\mathcal{X}_i$ 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 $\mathcal{X}_i$. Given an (s_i, t_i)-path $p \in \mathcal{P}_i$ consider the point x_{p_i} of the polytope with $x_e^{p_i} = 1$ for all $e \in p_i$ and $x_e^{p_i} = 0$ otherwise. Let assume that x_{p_i} is not an extreme point of $\mathcal{X}_i$. 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 $\mathcal{X}_i$.

On the opposite direction we show that for any extreme point $x \in \mathcal{X}_i$ 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 $x_{\min} := \min_{e : x_e > 0} x_e$. Consider an (s_i, t_i)-path containing edge $e_{\min} := \arg\min_{e : x_e > 0} x_e$. Notice that for the point $x \in \mathcal{X}_0^i$ the following holds,

$$x = x_{\min} \cdot p_{\min} + (1 - x_{\min}) \cdot \underbrace{\frac{x - p_{\min}}{1 - x_{\min}}}_{y}$$

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

[1] Note 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 $\mathcal{X}_i^e = \{x \in \mathcal{X}_i : x_e = 1\}$ and we check whether it is empty. Notice that $\mathcal{X}_i^e$ 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 $x \in \mathcal{X}_i^e$ 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, $\mathcal{X}_i^e$ 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 $x_{p_i} \in \mathcal{X}_i^e$. □

C.2. Proof of Lemma 2

Lemma 2. *The set $\mathcal{X}_i^\mu$ 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 $p_i^e \in \mathcal{P}_i$ and update y as $y \leftarrow y + \mu \cdot x_{p_i^e}$. The consider $x := y/|E_i|$. Notice that $y \in \mathcal{X}_i$ as convex combination $x_{p_i^e} \in \mathcal{X}_i$. At the same time, $y_e \geq 1/|E_i|$ for each $e \in E_i$. Thus, the set $\mathcal{X}_i^\mu$ is not empty for any $\mu \leq 1/|E_i|$. □

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

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 $\mathcal{O}(|V| + |E|)$ operations. Therefore, every iteration of the outermost loop require $\mathcal{O}(|E| + |V|)$ 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 $\sum_{e \in \mathrm{Out}(s)} x_e = 0$. This conclude the proof because the total operation complexity is $\mathcal{O}(|E| \left(|E| + |V|\right))$ as stated in the main text. □

C.3. Proof of Lemma 4

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

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

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

$$|\hat{c}_e^t| = \left|\frac{c_e^t}{x_e^t}\mathbb{1}\left[e \in p^t\right]\right| \leq \frac{c_{\max}}{\mu}$$

D. Proofs of Section 4

D.1. Proof of Lemma 5

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

$$\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t = \sum_{t=1}^{T} \left\langle c^t, x_i^t \right\rangle + \mathcal{O}\left(\sqrt{c_{\max}} \log(m/\delta)\sqrt{T}\right).$$

Proof. Since the objective is a linear function $\min_{x\in\mathcal{X}}\sum_{t=1}^{T}\langle c^t, x\rangle = \min_{p\in\mathcal{P}}\sum_{t=1}^{T}\sum_{e\in p} c_e^t$, then we have that

$$\sum_{t=1}^{T} X_t = \sum_{t=1}^{T}\sum_{e\in E} c_e^t \cdot \left(\mathbb{E}_t\left[\mathbb{1}[e\in p^t]\right] - \mathbb{1}[e\in p^t]\right)$$

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

$$\sum_{t=1}^{T} X_t = \sum_{e\in E}\sum_{t=1}^{T}\left(\mathbb{E}_t\left[c_e^t\mathbb{1}[e\in p^t]\right] - c_e^t\mathbb{1}[e\in p^t]\right)$$

Hence, we recognize the martingale difference sequence $\left(\left(\mathbb{E}_t\left[c_e^t\mathbb{1}[e\in p^t]\right] - c_e^t\mathbb{1}[e\in p^t]\right)\right)_{t=0}^{T}$ that satisfies $\left|\mathbb{E}_t\left[c_e^t\mathbb{1}[e\in p^t]\right] - c_e^t\mathbb{1}[e\in p^t]\right| \le c_{\max}$ therefore by Azuma-Hoeffding inequality we can conclude that with probability $1-\delta_1$ for every resource $e\in E$

$$\sum_{t=1}^{T}\left(\mathbb{E}_t\left[c_e^t\mathbb{1}[e\in p^t]\right] - c_e^t\mathbb{1}[e\in p^t]\right) \le \sqrt{\frac{1}{2}c_{\max}\log\frac{m}{\delta_1}T}.$$

□

Lemma 6. *Let the sequences $x_i^1, \ldots, x_i^T \in \mathcal{X}_i$ and $\hat{c}^1, \ldots, \hat{c}^T$ produced by Algorithm 2. Then,*

$$\sum_{t=1}^{T}\left\langle \hat{c}^t, x_i^t - x^\star_{\mu_t}\right\rangle \le \frac{2m}{\gamma_T} + m c_{\max}^2 \sum_{t=1}^{T}\frac{\gamma_t}{\mu_t^2}.$$

Proof.

$$\begin{aligned}
\left\langle \hat{c}^t, x_i^t - x^\star_{\mu_t}\right\rangle &\le \frac{\left\langle x_i^t - x_i^{t+1}, x_i^t - x^\star_{\mu_t}\right\rangle}{\gamma_t} \\
&= \frac{1}{2\gamma_t}\left(\left\|x^\star_{\mu_t} - x_i^t\right\|^2 - \left\|x^\star_{\mu_t} - x_i^{t+1}\right\|^2 + \left\|x_i^{t+1} - x_i^t\right\|^2\right) \\
&\le \frac{1}{2\gamma_t}\left(\left\|x^\star_{\mu_t} - x_i^t\right\|^2 - \left\|x^\star_{\mu_t} - x_i^{t+1}\right\|^2\right) + \frac{\gamma_t}{2}\left\|\hat{c}^t\right\|^2 \\
&\le \frac{1}{2\gamma_t}\left(\left\|x^\star_{\mu_t} - x_i^t\right\|^2 - \left\|x^\star_{\mu_t} - x_i^{t+1}\right\|^2\right) + \frac{c_{\max}^2 m}{2}\frac{\gamma_t}{\mu_t^2}.
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle c^t, x_i^t - x^\star_{\mu_t}\right\rangle &\le \sum_{t=1}^{T}\frac{1}{2\gamma_t}\left(\left\|x^\star_{\mu_t} - x_i^t\right\|^2 - \left\|x^\star_{\mu_t} - x_i^{t+1}\right\|^2\right) + \frac{c_{\max}^2 m}{2}\sum_{t=1}^{T}\frac{\gamma_t}{\mu_t^2} \\
&\le \sum_{t=1}^{T}\left(\frac{1}{2\gamma_t}\left\|x^\star_{\mu_t} - x_i^t\right\|^2 - \frac{1}{2\gamma_{t+1}}\left\|x^\star_{\mu_{t+1}} - x_i^{t+1}\right\|^2\right) \\
&\quad + \sum_{t=1}^{T}\left(\frac{\left\|x^\star_{\mu_{t+1}} - x_i^{t+1}\right\|^2}{2\gamma_{t+1}} - \frac{\left\|x^\star_{\mu_t} - x_i^{t+1}\right\|^2}{2\gamma_t}\right) + \frac{c_{\max}^2 m}{2}\sum_{t=1}^{T}\frac{\gamma_t}{\mu_t^2} \\
&\le \frac{\left\|x^\star_{\mu_T} - x_i^T\right\|^2}{2\gamma_T} - \frac{\left\|x^\star_{\mu_0} - x_i^1\right\|^2}{2\gamma_1} + \frac{3m}{2\gamma_T} + \frac{c_{\max}^2 m}{2}\sum_{t=1}^{T}\frac{\gamma_t}{\mu_t^2} \\
&\le \frac{2m}{\gamma_T} + \frac{c_{\max}^2 m}{2}\sum_{t=1}^{T}\frac{\gamma_t}{\mu_t^2}
\end{aligned}$$

Where in the second last inequality we used

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{\left\|x^{\star}_{\mu_{t+1}}-x_i^{t+1}\right\|^2}{2\gamma_{t+1}}-\frac{\left\|x^{\star}_{\mu_t}-x_i^{t+1}\right\|^2}{2\gamma_t}\right) &= \sum_{t=1}^{T}\left(\frac{\left\|x^{\star}_{\mu_{t+1}}-x_i^{t+1}\right\|^2}{2\gamma_{t+1}}-\frac{\left\|x^{\star}_{\mu_{t+1}}-x_i^{t+1}\right\|^2}{2\gamma_t}\right)\\
&\quad+\sum_{t=1}^{T}\frac{\left\|x^{\star}_{\mu_{t+1}}-x_i^{t+1}\right\|^2-\left\|x^{\star}_{\mu_t}-x_i^{t+1}\right\|^2}{2\gamma_t}\\
&= \sum_{t=1}^{T}\left(\left\|x^{\star}_{\mu_{t+1}}-x_i^{t+1}\right\|^2\left(\frac{1}{2\gamma_{t+1}}-\frac{1}{2\gamma_t}\right)\right)\\
&\quad+\sum_{t=1}^{T}\frac{\left\langle x^{\star}_{\mu_t}+x^{\star}_{\mu_{t+1}}-2x_i^{t+1},x^{\star}_{\mu_{t+1}}-x^{\star}_{\mu_t}\right\rangle}{2\gamma_t}\\
&\le m\sum_{t=1}^{T}\left(\frac{1}{2\gamma_{t+1}}-\frac{1}{2\gamma_t}\right)+\sum_{t=1}^{T}\frac{\left\|x^{\star}_{\mu_t}+x^{\star}_{\mu_{t+1}}-2x_i^{t+1}\right\|\left\|x^{\star}_{\mu_{t+1}}-x^{\star}_{\mu_t}\right\|}{2\gamma_t}\\
&\le \frac{m}{2\gamma_T}+\sqrt{m}\sum_{t=1}^{T}\frac{\|(1-m\mu_t)x_i^{\star}+m\mu_t s-(1-m\mu_{t+1})x_i^{\star}-m\mu_{t+1}s\|}{\gamma_t}\\
&= \frac{m}{2\gamma_T}+m^{3/2}\sum_{t=1}^{T}\frac{(\mu_t-\mu_{t+1})\left\|x^{\star}-s\right\|}{\gamma_t}\\
&\le \frac{m}{2\gamma_T}+m^2\sum_{t=1}^{T}\frac{(\mu_t-\mu_{t+1})}{\gamma_t}\\
&\le \frac{m}{2\gamma_T}+\frac{m^2}{\gamma_T}\sum_{t=1}^{T}(\mu_t-\mu_{t+1})\\
&\le \frac{m}{2\gamma_T}+\frac{m^2\mu_1}{\gamma_T}\\
&\le \frac{m}{2\gamma_T}+\frac{m}{\gamma_T}=\frac{3m}{2\gamma_T}
\end{aligned}$$

□

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

$$\sum_{t=1}^{T}\left\langle c^t, x^{\star}_{\mu_t}-x_i^{\star}\right\rangle \le m^2 c_{\max}\sum_{t=1}^{T}\mu_t.$$

where $x_i^{\star}$ and $x^{\star}_{\mu_t}$ are introduced in Definition 4.1.

Proof.

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle c^t, x^{\star}_{\mu_t}-x_i^{\star}\right\rangle &\le \sum_{t=1}^{T}\left\|c^t\right\|_\infty\left\|x_i^{\star}-x^{\star}_{\mu_t}\right\|_1\\
&\le m^2 c_{\max}\sum_{t=1}^{T}\mu_t
\end{aligned}$$

where we used the following bound on $\left\|x_i^{\star}-x^{\star}_{\mu_t}\right\|_1$. Recall that s is the vector as constructed in Remark 4.2.

$$\left\|x_i^{\star}-x^{\star}_{\mu_t}\right\|_1=\left\|x_i^{\star}-(1-m\mu_t)x_i^{\star}-m\mu_t s\right\|_1$$

$$
\begin{aligned}
&= \left\| m\mu_t (x_i^\star - s) \right\|_1 \\
&\leq m^2 \mu_t
\end{aligned}
$$

□

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

$$\sum_{t=1}^{T} \left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \leq \frac{m c_{\max}}{\mu_T} \sqrt{T \log(2m/\delta)}.$$

Proof. We denote $\mathcal{F}^t$ a filtration adapted to the sigma algebra induced by the random variables $\left\{x_i^l\right\}_{l=1}^t$ and let $\mathbb{E}_t$ denote expectation conditioned on $\mathcal{F}^t$. We recognize that $\left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle$ is a bounded martingale difference sequence. Indeed,

$$\mathbb{E}_t \left[\left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \right] = \mathbb{E}_t \left[\left\langle \mathbb{E}_t \left[\hat{c}^t\right] - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \right] = \left\langle \mathbb{E}_t \left[\hat{c}^t\right] - \mathbb{E}_t \left[\hat{c}^t\right], x_i^t - x_{\mu_t}^\star \right\rangle = 0$$

where the last equality holds because x_i^t is $\mathcal{F}^t$-measurable. In addition, $\left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle$ can be bounded as

$$\left| \left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \right| \leq \left\| c^t - \hat{c}^t \right\|_\infty \left\| x_i^t - x_{\mu_t}^\star \right\|_1 \leq 2 m c_{\max} (1 + 1/\mu_t)$$

where we used $\left\| x_i^t - x_{\mu_t}^\star \right\|_1 \leq 2m$ and $\left\| c^t - \hat{c}^t \right\|_\infty \leq c_{\max}(1 + 1/\mu_t)$. Therefore, by Azuma-Hoeffding, with probability $1 - \delta$,

$$\sum_{t=1}^{T} \left\langle c^t - \hat{c}^t, x_i^t - x_{\mu_t}^\star \right\rangle \leq 2 m c_{\max} \sqrt{2 \sum_{t=1}^{T} (1 + 1/\mu_t)^2 \log(1/\delta)} \leq 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)}$$

□

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 $p_i^1, \dots, p_i^T \in \mathcal{P}_i$ the sequence of strategies produced by Algorithm 2 given as input the costs $c^1, \dots, c^T$ with $\left\| c^t \right\|_\infty \leq c_{\max}$. Then with probability $1 - \delta$,*

$$\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t - \min_{p_i^\star \in \mathcal{P}_i} \sum_{e \in p_i^\star} c_e^t \leq \mathcal{O}\left(m c_{\max}^2 T^{4/5} \log(1/\delta) \right)$$

Proof. Combining the previous theorems we can obtained the following bounds

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t - \min_{p_i^\star \in \mathcal{P}_i} \sum_{e \in p_i^\star} c_e^t \leq \sqrt{\frac{1}{2} c_{\max} \log \frac{m}{\delta_1} T} + \frac{m}{\gamma_T} + \frac{c_{\max}^2 m}{2} \sum_{t=1}^{T} \frac{\gamma_t}{\mu_t^2} + m c_{\max} \sum_{t=1}^{T} \mu_t \\
+ 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)}
\end{aligned}
$$

Now, replacing $\gamma_t = t^{-3/5}$ and $\mu_t = \min\left\{1/m, t^{-1/5}\right\}$, we obtain

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{e \in p_i^t} c_e^t - \min_{p_i^\star \in \mathcal{P}_i} \sum_{e \in p_i^\star} c_e^t &\leq \sqrt{\frac{1}{2} c_{\max} T \log \frac{m}{\delta_1}} + \frac{m}{T^{-3/5}} + \frac{c_{\max}^2 m}{2} \sum_{t=1}^{T} \frac{t^{2/5}}{t^{3/5}} + \frac{c_{\max}^2 m}{2} \sum_{t=1}^{m^{1/5}} \frac{m^2}{t^{3/5}} + m c_{\max} \sum_{t=1}^{T} \frac{1}{t^{1/5}} \\
&\quad + m c_{\max} \sum_{t=1}^{m^{1/5}} \frac{1}{m} + 2 m c_{\max} (1 + 1/\mu_T) \sqrt{2T \log(1/\delta)} \\
&= \sqrt{\frac{1}{2} c_{\max} T \log \frac{m}{\delta_1}} + m^2 T^{3/5} + \frac{6 c_{\max} m}{8} T^{4/5} + \frac{c_{\max} m}{2} m^{11/5} + m c_{\max} \frac{5}{4} T^{4/5} \\
&\quad + m^{1/5} c_{\max} + 2 m c_{\max} (1 + T^{1/5}) \sqrt{2T \log(1/\delta)}
\end{aligned}
$$

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,*

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

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 $\lambda \triangleq (2n^2 c_{\max}\sqrt{m})^{-1}$

Proof. We start by taking a first partial derivative of the potential function for the fractional cost $x_{\bar{i}e}$, that is

$$\frac{\partial}{\partial x_{\bar{i}e}}\left[\Phi(x)\right] = \sum_{\mathcal{S}\subset[n],\bar{i}\in\mathcal{S}} \prod_{j\in\mathcal{S},j\neq\bar{i}} x_{je} \prod_{j\notin\mathcal{S}}(1 - x_{je}) \sum_{i=0}^{|\mathcal{S}|} c_e(i) - \sum_{\mathcal{S}\subset[n],\bar{i}\notin\mathcal{S}} \prod_{j\in\mathcal{S}} x_{je} \prod_{j\notin\mathcal{S},j\neq\bar{i}} (1 - x_{je}) \sum_{i=0}^{|\mathcal{S}|} c_e(i)$$

Taking a second a partial derivative with respect to $x_{\bar{j}\bar{e}}$ we obtain

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

Notice that the factor $\prod_{j\in\mathcal{S},j\neq\bar{j}} x_{je} \prod_{j\notin\mathcal{S},j\neq\bar{i}}(1-x_{je})$ 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 $\sum_{\mathcal{S}\subset[n],\bar{i}\in\mathcal{S},\bar{j}\in\mathcal{S}} \prod_{j\in\mathcal{S},j\neq\bar{i},j\neq\bar{j}} x_{je} \prod_{j\notin\mathcal{S}}(1-x_{je}) = 1$. Similar observations hold for the other three terms in the nonzero partial derivatives. Then it follows that $-2nc_{\max} \leq \frac{\partial^2}{\partial x_{\bar{j}\bar{e}}\partial x_{\bar{i}e}}\left[\Phi(x)\right] \leq 2nc_{\max}$ 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 $\left|\frac{\partial^2}{\partial x_{\bar{j}\bar{e}}\partial x_{\bar{i}e}}\left[\Phi(x)\right]\right|^2 \leq 4n^2 c_{\max}^2$. Therefore, we can bound the Hessian Frobenius norm as

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

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 $2n^2 c_{\max}\sqrt{m}$ and therefore the potential function is $2n^2 c_{\max}\sqrt{m}$-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 $\mathcal{X}_i^\mu$*, denoted as $\Delta(\mathcal{P}_i^\mu)$, the set all all probability distributions $\pi_i \in \mathcal{P}_i$ such that

$$\mathbb{P}_{p_i\sim\pi_i}\left[e \in p_i\right] \geq \mu \quad \text{for all } e \in E_i.$$

Now, we can state and prove the auxiliary lemma which relates ϵ, μ stationary point to $4n^2 mc_{\max}\epsilon$-Nash equilibrium for strategies profile in the Caratheodory's decomposition of the set $\mathcal{X}^\mu$.

Lemma 13. *Let $(\pi_i, \pi_{-i}) \in \Delta(\mathcal{P}_1) \times \ldots \times \Delta(\mathcal{P}_n)$ be a Caratheodory decomposition of $x \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ and let x be a (ϵ, μ)-stationary point according to Definition 2. Then for each agent $i \in [n]$,*

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

where $\Delta(\mathcal{P}_i^\mu) \subset \Delta(\mathcal{P}_i)$ is the Caratheodory's decomposition image of the set $\mathcal{X}_i^\mu$.

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

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

Let λ be the inverse of the smoothness parameter of Φ that is $\lambda = \frac{1}{2n^2 c_{\max}\sqrt{m}}$. To simplify notation let $\epsilon := \|\Pi_{\mathcal{X}^\mu}[x - \lambda \nabla \Phi(x)] - x\|_2$ meaning that x is trivially an (ϵ, μ)-stationary point. Then, by the result (Agarwal et al., 2019, Proposition B.1) and (Ghadimi & Lan, 2016, Lemma 3), we have that

$$\max_{\delta \in \Delta} -\delta^T \nabla_{x_i} g(x_i) \le 2\frac{\epsilon}{\lambda}\delta_{\max} \quad \forall i \in [n]$$

with $\Delta \triangleq \{\delta \quad \text{such that} \quad x_i + \delta \in \mathcal{X}_i^\mu, \|\delta\| \le \delta_{\max}\}$ and $g(x_i) := \Phi(x_i, x_{-i})$ is equal to the potential function $\Phi(\cdot)$ when all the players (but the player i) keep their strategy profile fixed. Since $\|x_i - x_i'\| \le \sqrt{m}$ for any $x, y \in \mathcal{X}_i$ we get that $\delta_{\max} \le \sqrt{m}$. Now by plugging the value of λ, setting $\delta = x_i - \bar{x}_i$ on the left hand side, we get that for any player $i \in [n]$:

$$(x_i - \bar{x}_i)^T \nabla_{x_i} g(x_i) \le 4mn^2 c_{\max}\epsilon$$

The function $g(x_i)$ is a linear function (see Definition 2). By linearity of $g(x_i)$ it holds that

$$\begin{aligned}\Phi(x) - \Phi(\bar{x}_i, x_{-i}) &= g(x_i) - g(\bar{x}_i)\\ &= (x_i - \bar{x}_i)^T \nabla_{x_i} g(x_i)\\ &\le 4mn^2 c_{\max}\epsilon\end{aligned}$$

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

$$c_i(\pi_i, \pi_{-i}) - c_i(\bar{\pi}_i, \pi_{-i}) \le 4mn^2 c_{\max}\epsilon$$

□

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

Lemma 14. *Let $\pi = (\pi_1, \ldots, \pi_n) \in \Delta(\mathcal{P}_1) \times \ldots \times \Delta(\mathcal{P}_n)$ and $x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ such that for all resources $e \in E$,*

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

Then the following holds,

$$c_i(\pi_i, \pi_{-i}) - \min_{\bar{\pi}_i \in \Delta(\mathcal{P}_i^\mu)} c_i(\bar{\pi}_i, \pi_{-i}) \le 4n^2 m c_{\max} \cdot \left\| x - \Pi_{\mathcal{X}^\mu} \left[x - \frac{1}{2n^2 c_{\max}\sqrt{m}} \nabla \Phi(x) \right] \right\| + 2m^2 n c_{\max}\mu$$

Proof. Lemma 13 implies that

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

As a result, we just need to bound

$$\min_{\bar{\pi}_i \in \Delta(\mathcal{P}_i^\mu)} c_i(\bar{\pi}_i, \pi_{-i}) - \min_{\bar{\pi}_i \in \Delta(\mathcal{P}_i)} c_i(\bar{\pi}_i, \pi_{-i})$$

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

$$
\begin{aligned}
\min_{\bar{\pi}_i \in \Delta(\mathcal{P}_i^\mu)} c_i(\bar{\pi}_i, \pi_{-i}) - \min_{\bar{\pi}_i \in \Delta(\mathcal{P}_i)} c_i(\bar{\pi}_i, \pi_{-i}) &= \min_{\bar{\pi}_i \in \Delta(\mathcal{P}_\mu^i)} c_i(\bar{\pi}_i, \pi_{-i}) - c_i(\bar{\pi}_i^\star, \pi_{-i}) \\
&\leq c_i(\bar{\pi}_{i,\mu}^\star, \pi_{-i}) - c_i(\bar{\pi}_i^\star, \pi_{-i}) \\
&= \Phi(\bar{x}_{i,\mu}^\star, x_{-i}) - \Phi(\bar{x}_i^\star, x_{-i}) \\
&= \sum_{e \in E} \sum_{\mathcal{S} \subset [n], i \in \mathcal{S}} (\bar{x}_{ie}^\star - (1 - m\mu)\bar{x}_{ie}^\star - m\mu s) \prod_{j \in \mathcal{S}, j \neq i} x_{je} \prod_{j \notin \mathcal{S}} (1 - x_{je}) \sum_{l=0}^{|\mathcal{S}|} c_e(l) \\
&\quad + \sum_{e \in E} \sum_{\mathcal{S} \subset [n], i \notin \mathcal{S}} (-\bar{x}_{ie}^\star + (1 - m\mu)\bar{x}_{ie}^\star + m\mu s) \prod_{j \in \mathcal{S}} x_{je} \prod_{j \notin \mathcal{S}, j \neq i} (1 - x_{je}) \sum_{l=0}^{|\mathcal{S}|} c_e(l) \\
&\leq \sum_{e \in E} (\bar{x}_{ie}^\star - (1 - m\mu)\bar{x}_{ie}^\star - m\mu s)\, n c_{\max} \underbrace{\sum_{\mathcal{S} \subset [n], i \in \mathcal{S}} \prod_{j \in \mathcal{S}, j \neq i} x_{je} \prod_{j \notin \mathcal{S}} (1 - x_{je})}_{=1} \\
&\quad + \sum_{e \in E} (-\bar{x}_{ie}^\star + (1 - m\mu)\bar{x}_{ie}^\star + m\mu s)\, n c_{\max} \underbrace{\sum_{\mathcal{S} \subset [n], i \notin \mathcal{S}} \prod_{j \in \mathcal{S}} x_{je} \prod_{j \notin \mathcal{S}, j \neq i} (1 - x_{je})}_{=1} \\
&= 2m^2 n c_{\max} \mu
\end{aligned}
$$

. □

E.2. Proof of Theorem 5.1

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

$$
x^{t+1} = \Pi_{\mathcal{X}^{\mu_{t+1}}} \left[x^t - \gamma_t \cdot \nabla_t \right] \tag{3}
$$

where the estimator $\nabla_t \triangleq [\hat{c}_1^t, \ldots, \hat{c}_n^t]$ ($\hat{c}_i^t$ 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.* $\mathbb{E}\left[\|\nabla_t\|^2 \right] \leq \frac{n c_{\max}^2 m}{\mu_t}$.

Proof. In the first part of the proof we show that the projection operator is separable. That is, for a generic set $\mathcal{X}$ $\Pi_{\mathcal{X}}(z) = [\Pi_{\mathcal{X}_1}[z_1], \ldots, \Pi_{\mathcal{X}_n}[z_n]]^\intercal$ 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

$$
\Pi_{\mathcal{X}}[z] = \arg\min_{z' \in \mathcal{X}} \|z - z'\|^2 = \arg\min_{z' \in \mathcal{X}} \sum_{i=1}^n \|z_i - z_i'\|^2 = \sum_{i=1}^n \arg\min_{z_i' \in \mathcal{X}_i} \|z_i - z_i'\|^2 = [\Pi_{\mathcal{X}_1}[z_1], \ldots, \Pi_{\mathcal{X}_n}[z_n]]
$$

Let (π_i^t, π_{-i}^t) denote a Caratheodory Decomposition for $x^t := (x_i^t, x_{-i}^t)$. Then,

$$
\begin{aligned}
\mathbb{E}\left[[\nabla_t]_{ie}\right] &= \mathrm{Pr}_{\pi_i^t}\left[\text{agent } i \text{ selects resource } e \text{ at round } t\right] \cdot \mathbb{E}\left[c_e^t / x_{ie}^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right] \\
&= x_{ie}^t \cdot \mathbb{E}\left[c_e^t / x_{ie}^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right] \\
&= \mathbb{E}\left[c_e^t \mid \text{agent } i \text{ selects resource } e \text{ at round } t \right] \\
&= \sum_{\mathcal{S}_{-i} \subseteq [n-1]} \Bigg(\prod_{j \in \mathcal{S}_{-i}} \mathrm{Pr}_{\pi_j^t}\left[\text{agent } j \text{ selects } e \text{ at round } t\right] \\
&\quad \cdot \prod_{j \notin \mathcal{S}_{-i}} \mathrm{Pr}_{\pi_j^t}\left[\text{agent } j \text{ does not select } e \text{ at round } t\right] c_e(|\mathcal{S}_{-i}| + 1) \Bigg)
\end{aligned}
$$

$$= \sum_{\mathcal{S}_{-i}\subseteq[n-1]} \prod_{j\in\mathcal{S}_{-i}} x_{je}^t \prod_{j\notin\mathcal{S}_{-i}} (1-x_{je}^t) c_e\left(|\mathcal{S}_{-i}|+1\right)$$

At the same time,

$$\begin{aligned}
\frac{\partial\Phi(x)}{\partial x_{ie}} &= \sum_{\mathcal{S}_{-i}\subseteq[n-1]} \prod_{j\in\mathcal{S}_{-i}} x_{je} \prod_{j\notin\mathcal{S}_{-i}} (1-x_{je}) \sum_{\ell=0}^{|\mathcal{S}_{-i}|+1} c_e(\ell) - \sum_{\mathcal{S}_{-i}\subseteq[n-1]} \prod_{j\in\mathcal{S}_{-i}} x_{je} \prod_{j\notin\mathcal{S}_{-i}} (1-x_{je}) \sum_{\ell=0}^{|\mathcal{S}_{-i}|} c_e(\ell) \\
&= \sum_{\mathcal{S}_{-i}\subseteq[n-1]} \prod_{j\in\mathcal{S}_{-i}} x_{je}^t \prod_{j\notin\mathcal{S}_{-i}} (1-x_{je}^t) c_e\left(|\mathcal{S}_{-i}|+1\right) \\
&= \mathbb{E}\left[[\nabla_t]_{ie}\right]
\end{aligned}$$

The second part of the proof concerns bounding the norm of the stochastic gradients of the potential function. More precisely, we show that $\mathbb{E}\left[\|\nabla_t\|^2\right] \leq \frac{nc_{\max}^2 m}{\mu_t}$

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_t\|^2\right] &= \sum_{i=1}^{N}\sum_{e\in E_i} \Pr_{\pi_i^t}\left[\text{agent } i \text{ selects resource } e \text{ at round } t\right] \cdot \mathbb{E}\left[(c_e^t/x_{ie}^t)^2 \mid \text{agent } i \text{ selects resource } e \text{ at round } t\right] \\
&= \sum_{i=1}^{n}\sum_{e\in E_i} x_{ie}^t \cdot \mathbb{E}\left[(c_e^t/x_{ie}^t)^2 \mid \text{agent } i \text{ selects resource } e\right] \\
&= \sum_{i=1}^{n}\sum_{e\in E_i} \mathbb{E}\left[(c_e^t)^2/x_{ie}^t \mid \text{agent } i \text{ selects resource } e\right] \\
&\leq \sum_{i=1}^{n}\sum_{e\in E_i} c_{\max}/\mu_t \\
&= \frac{\sum_{i=1}^{n}|E_i|\, c_{\max}^2}{\mu_t} \leq \frac{nmc_{\max}^2}{\mu_t}
\end{aligned}$$

□

E.3. Proof of Theorem 5.2

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

$$2\sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T} + \frac{\lambda c_{\max}^2 nm\sum_{t=1}^{T}\frac{\gamma_t^2}{\mu_t}}{2T\gamma_T} + \frac{8\sqrt{nm^3}}{T}\sum_{t=1}^{T}\mu_t}.$$

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

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

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

$$\begin{aligned}
\phi_{\lambda\mathcal{X}^{\mu_{t+1}}}(x^{t+1}) &\leq \Phi(y^{t+1}) + \frac{1}{\lambda}\left\|x^{t+1}-y^{t+1}\right\|^2 \\
&\leq \Phi(y^{t+1}) + \frac{1}{\lambda}\left\|x^t - \gamma_t\nabla_t - y^{t+1}\right\|^2 \\
&= \Phi(y^{t+1}) + \frac{1}{\lambda}\left\|x^t-y^{t+1}\right\|^2 + \frac{\gamma_t^2}{\lambda}\|\nabla_t\|^2 - \frac{2\gamma_t}{\lambda}(x^t-y^{t+1})^T\nabla_t
\end{aligned}$$

$$= \phi_{\lambda \mathcal{X}^{\mu_{t+1}}}(x^t) + \frac{\gamma_t^2}{\lambda} \left\|\nabla_t\right\|^2 - \frac{2\gamma_t}{\lambda}(x^t - y^{t+1})^T \nabla_t.$$

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 $y^{t+1} = \arg\min_{y \in \mathcal{X}^{\mu_{t+1}}} \left(\Phi(y) + \frac{1}{\lambda}\left\|x^t - y\right\|^2\right)$.

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

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

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

$$\mathbb{E}\left[\phi_{\lambda \mathcal{X}^{\mu_{t+1}}}(x^{t+1})\right] \leq \mathbb{E}\left[\phi_{\lambda \mathcal{X}^{\mu_{t+1}}}(x^t)\right] + \frac{\gamma_t^2}{\lambda}\frac{c_{\max}^2 E}{\mu_t} - \frac{2\gamma_t}{\lambda}\mathbb{E}\left[(x^t - y^{t+1})^T \nabla\Phi(x^t)\right]$$

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

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

which implies that

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

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

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

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

$$\begin{aligned} -\frac{2}{\lambda T}\sum_{t=1}^{T}\gamma_t \mathbb{E}\left[\Phi(y^{t+1}) - \Phi(x^t) + \frac{1}{2\lambda}\left\|y^{t+1} - x^t\right\|^2\right] &\leq \frac{\mathbb{E}\left[\phi_{\lambda \mathcal{X}^{\mu_1}}(x_1)\right] - \mathbb{E}\left[\phi_{\lambda \mathcal{X}^{\mu_T}}(x_T)\right]}{T} + \frac{c_{\max}^2 nm}{\lambda T}\sum_{t=1}^{T}\frac{\gamma_t^2}{\mu_t} \\ &\leq \frac{\Phi_{\max}}{T} + \frac{c_{\max}^2 nm}{\lambda T}\sum_{t=1}^{T}\frac{\gamma_t^2}{\mu_t}. \end{aligned} \tag{4}$$

since $\phi_{\lambda \mathcal{X}^{\mu_1}}(x_1) \leq \Phi(x_1) \leq nmc_{\max}$.

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

$$\begin{aligned} \Phi(x^t) - \Phi(y^{t+1}) - \frac{1}{2\lambda}\left\|y^{t+1} - x^t\right\|^2 &= H(x^t) - H(y^{t+1}) + \frac{1}{2\lambda}\left\|y^{t+1} - x^t\right\|^2 \\ &\geq \nabla H(y^{t+1})^T(x^t - y^{t+1}) + \frac{1}{2\lambda}\left\|y^{t+1} - x^t\right\|^2 \\ &\geq \frac{1}{2\lambda}\left\|y^{t+1} - x^t\right\|^2 \end{aligned}$$

Therefore, plugging into (4), it holds that

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

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

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

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

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

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

$$\begin{aligned}
\left\|y^{t+1}-\tilde{y}^{t+1}\right\| &\leq \left\|\Pi_{\mathcal{X}}[w^t] - \Pi_{\mathcal{X}^{\mu_{t+1}}}[w^t]\right\| + \left\|\Pi_{\mathcal{X}}[\tilde{w}^t] - \Pi_{\mathcal{X}^{\mu_T}}[\tilde{w}^t]\right\| + \left\|\Pi_{\mathcal{X}}[w^t] - \Pi_{\mathcal{X}}[\tilde{w}^t]\right\| \\
&\leq \sqrt{nm^3}\mu_{t+1} + \sqrt{nm^3}\mu_T + \left\|w^t - \tilde{w}^t\right\| \\
&= \sqrt{nm^3}\mu_{t+1} + \sqrt{nm^3}\mu_T + \frac{\lambda}{2}\left\|\nabla\Phi(y^{t+1}) - \nabla\Phi(\tilde{y}^{t+1})\right\| \\
&\leq \sqrt{nm^3}\mu_{t+1} + \sqrt{nm^3}\mu_T + \frac{1}{2}\left\|y^{t+1}-\tilde{y}^{t+1}\right\|
\end{aligned}$$

where in the first inequality, we used the bound distance between a point in $\mathcal{X}$ and its projection in $\mathcal{X}^\mu$ 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

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

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

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|x^t-\tilde{y}^{t+1}\right\| &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|y^{t+1}-\tilde{y}^{t+1}\right\| + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|y^{t+1}-x^t\right\| \\
&\leq \frac{1}{T}\sum_{t=1}^T 2\sqrt{nm^3}(\mu_{t+1}+\mu_T) + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T}+\frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^T \frac{\gamma_t^2}{\mu_t}} \\
&\leq \frac{4\sqrt{nm^3}}{T}\sum_{t=1}^T \mu_t + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T}+\frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^T \frac{\gamma_t^2}{\mu_t}}
\end{aligned}$$

Finally, we conclude the proof with the following steps

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|G(x^t)\right\|_2 &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\Pi_{\mathcal{X}^{\mu_t}}\left[x^t - \frac{\lambda}{2}\nabla\Phi(x^t)\right] - x^t\right\|_2 \\
&\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\Pi_{\mathcal{X}^{\mu_t}}\left[x^t - \frac{\lambda}{2}\nabla\Phi(x^t)\right] - \tilde{y}^{t+1}\right\|_2 + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\tilde{y}^{t+1}-x^t\right\|_2 \\
&\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\Pi_{\mathcal{X}^{\mu_t}}\left[x^t - \frac{\lambda}{2}\nabla\Phi(x^t)\right] - \Pi_{\mathcal{X}^{\mu_t}}\left[x^t - \frac{\lambda}{2}\nabla\Phi(\tilde{y}^{t+1})\right]\right\|_2 \\
&\quad + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T}+\frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^T \frac{\gamma_t^2}{\mu_t}} + \frac{4\sqrt{nm^3}}{T}\sum_{t=1}^T \mu_t \\
&\leq \frac{\lambda}{2T}\sum_{t=1}^T \mathbb{E}\left\|\nabla\Phi(x^t) - \nabla\Phi(\tilde{y}^{t+1})\right\|_2 + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T}+\frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^T \frac{\gamma_t^2}{\mu_t}} + \frac{4\sqrt{nm^3}}{T}\sum_{t=1}^T \mu_t
\end{aligned}$$

$$\leq \frac{\lambda}{2T}\frac{1}{\lambda}\sum_{t=1}^{T}\mathbb{E}\left\|x^t - \tilde{y}^{t+1}\right\|_2 + \sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^{T}\frac{\gamma_t^2}{\mu_t} + \frac{4\sqrt{nm^3}}{T}\sum_{t=1}^{T}\mu_t}$$

$$\leq 2\sqrt{\frac{\lambda^2 nmc_{\max}}{2T\gamma_T} + \frac{\lambda c_{\max}^2 nm}{2T\gamma_T}\sum_{t=1}^{T}\frac{\gamma_t^2}{\mu_t} + \frac{8\sqrt{nm^3}}{T}\sum_{t=1}^{T}\mu_t}$$

that concludes the proof. □

E.4. Proof of Theorem 3.8

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

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

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

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

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

As a result,

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

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

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

To simplify notation let

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

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

$$\begin{aligned}
\text{(A)} &\leq 4\sqrt{m}\sqrt{\frac{nmc_{\max}}{2T^{2/5}C_\gamma} + \frac{c_{\max}^2 nmC_\gamma}{2T^{2/5}\lambda C_\mu}\sum_{t=1}^{m^{1/5}}\frac{1}{mt^{6/5}} + \frac{c_{\max}^2 nmC_\gamma}{2T^{2/5}\lambda C_\mu}\sum_{t=1}^{T}\frac{t^{1/5}}{t^{6/5}}} \\
&\quad + \frac{20\sqrt{n}m^2C_\mu}{\lambda T^{1/5}} + \frac{2m^2nc_{\max}C_\mu}{T^{1/5}} \\
&\leq 4\sqrt{m}\sqrt{\frac{nmc_{\max}}{2T^{2/5}C_\gamma} + \frac{c_{\max}^2 nmC_\gamma}{2T^{2/5}\lambda C_\mu}(\log T + 1)} + \frac{20\sqrt{n}m^2C_\mu}{\lambda T^{1/5}} + \frac{2m^2nc_{\max}C_\mu}{T^{1/5}} \\
&= \left(4\sqrt{m}\sqrt{\frac{nmc_{\max}}{2C_\gamma} + \frac{c_{\max}^2 nmC_\gamma}{2\lambda C_\mu}(\log T + 1)} + \frac{20\sqrt{n}m^2C_\mu}{\lambda} + 2m^2nc_{\max}C_\mu\right)T^{-1/5}
\end{aligned}$$

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

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

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:

$$\begin{aligned}(\mathrm{A}) &\leq \left(4\sqrt{m}\sqrt{\frac{c_{\max}^2 n^{13/5}m^{9/5}}{2} + c_{\max}^2 n^{13/5}m^{9/5}(\log T + 1)} + 80m^{14/10}n^{13/10}c_{\max}\right)T^{-1/5}\\ &\leq \frac{88m^{14/10}n^{13/10}c_{\max}(\log T + 1)}{T^{1/5}}\end{aligned}$$

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

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

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

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

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

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

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

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

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

F. Auxiliary Lemmas

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

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

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

Proof. By the Caratheodory's theorem, there exists a collection of simple paths $\{\hat{p}_1, \dots, \hat{p}_{m+1}\}$ and scalars $\lambda_1, \dots, \lambda_{m+1}$ such that $\lambda_i \geq 0 \quad \forall i \in \{1, m+1\}$ 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 $\hat{p}_{je} = 0$ for all $j \in \{1, m+1\}$. This implies that $x_e = 0$ which is a clear contradiction. That means that there exist $j^\star \in \{1, m+1\}$ such that $e \in \hat{p}_{j^\star}$ proving part (i). In addition it must be true that the weight in the convex combination is positive, i.e. $\lambda_{j^\star} > 0$. This implies that $x_e \geq \lambda_{j^\star}\hat{p}_{j^\star e}$. Therefore, for the edges e s.t. $\hat{p}_{j^\star e} = 1$, it holds that $x_e \geq \lambda_{j^\star} > 0$. □

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

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

Proof. We start by manipulating the cost difference

$$\begin{aligned}
c_i\left(\bar{\pi}_i, \pi_{-i}\right) - c_i\left(\pi_i, \pi_{-i}\right) &= \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \left(\mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p_i\right] C_i(p_i, p_{-i}) - \mathop{\mathbb{P}}_{\pi_i}\left[p_i\right] C_i(p_i, p_{-i}) \right) \\
&= \sum_{p_i \in \mathcal{P}_i} \sum_{p'_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p'_i\right] \mathop{\mathbb{P}}_{\pi_i}\left[p_i\right] \left(C_i(p'_i, p_{-i}) - C_i(p_i, p_{-i}) \right) \\
&= \sum_{p_i \in \mathcal{P}_i} \sum_{p'_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p'_i\right] \mathop{\mathbb{P}}_{\pi_i}\left[p_i\right] \left(\Phi(p'_i, p_{-i}) - \Phi(p_i, p_{-i}) \right) \\
&= \sum_{e \in E} \sum_{p_i \in \mathcal{P}_i} \sum_{p'_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p'_i\right] \mathop{\mathbb{P}}_{\pi_i}\left[p_i\right] \left(\sum_{i=0}^{l_e(p'_i, p_{-i})} c_e(i) - \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) \right) \\
&= \sum_{e \in E} \sum_{p'_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p'_i\right] \sum_{i=0}^{l_e(p'_i, p_{-i})} c_e(i) \\
&\quad - \sum_{e \in E} \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\pi_i}\left[p_i\right] \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i)
\end{aligned}$$

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

$$\begin{aligned}
\sum_{e \in E} \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p_i\right] \sum_{i=0}^{l_e(p_i, p_{-i})} c_e(i) &= \sum_{s=1}^{N} \sum_{e \in E} \sum_{p_i \in \mathcal{P}_i} \sum_{p_{-i} \in \mathcal{P}_{-i}} \mathop{\mathbb{P}}_{\pi_{-i}}\left[p_{-i}\right] \mathop{\mathbb{P}}_{\bar{\pi}_i}\left[p_i\right] \mathbb{1}\left\{ l_e(p_i, p_{-i}) = s \right\} \sum_{i=0}^{s} c_e(i) \\
&= \sum_{e \in E} \sum_{s=1}^{N} \mathop{\mathbb{P}}_{\pi_i, \bar{\pi}_i}\left[\text{Exactly } s \text{ agents select edge } e\right] \sum_{i=0}^{s} c_e(i) \\
&= \sum_{e \in E} \sum_{s=1}^{N} \sum_{\mathcal{S} \subset \binom{[n]}{s}} \prod_{j \in \mathcal{S}} x_{je} \prod_{j \notin \mathcal{S}} (1 - x_{je}) \sum_{i=0}^{s} c_e(i) \\
&= \sum_{e \in E} \sum_{\mathcal{S} \subset [n]} \prod_{j \in \mathcal{S}} x_{je} \prod_{j \notin \mathcal{S}} (1 - x_{je}) \sum_{i=0}^{|\mathcal{S}|} c_e(i)
\end{aligned}$$

□

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 $x \notin \Delta_n^\mu$.

Example for $x \in \Delta_n^\mu$: Consider $\epsilon = \mu$ and $x = (2/3, 1/3)$. In this case ϵ-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 ϵ-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 $\mu \leq \frac{0.4}{3}$ we can use the KKT conditions of the problem to derive that $x' = (x_1', x_2', x_3')$ must satisfy the following system

$$\begin{cases} 2(x_1' - 0.8) + \lambda = 0 \\ 2(x_2' - 0.2) + \lambda = 0 \\ 2(x_3' - 0.0) + \lambda \geq 0 \\ -x_1' + \mu < 0 \\ -x_2' + \mu < 0 \\ -x_3' + \mu = 0 \end{cases} \tag{7}$$

which admits the solution $x' = (0.8 - \mu/2, 0.2 - \mu/2, \mu)$ which is a different transformation than the one obtained with ϵ-greedy. To see this, take for example $\epsilon = \mu = 0.12$. ϵ 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 $(0.8 - \mu/2, 0.2 - \mu/2, \mu) = (0.74, 0.14, 0.12)$.