Online Learning of Coalition Structures by Selfish Agents

Saar Cohen, Noa Agmon

Department of Computer Science, Bar-Ilan University, Israel
saar30@gmail.com, agmon@cs.biu.ac.il

Abstract

Coalition formation concerns autonomous agents that strategically interact to form self-organized coalitions. When agents lack initial sufficient information to evaluate their preferences *before* interacting with others, they learn them *online* through repeated feedback while iteratively forming coalitions. In this work, we introduce online learning in coalition formation from a *non-cooperative* perspective, studying the impact of *collective data utilization* where selfish agents aim to accelerate their learning by leveraging a shared data platform. Thus, the efficiency and dynamics of the learning process are affected by each agent's local feedbacks, motivating us to explore the tension between *semi-bandit* and *bandit* feedback, which differ in the granularity of utility information observed by each agent. Under our non-cooperative viewpoint, we evaluate the system by means of *Nash stability*, where no agent can improve her utility by unilaterally deviating. Our main result is a sample-efficient algorithm for selfish agents that aims to minimize their Nash regret under both semi-bandit and bandit feedback, implying approximately Nash stable outcomes. Under both feedback settings, our algorithm enjoys Nash regret and sample complexity bounds that are ***optimal*** up to logarithmic factors.

1 Introduction

A set of students begins their Bachelor's degree, driven by their individual desires to enhance their academic successes via forming study groups during the semester. Students initially lack clarity on their own preferences about group size, study methods, or topics. By joining study sessions based on their self-interest and interacting with others, students learn about their preferences via feedback on productivity and satisfaction with group dynamics. This adaptive process allows them to selfishly optimize their own studying experiences, making informed decisions to join study groups aligned with their individual working styles and academic goals. Such cases and many other real-world scenarios noticed in our social, economic, and politic life, fall within the phenomenon of *coalition formation*, where *agents* perform activities in *coalitions* rather than on their own.

A prominent framework for studying coalition formation is that of *hedonic games* (Dreze and Greenberg 1980), where the agents' utilities solely depend on the coalition they are part of, without caring about the structure of other coalitions, i.e., *externalities* are ignored. The outcome of such games is a set of disjoint coalitions (hereafter, *partition*). The desirability of partitions is often measured by the common criterion of *stability* (Aziz and Savani 2016; Bullinger and Romen 2024), capturing the prospect of agents maintaining their coalitions. The hedonic games literature on stability typically concerns only the final outcome of coalition formation, ignoring the process of reaching stable partitions. Specifically, in most existing works it is implicitly assumed that a central authority can obtain the agents' preferences, find a stable partition and enforce it on the agents. On the contrary, recent studies consider a *dynamic* process where, starting from a given partition, agents deliberately move between coalitions based on their individual preferences (Boehmer, Bullinger, and Kerkmann 2023; Brandt, Bullinger, and Tappe 2022).

However, in many realistic scenarios as our study groups example, agents' preferences toward others are initially *unknown* prior to interactions (see, e.g., (Cohen and Agmon 2023a)). Thus, each agent must make decisions individually and perform online learning of her own preferences from repeated feedback by iteratively joining coalitions. To reflect such settings, Cohen and Agmon (2024b) study online learning in hedonic games, aiming to maximize social welfare.

In contrast, in this paper we present and study a *new* model for online learning in coalition formation under a *non-cooperative* perspective, evaluating partitions by means of *Nash stability*, where no agent can improve her utility by unilaterally deviating (e.g., no student can improve in her studies by changing groups unilaterally). This stability notion is well-suited to our context as it captures the selfish incentives of single agents to perform improving deviations. We exhibit our results for *additively separable hedonic games* (ASHGs) with *symmetric* preferences (Bogomolnaia and Jackson 2002), where an agent's utility for a coalition is the sum of her utilities from other coalition members. The existence of a Nash stable outcome in those games is guaranteed (Bogomolnaia and Jackson 2002), but computing such partitions is PLS-complete (Gairing and Savani 2019).

In many realistic scenarios, agents interact in a shared environment, as in our study groups example. This prompts us to explore how online learning algorithms are affected

by *collective data utilization*, where selfish agents aim to enhance their learning by exploiting a common data platform. An algorithm's learning process is thus affected by each agent's local feedbacks, motivating us to study the interplay between *semi-bandit* feedback, where agents obtain feedbacks from interactions within their coalition, and *bandit* feedback, where agents only observe the overall utility received from their coalition. To quantify how far a strategy produced by an algorithm is from being Nash stable, we use the popular metric of *Nash regret* (Ding et al. 2022; Liu et al. 2021b), evaluating each agent's learnt strategy with her best response strategy at any round. A sublinear Nash regret translates to low sample complexity, as it implies best-iterate convergence to an approximate Nash stable solution with fewer samples due to, e.g., (Jin et al. 2018).

Contributions. We devise sample-efficient online learning algorithms for selfish agents that minimize their Nash regret, obtaining approximately Nash stable partitions. Though our model is proven to be a sequence of potential games, the state-of-the-art method by Liu et al. (2021b) for potential games with collective data utilization is unsuited to our context: its Nash regret and sample complexity bounds are linear in the size of the joint strategies' space, which, in our setting, is *exponential* in the number of agents. As a remedy, we propose an algorithm whose Nash regret and sample complexity bounds are not only *polynomial* in the number of agents under either semi-bandit or bandit feedback, but they are also ***optimal*** up to logarithmic factors under both feedback models. Based on the principle of *"optimism in the face of uncertainty"*, our algorithm optimistically estimates the agents' preferences using upper confidence bounds (UCBs), a common approach in many bandit algorithms (Lattimore and Szepesvári 2020). For both feedback settings, we prove that the agents' UCB estimates induce a potential game at each round. Hence, we update the agents' joint strategy to be an approximate Nash stable strategy of this game, which we prove can be found in a *polynomial* number of steps by a natural better-response dynamics induced by *approximate* Nash deviations. This is in contrast to better-response dynamics based on *exact* Nash deviations in symmetric ASHGs, which may converge to an *exact* NS partition only after an *exponential* number of steps (Brandt, Bullinger, and Tappe 2024). All omitted proofs can be found in the supplementary materials (Cohen and Agmon 2025).

2 Related Work

Hedonic games have been introduced by Drèze and Greenberg (1980), and later extended to the study of various solution concepts like stability, fairness, and optimality (see, e.g., (Aziz and Savani 2016; Woeginger 2013)). One major concern is designing computationally manageable classes of hedonic games, which led to an abundance of game representations. Some are *ordinal* and can *fully* express any preference over coalitions (Bouveret et al. 2010; Elkind and Wooldridge 2009), but may require exponential space. In contrast, *cardinal* hedonic games, based on weighted graphs (Aziz et al. 2019; Bogomolnaia and Jackson 2002), are *not* fully expressive, yet only require *polynomial* space for reasonable weights. Our work focuses on non-cooperativeness in *additively separable hedonic games* (ASHGs) with *symmetric* preferences (Bogomolnaia and Jackson 2002), where a vast body of work evaluates the system in terms of Nash stability (Aziz, Brandt, and Seedig 2011; Banerjee, Konishi, and Sönmez 2001; Ballester 2004). Bogomolnaia and Jackson (2002) proved that Nash stable partitions may not exist in general ASHGs, while Sung and Dimitrov (2010) showed that checking if an instance admits such partition is NP-complete in the strong sense. Yet, for *symmetric* preferences, the existence of a Nash stable outcome is guaranteed by potential function argument (Bogomolnaia and Jackson 2002), but computing such partitions is PLS-complete (Gairing and Savani 2019), where the complexity class PLS (Polynomial Local Search) consists of local search problems with polynomially verifiable local optimality (Johnson, Papadimitriou, and Yannakakis 1988). However, the above works consider ASHGs in an *offline* setting, while we consider an *online* setting.

Our work is thus closely tied to *time-dependent* models in hedonic games, including their *online* version presented by Flammini *et al.* (2021b), where agents arrive one at a time and should be *immediately* and *irrevocably* assigned to coalitions with the goal of maximizing social welfare. This problem was recently extended to other setups (Cohen and Agmon 2024a; Bullinger and Romen 2023), with Bullinger and Romen (2024) also exploring various *stability* concepts. Yet, the assumption that the agents are partitioned by an *external* authority may be unrealistic as agents usually make decisions individually. In contrast, we take a *non-cooperative* approach, where agents individually and selfishly choose to form a new coalition or to join an existing one by leveraging a shared data platform.

Recently, dynamic approaches to hedonic games also received increased attention, exploring deviation dynamics for either single-agent stability (Bilò et al. 2018; Boehmer, Bullinger, and Kerkmann 2023; Brandt, Bullinger, and Tappe 2022) or group stability (Carosi, Monaco, and Moscardelli 2019; Fanelli, Monaco, and Moscardelli 2021). Brandt et al. (2024) show that the above PLS-completeness result of Nash stability implies that dynamics based on Nash deviations in symmetric ASHGs may converge to an *exact* NS partition only after an *exponential* number of steps. However, we show that this negative result can be mitigated: our proposed algorithm employs an *approximate* dynamics, proven to always converge to an *approximate* Nash stable partition after *polynomially* many steps.

However, the above studies on online and dynamic hedonic games unrealistically demand that the agents' preferences are fully known. Existing works on PAC learnability in hedonic games tackle this issue (Sliwinski and Zick 2017; Fioravanti et al. 2023). Unlike our work, they take a *cooperative* approach with the goal of efficiently inferring preferences from a limited and fixed number of offline samples. Yet, they all assume exact knowledge of preferences *before* making decisions, limiting practicality as agents often require time to learn their own preferences from social interactions as in our study groups example. Particularly, the PAC learning approach is unsuited to our *dynamic* setting due to its *static* nature, requiring a fixed set of preferences

that are available upfront. Unlike prior studies, we thus consider realistic scenarios where each agent *dynamically* learns her own preferences through repeated interactions, allowing her to adapt to changing situations. Our repeated game approach allows agents to learn the coalitions proven most relevant and effective for their selfish desires.

We offer a novel framework that resolves those issues by studying online learning in coalition formation from a *non-cooperative* viewpoint, opposed to the online learning approach to hedonic games introduced by Cohen and Agmon (2024b), whose goal is maximizing social welfare. Our work contributes to the growing focus on online learning in combinatorial domains such as bandits (Tekin and Van Der Schaar 2015), online task allocation (Cohen and Agmon 2023b), congestion games (Cui et al. 2022; Panageas et al. 2023) and many more. Traditional literature on learning in games mainly regards how various dynamics asymptotically converge to a Nash equilibrium (e.g., no-regret dynamics (Daskalakis, Fishelson, and Golowich 2021; Chen and Peng 2020) and fictitious play (Daskalakis and Pan 2014; Leslie and Collins 2006)). In contrast, we focus on non-asymptotic convergence, as done in recent studies on multi-agent reinforcement learning where the central performance measure is *Nash regret* (Ding et al. 2022; Liu et al. 2021b), comparing an agent's learnt strategy with her best response strategy at any *individual* round. This metric is thus suitable to our setting, which is non-stationary from the standpoint of each agent. Further, by classic online-to-batch conversion (e.g., (Jin et al. 2018, Section 3.1)), sublinear Nash regret yields low sample complexity, leading to best-iterate convergence to an approximate Nash stable solution with fewer samples.

Highly tied to our work are *matching markets*, which can be viewed as a *constrained* version of hedonic games where coalitions are limited to be of size at most 2. Unlike prior work on matching markets (Maheshwari, Sastry, and Mazumdar 2022; Zhang, Wang, and Fang 2022; Liu, Mania, and Jordan 2020; Eichhorn, Banerjee, and Kempe 2022; Sentenac et al. 2021; Rodet and Gaudel 2022), our setting is more challenging as it is *not* confined to matchings and we consider more practical settings with broader utility functions under either semi-bandit or bandit feedback.

As noted above, our proposed algorithm optimistically estimates agents' preferences via upper confidence bounds (UCBs), a popular approach in many bandit algorithms (Lattimore and Szepesvári 2020). Existing works on matchings have also employed UCBs, yet they are limited to either restrictive conditions on preferences (Sankararaman, Basu, and Sankararaman 2021; Maheshwari, Sastry, and Mazumdar 2022) or a less strict notion of regret (Liu et al. 2021a). Our work is further connected to potential games with collective data utilization, where the state-of-the-art method by Liu et al. (2021b) also uses UCBs. Yet, its Nash regret and sample complexity bounds are linear in the size of the joint strategies' space, which, in our context, is *exponential* in the number of agents. Congestion games, a special subclass of potential games, also face an *exponential* strategy space, and though Cui et al. (2022) remove this dependency via a UCB-based approach, their Nash regret and sample complexity bounds under bandit feedback are *suboptimal*, and they lack a tight analysis for both semi-bandit and bandit feedback. In contrast, the Nash regret and sample complexity bounds of our algorithm are *polynomial* in the number of agents, asymptotically tight and ***optimal*** up to logarithmic factors under both semi-bandit and bandit feedback. Further, unlike other methods for potential games where agents *independently* update their policies (Ding et al. 2022; Leonardos et al. 2022), our algorithm eliminates the linear dependency on the number of actions an agent can take.

3 Preliminaries

We study an *online learning* version of hedonic games, where selfish agents with initially *unknown* preferences partition themselves into disjoint subsets (i.e, *coalitions*) over T rounds. Formally, our non-cooperative strategic game is given by a finite set $N = \{1, \dots, n\}$ of n self-interested agents with *unknown* preferences, while we hereafter denote $[k] := \{1, \dots, k\}$ for $k \in \mathbb{N}$ and $[0] = \{0\}$. At any time $t \in [T]$, each agent can join one of n *candidate* coalitions since there are n agents (i.e., a partition can contain between 1 to n coalitions). In our study groups example, this can be thought of as if each student chooses which room to enter among n rooms. Hence, at any time t, each agent i chooses to join a certain coalition among n candidate ones following a ***mixed strategy*** φ_i^t, built on all the information available to her from past rounds. Time $t = 1$ is exceptional, where each agent i arbitrarily initializes her strategy φ_i^1. Formally, $\varphi_i^t \in \Delta([n])$ where $\Delta([n])$ is the probability simplex over $[n]$, i.e., for any $x \in [n]$, agent i picks the xth candidate coalition with probability $\varphi_i^t(x) \in [0, 1]$. Once each agent has played her strategy, let $\varphi^t = (\varphi_i^t)_{i \in N} \in \Delta([n])^n$ be the agents' ***joint mixed strategy*** at time t. At any time t, each agent i then samples an ***assignment*** $x_i \in [n]$ from φ_i^t independently from other agents, forming a ***joint assignment*** $\mathbf{x} = (x_i)_{i \in N}$. Thus, the agents' iterative decision process during time t unfolds as follows:

1. Each agent i samples $x_i \sim \varphi_i^t$ and joins the x_ith candidate coalition, forming a joint assignment $\mathbf{x} = (x_i)_{i \in N}$.
2. Each agent i observes the other members in the coalition she joined and gets feedback(s) about the utility gained from her own coalition.
3. Agent i's chosen candidate coalition and resulting utility feedback(s) are propagated to a shared data platform accessible by all agents.
4. Using the platform's information, agent i updates her strategy to be φ_i^{t+1} and moves to the next time step $t+1$.

We term the game formed by this learning process as ***online learning ASHGs*** (OL-ASHGs). Next, we elaborate on this process in detail. The formed joint assignment $\mathbf{x}$ induces a partition of the agents $\pi^{\mathbf{x}} = (C_\ell^{\mathbf{x}})_{\ell \in [n]}$, where, for any $\ell \in [n]$, $C_\ell^{\mathbf{x}}$ is the set of agents joining the ℓth candidate coalition, i.e., $C_\ell^{\mathbf{x}} = \{i \in N : x_i = \ell\}$. As the number of candidate coalitions equals to the number of agents, some coalitions may be empty. Accordingly, we denote by $|\pi^{\mathbf{x}}|$ the number of non-empty coalitions in $\pi^{\mathbf{x}}$. After her assignment, notice that agent i becomes aware of the other members within the coalition she joined. We thus denote the coalition in $\pi^{\mathbf{x}}$ containing agent i as $\pi^{\mathbf{x}}(i)$.

Afterwards, each agent obtains a utility derived from her chosen strategy, determined by aggregating her valuations of other agents. We focus on *additively separable hedonic games* (ASHGs) with *symmetric* preferences, where any pair of agents assign the same numerical value toward each other, indicating the intensity by which they prefer each other to another agent. As common in the literature (see, e.g., (Flammini et al. 2021a)), we assume that agents' valuations are within $[-1, 1]$. Recall that preferences are *unknown* and even the agents themselves may not be aware of them. Thus, for any pair of distinct agents i, j, the uncertainty about their mutual valuation is captured by an *unknown* and *fixed* distribution $\mathcal{D}_{i,j}$ over $[-1, 1]$ with mean $d_{i,j}$, which the agents i and j aim to learn. At each time t, the utility $v_{i,j}^t$ of agents i, j for each other is then independently drawn from $\mathcal{D}_{i,j}$. We use the convention that $v_{i,i}^t = d_{i,i} = 0$ for any agent i. For any joint assignment $\mathbf{x}$ sampled at time t, agent i's utility from the partition induced by $\mathbf{x}$ is then given by $v_i^t(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i)} v_{i,j}^t$, whose mean is $d_i(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i)} d_{i,j}$. Agent i's utility from her strategy φ_i^t at time t is thus defined as $V_i(\varphi^t) := \mathbb{E}_{\mathbf{x} \sim \varphi^t}[d_i(\mathbf{x})]$.

Each agent joins a coalition with the goal of maximizing her own utility. We thus want to study stability under single agents' incentives to deviate between coalitions. When strategies are *pure*, each agent i's strategy is joining a *single* candidate coalition at time t by only selecting some $x_i \in [n]$. Let $\mathbf{x}_{-i} = (x_j)_{j \neq i}$ be the joint strategy of all agents except for agent i. Agent i can then ***deviate*** by moving from her selected coalition to another one with index $y_i \in [n]$, which is a ***Nash deviation*** if it improves her utility, i.e., $V_i^t(\mathbf{x}_{-i}, y_i) > V_i^t(\mathbf{x}_{-i}, x_i)$. When working with *mixed* strategies, we define another notion of Nash deviations. Consider the joint strategy φ^t at time t. Letting $\varphi_{-i}^t = (\varphi_j^t)_{j \neq i}$ for any agent i, agent i may perform a ***(mixed single-agent) deviation*** from her strategy φ_i^t to another strategy $\phi_i \in \Delta([n])$, which is a ***mixed Nash deviation*** only if it immediately makes her better off, i.e., $V_i(\varphi_{-i}^t, \phi_i) > V_i(\varphi^t)$. Hence, a (*pure* or *mixed*) joint strategy for which no Nash deviation is possible is said to be ***Nash stable*** (NS), also called a *Nash equilibrium* (NE). We further consider an *approximate* notion of Nash stability. At each time t, note that agent i's ***best response*** to the other agents' strategies is a Nash deviation given by a strategy $\phi_i^{\star,t}$ satisfying $V_i^\star(\varphi_{-i}^t) := V_i(\varphi_{-i}^t, \phi_i^{\star,t}) = \max_{\phi \in \Delta([n])} V_i(\varphi_{-i}^t, \phi)$. Thus, for any $\varepsilon \geq 0$, the agents' joint strategy φ^t at time t is ***ε-approximate Nash stable*** (ε-NS) if $\max_{i \in N}(V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t)) \leq \varepsilon$. Here, the quantity $\max_{i \in N}(V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t))$ measures the ***worst agent's local gap*** between the expected utilities she gets from her best response and her current strategy at time t, respectively.

We consider the most general setting under ***collective data utilization***, where each agent should learn her own preferences from repeated feedback so as to calculate a strategy yielding an ε-NS partition for some $\varepsilon \geq 0$. At each time t and for any joint assignment $\mathbf{x}$, we assume that each agent i can receive either *semi-bandit* or *bandit feedback*. The ***semi-bandit feedback*** setting models scenarios where each agent can observe feedback from her interaction with any other member of her chosen coalition, i.e., each agent i attains her realized utility $v_{i,j}^t$ for each agent $i \neq j \in \pi^{\mathbf{x}}(i)$. In contrast, ***bandit feedback*** refers to cases where each agent can only observe the overall utility she receives from her coalition, i.e., each agent i only obtains her utility from the partition induced by $\mathbf{x}$ (i.e., $v_i^t(\mathbf{x})$), with no information about the individual utility $v_{i,j}^t$ assigned to any agent $j \neq i$.

Our goal is devising an online learning algorithm that learns an ε-NS joint strategy for some $\varepsilon \geq 0$. At any time t, each agent i updates her strategy φ_i^{t-1} based on the history of play recorded on the shared platform up to time $t-1$, including all partitions and all agents' utility feedbacks until that time. We analyze an algorithm's performance in terms of *Nash regret* (Ding et al. 2022; Liu et al. 2021b), comparing each agent's learnt strategy with her best response strategy at any round. Formally, given a sequence of joint strategies $\{\varphi^t\}_{t=1}^T$, the ***Nash regret*** after T rounds is:

$$\mathcal{R}^T := \sum_{t=1}^T \max_{i \in N}(V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t)) \qquad (1)$$

Intuitively, Nash regret measures how far the agents' strategies are from an approximate Nash equilibrium, as it sums the worst agent's local gap between her learned strategy and her best response to others' strategies across all rounds. For any $\varepsilon \geq 0$, it is well-known that this correlation can be used to show that an algorithm with a Nash regret bound of ε obtains an ε-NS strategy (see, e.g., (Ding et al. 2022)). Hence, one of our goals is minimizing the Nash regret, i.e., we aim to devise algorithms with a ***Nash regret bound*** that is sublinear in the number of rounds T and polynomial in the number of agents n. Another goal is finding an ε-NS strategy using a number of episodes that is small in its dependency on the number of agents n and $1/\varepsilon$, guaranteeing a ***(PAC) sample complexity bound***. In fact, note that any algorithm with sublinear Nash regret can be directly converted to a polynomial-sample algorithm via the standard online-to-batch conversion (see, e.g., (Jin et al. 2018)).

3.1 OL-ASHGs are Sequences of Potential Games

We begin with discussing a useful connection of our model to potential games. In case of a single round (i.e., $T = 1$) and *no* uncertainty, a Nash stable strategy is guaranteed for ASHGs with *known* symmetric preferences through a potential function argument (Bogomolnaia and Jackson 2002). In general OL-ASHGs, where the symmetric preferences are *unknown* at any time t, we show that the game associated with each time t is also a potential game:

Lemma 1. *At any time t, the ASHG with symmetric and unknown preferences during time t is a* potential game.

Proof. Consider a joint mixed strategy φ^t at time t. Given a joint assignment $\mathbf{x} \sim \varphi^t$, note that $\Phi^t(\mathbf{x}) = \sum_{i \in N} v_i^t(\mathbf{x})$ is a potential function for *pure* strategies as $\Phi^t(\mathbf{x}_{-i}, x_i) - \Phi^t(\mathbf{x}_{-i}, x_i') = v_i^t(\mathbf{x}_{-i}, x_i) - v_i^t(\mathbf{x}_{-i}, x_i')$ for any agent i and another assignment $x_i' \in [n]$ of agent i. Therefore, by slight abuse of notation, $\Phi(\varphi^t) = \sum_{i \in N} V_i(\varphi^t)$ is a potential function for *mixed* strategies since $\Phi(\varphi_{-i}^t, \varphi_i^t) - \Phi(\varphi_{-i}^t, \phi_i) = V_i(\varphi_{-i}^t, \varphi_i^t) - V_i(\varphi_{-i}^t, \phi_i)$ for any agent i and another strategy $\phi_i \in \Delta([n])$ of agent i. □

Algorithm 1: **UCB-NS**

Input: T rounds; n agents; $\delta, \varepsilon \in (0, 1]$.

1: Initialize an arbitrary joint mixed strategy φ^1.
2: **for each** time $t = 1, \ldots, T$ **do**
3: **for each** agent $i \in N$ **do**
4: Agent i joins a coalition by sampling $x_i^t \sim \varphi_i^t$.
5: **Bandit Feedback:** Agent i gets a utility v_i^t
6: from her coalition.
7: Set $\bar{u}_i^t(\mathbf{x}) = \bar{v}_i^t(\mathbf{x}) + \bar{b}_i^t(\mathbf{x})\ \forall \mathbf{x} \in [n]^n$ by (2)-(3).
8: **Semi-Bandit Feedback:** Agent i gets a utility
9: $v_{i,j}^t$ from each agent j in her coalition.
10: Set $\hat{u}_i^t(\mathbf{x}) = \hat{v}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})\ \forall \mathbf{x} \in [n]^n$ by (3)-(4).
11: Via Algorithm 2, compute an ε-NS strategy φ^{t+1} of:
12: **Bandit Feedback:** $\mathcal{G}^t$ from Lemma 2.
13: **Semi-Bandit Feedback:** $\bar{\mathcal{G}}^t$ from Lemma 3.

Online learning algorithms for our context should thus adhere to the principle that any ***Nash stable*** strategy is a stationary point of the potential function from Lemma 1. Computational efficiency and scalability are other key design factors: a critical challenge in our setting is that the joint strategies' space is *exponential* in the number of agents n. Hence, an efficient algorithm should have Nash regret and sample complexity polynomial in the number of agents n, without dependence on the size of the strategy space. As mentioned in Section 2, the existing state-of-the-art centralized method for potential games fails to satisfy this criterion (Liu et al. 2021b). In Section 4, we thus design an online learning algorithm that is tailored specifically to our setting and operates under either semi-bandit or bandit feedback.

4 Optimistic Online Learning Algorithms

In this section, we devise **UCB-NS** (Algorithm 1), an online learning algorithm that obtains a sublinear Nash regret that is also polynomial in the number of agents n under either semi-bandit or bandit feedback. Following the principle of *"optimism in the face of uncertainty"*, stating that one should act as if the environment is as nice as *plausibly possible*, UCB-NS optimistically estimates the agents' preferences using upper confidence bounds (UCBs), a well-known approach underlying many bandit algorithms (Lattimore and Szepesvári 2020). First, we describe our algorithm's main ingredients under either bandit feedback (Section 4.1) or semi-bandit feedback (Section 4.2), which mainly differ in the employed UCBs. Under both settings, we show that the agents' UCB estimates induce a potential game at each round. We thus update the agents' strategies to be an ε-NS strategy of this game, which we prove can be found in a polynomial number of steps by a natural dynamics induced by approximate Nash deviations. We then prove that UCB-NS has Nash regret and sample complexity bounds that are ***optimal*** up to logarithmic factors under both semi-bandit and bandit feedback (Section 4.3).

Algorithm 2: ε-**BRD**

Input: A game $\mathcal{G}^t = (N, (u_i^t)_{i \in N})$; $\varepsilon \in (0, 1]$.

1: Initialize an arbitrary joint pure strategy $\varphi^1 = \mathbf{x} \in [n]^n$.
2: **for each** step $s = 1, 2, 3, \ldots$ **do**
3: Pick $y_i^s \in \arg\max_{z_i \in [n]} u_i^t(\varphi_{-i}^s, z_i) - u_i^t(\varphi^s)\ \forall i \in N$ and an agent $j \in \arg\max_{i \in [n]} u_i^t(\varphi_{-i}^s, y_i^s) - u_i^t(\varphi^s)$.
4: Update $\varphi_j^{s+1} = y_j$, $\varphi_i^{s+1} = \varphi_i^s\ \forall i \neq j$.
5: **if** φ^{s+1} is an ε-NS strategy **then return** φ^{s+1}

4.1 Bandit Feedback

Under *bandit* feedback, each agent i cannot estimate her mean utility from any other agent $i \neq j \in N$. Instead, when running UCB-NS (Algorithm 1) under *bandit feedback*, she uses her feedback to *directly* construct her UCB estimates for each candidate coalition. Formally, each agent i first initializes an arbitrary mixed strategy $\varphi_i^1 \in \Delta([n])$ (line 1). At any time t, each agent i first samples her chosen candidate coalition $x_i^t \in [n]$ from her strategy φ_i^t (line 4), establishing a joint assignment $\mathbf{x}^t = (x_i^t)_{i \in N}$. After the partition $\pi^t := \pi^{\mathbf{x}^t}$ is formed, agent i obtains a utility feedback v_i^t from her coalition $\pi^t(i)$ (lines 5-6). At time t, agent i then uses those feedbacks to estimate her mean utility from each candidate coalition $\ell \in [n]$ by her empirical average of the utilities she received (if any) from that coalition until time t. Namely, for any joint assignment $\mathbf{x} \in [n]^n$, let $M_i^t(\mathbf{x}) = \sum_{\tau=1}^t \mathbb{1}\{\pi^\tau(i) = \pi^{\mathbf{x}}(i)\}$ be the number of times agent i's coalition was $\pi^{\mathbf{x}}(i)$ up to time t, where, for any time $\tau \in [t]$, $\mathbb{1}\{\pi^\tau(i) = \pi^{\mathbf{x}}(i)\}$ equals to 1 if $\pi^\tau(i) = \pi^{\mathbf{x}}(i)$ and 0 otherwise. As such, agent i can estimate her utility from the partition induced by any possible joint assignment $\mathbf{x} = (x_i)_{i \in N} \in [n]^n$ via her empirical mean utility from the x_ith candidate coalition until time t:

$$\bar{v}_i^t(\mathbf{x}) = \frac{\sum_{\tau=1}^t v_i^\tau \mathbb{1}\{\pi^\tau(i) = \pi^{\mathbf{x}}(i)\}}{M_i^t(\mathbf{x}) \vee 1} \tag{2}$$

Every agent then explores a joint assignment $\mathbf{x} \in [n]^n$ more often if it is either promising or not explored enough. Hence, any agent i constructs her UCB estimate for any joint assignment $\mathbf{x}$ based on two terms (line 7): (1) the estimated utility $\bar{v}_i^t(\mathbf{x})$ that captures the exploitation aspect, and (2) an exploration bonus that decreases with the increase in $M_i^t(\mathbf{x})$. Formally, at any time t, agent i's UCB estimate of any joint assignment $\mathbf{x}$ is $\bar{u}_i^t(\mathbf{x}) = \bar{v}_i^t(\mathbf{x}) + \bar{b}_i^t(\mathbf{x})$, where the exploration bonus $\bar{b}_i^t(\mathbf{x})$ is given by:

$$\bar{b}_i^t(\mathbf{x}) := \sqrt{\frac{n^3 \log(4(n^2+1)T/\delta)}{M_i^t(\mathbf{x}) \vee 1}} \tag{3}$$

for some confidence level $\delta \in (0, 1]$ that captures the degree of certainty. As we will see in Section 4.3, we carefully designed (3) based on Hoeffding's inequality to ensure a Nash regret of $\tilde{\mathcal{O}}(\sqrt{T})$. Unlike the standard analysis of UCB, we next prove that the UCB estimates induce a potential game:

Lemma 2. *At any time t, let $\bar{\mathcal{G}}^t$ be the game with agent set N and each agent i's utility from any joint assignment $\mathbf{x} \in [n]^n$ is $\bar{u}_i^t(\mathbf{x}) = \bar{v}_i^t(\mathbf{x}) + \bar{b}_i^t(\mathbf{x})$. Then, $\bar{\mathcal{G}}^t$ is a* potential game.

Proof. As in Lemma 1, note that $\bar{\Psi}^t(\mathbf{x}) = \sum_{i \in N} \bar{u}_i^t(\mathbf{x})$ is a potential function for *pure* strategies in $\bar{\mathcal{G}}^t$ as $\bar{\Psi}^t(\mathbf{x}_{-i}, x_i) - \bar{\Psi}^t(\mathbf{x}_{-i}, x_i') = \bar{u}_i^t(\mathbf{x}_{-i}, x_i) - \bar{u}_i^t(\mathbf{x}_{-i}, x_i')$ for any agent i and another assignment $x_i' \in [n]$ of agent i. □

Exploiting this result, we then use Algorithm 2 to *efficiently* update the agents' joint strategy at time $t+1$ to be an ε-NS strategy of the game $\bar{\mathcal{G}}^t$ from Lemma 3 (lines 11-12). The idea behind Algorithm 2 is simple: given *any* such game $\mathcal{G}^t$, if a joint assignment $\mathbf{x}$ is *not* an ε-NS of $\mathcal{G}^t$, then there exists an agent i and a strategy x_i' of i that improves her utility (see line 3 in Algorithm 2). Such a deviation is called an *improving move*. This leads to Algorithm 2, which follows the natural process of finding approximate equilibria in games termed as *ε-better-response dynamics* (ε-BRD): starting from an arbitrary joint strategy, repeatedly perform improving moves. When no such move exists anymore, an ε-NS strategy has been reached (line 5), which may be *mixed* as the initial joint strategy is mixed. We prove that ε-BRD terminates for any input as it always converges to an ε-NS strategy in a polynomial number of steps:

Theorem 1. *For any game $\mathcal{G}^t = (N, (u_i^t)_{i \in N})$ with agent set N and each agent i's utility from any joint assignment $\mathbf{x} \in [n]^n$ is $u_i^t(\mathbf{x})$, for any $\varepsilon \in (0, 1]$ and for any initial pure strategy φ^1, ε-BRD converges to an ε-NS strategy in at most $s^\star := \lceil \frac{nW}{\varepsilon} \rceil$ steps, where $W = \max_{i \in N, \mathbf{x} \in [n]^n} |u_i^t(\mathbf{x})|$.*

Proof. (*Sketch*) Any game $\mathcal{G}^t$ admits the potential function $\bar{\Psi}^t$ specified in the proof of Lemma 3. In Appendix A, we show that $\bar{\Psi}^t(\varphi) - \bar{\Psi}^t(\varphi') \leq nW$ for any pair joint pure strategies $\varphi, \varphi' \in [n]^n$. Then, we assume by contradiction that the agents' joint strategy φ^s at any step $1 \leq s \leq s^\star$ is *not* an ε-NS strategy. Thus, this also holds at step $s^\star$, which we prove to yield $\bar{\Psi}^t(\varphi^{s^\star+1}) - \bar{\Psi}^t(\varphi^1) > nW$, contradicting our previous observation. □

Remark 1. *As formulated in Appendix B, symmetric ASHGs satisfy the conditions of Theorem 1 due to Lemma 1. Hence, Theorem 1 has major implications for deviation dynamics in symmetric ASHGs. As mentioned in Section 2, computing an* exact *NS partition is PLS-complete (Gairing and Savani 2019). Brandt et al. (2024) show that this yields that the better-response dynamics based on Nash deviations in symmetric ASHGs may converge to an* exact *NS partition only after an* exponential *number of steps. Yet, Theorem 1 indicates that this can be amended: the* approximate *better-response dynamics always converges to an* approximate *NS partition after* polynomially *many steps*

4.2 Semi-Bandit Feedback

Unlike the *bandit* feedback setting, by executing UCB-NS (Algorithm 1) under *semi-bandit* feedback, each agent exploits her feedbacks to approximate her mean utility from any other agent, which is then used to construct her UCB estimates for each candidate coalition. Formally, each agent i first initializes an arbitrary mixed strategy $\varphi_i^1 \in \Delta([n])$ (line 1). At any time t, each agent i first samples her chosen candidate coalition $x_i^t \in [n]$ from her strategy φ_i^t (line 4), establishing a joint assignment $\mathbf{x}^t = (x_i^t)_{i \in N}$. After forming the partition $\pi^t := \pi^{\mathbf{x}^t}$, agent i obtains a utility feedback $v_{i,j}^t$ from each agent $i \neq j \in \pi^t(i)$ (lines 8-9). Agent i then uses those feedbacks to estimate her mean utility from each other agent $i \neq j \in N$ by her empirical average of the utilities she received (if any) from agent j until time t. Namely, let $N_{i,j}^t = \sum_{\tau=1}^t \mathbb{1}\{x_j^\tau = x_i^\tau\}$ be the number of times any pair of agents i, j joined the same coalition up to time t, where, for any $\ell, \ell' \in [n]$, $\mathbb{1}\{\ell = \ell'\}$ equals to 1 if $\ell = \ell'$ and 0 otherwise. Thus, agent i's empirical mean utility from another agent $i \neq j \in N$ up to time t is $\hat{v}_i^t(j) = \frac{\sum_{\tau=1}^t v_{i,j}^\tau \mathbb{1}\{x_j^\tau = x_i^\tau\}}{N_{i,j}^t \vee 1}$. Afterwards, agent i can estimate her utility from the partition induced by any possible joint assignment $\mathbf{x} \in [n]^n$ via:

$$\hat{v}_i^t(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i)} \hat{v}_i^t(j) \tag{4}$$

Every agent then explores a joint assignment $\mathbf{x} \in [n]^n$ more often if it is either promising or not explored enough. Similarly to Section 4.1, at any time t, agent i's UCB estimate of any joint assignment $\mathbf{x}$ is $\hat{u}_i^t(\mathbf{x}) = \hat{v}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$, while the exploration bonus $\hat{b}_i^t(\mathbf{x})$ is defined similarly to (3):

$$\hat{b}_i^t(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i)} \sqrt{\frac{2n \log(4(n^2+1)T/\delta)}{N_{i,j}^t \vee 1}} \tag{5}$$

where $\delta \in (0, 1]$ is a confidence level that captures the degree of certainty. Unlike the classic analysis of UCB, the UCB estimates induce a potential game, as in Section 4.1:

Lemma 3. *At any time t, let $\hat{\mathcal{G}}^t$ be the game with agent set N and any agent i's utility from any joint assignment $\mathbf{x} \in [n]^n$ is $\hat{u}_i^t(\mathbf{x}) = \hat{v}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$. Then, $\hat{\mathcal{G}}^t$ is a* potential game.

Proof. As in Lemma 1, note that $\hat{\Psi}^t(\mathbf{x}) = \sum_{i \in N} \hat{u}_i^t(\mathbf{x})$ is a potential function for *pure* strategies in $\hat{\mathcal{G}}^t$ as $\hat{\Psi}^t(\mathbf{x}_{-i}, x_i) - \hat{\Psi}^t(\mathbf{x}_{-i}, x_i') = \hat{u}_i^t(\mathbf{x}_{-i}, x_i) - \hat{u}_i^t(\mathbf{x}_{-i}, x_i')$ for any agent i and another assignment $x_i' \in [n]$ of agent i. □

By Theorem 1, we obtain the following corollary:

Corollary 1. *Under* semi-bandit *feedback, we can use Algorithm 2 to* efficiently *update the agents' joint strategy at time $t+1$ to be an ε-NS strategy of the game $\hat{\mathcal{G}}^t$ from Lemma 3 in at most $\lceil \frac{nW}{\varepsilon} \rceil$ steps (lines 11 and 13), where $W = \max_{i \in N, \mathbf{x} \in [n]^n} |\hat{u}_i^t(\mathbf{x})|$.*

4.3 Analysis of Nash Regret

To establish the Nash regret bounds of Algorithm 1, we derive an upper bound on each agent's estimation error for her mean utility $d_i(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i)} d_{i,j}$ from the partition induced by any joint assignment $\mathbf{x}$ at any time t:

Lemma 4. *For any $\delta \in (0, 1]$, simultaneously for any time t, for each agent i and any joint assignment $\mathbf{x} \in [n]^n$, the estimation error is bounded as follows with probability at least $1 - \delta$ under* bandit *and* semi-bandit *feedback (resp.):*

$$|\bar{v}_i^t(\mathbf{x}) - d_i(\mathbf{x})| \leq \sqrt{\frac{n^3 \log(2n^2T/\delta)}{2M_i^t(\mathbf{x}) \vee 1}} \tag{6}$$

$$|\hat{v}_i^t(\mathbf{x}) - d_i(\mathbf{x})| \leq \sum_{i \neq j \in \pi^{\mathbf{x}}(i) : N_{i,j}^t \geq 1} \sqrt{\frac{2n \log(2n^2T/\delta)}{N_{i,j}^t \vee 1}} \tag{7}$$

which differ due to the the granularity of utility information observed by every agent under each feedback model.

Proof. (*Sketch*) We prove for semi-bandit feedback; the proof for bandit feedback is similar, and thus deferred to Appendix C.2. If $N_{i,j}^t = 0$ for any agent $i \neq j \in \pi^{\mathbf{x}}(i)$, then our statement clearly holds as $\hat{v}_i^t(\mathbf{x}) = 0$. Hereafter, we thus assume that $N_{i,j}^t(\mathbf{x}) \geq 1$ for at least one agent $i \neq j \in \pi^{\mathbf{x}}(i)$. Note that (4) can be rephrased as $\hat{v}_i^t(\mathbf{x}) = \sum_{i \neq j \in \pi^{\mathbf{x}}(i): N_{i,j}^t(\mathbf{x}) \geq 1} \hat{v}_i^t(j)$. In Appendix C.1, we apply Hoeffding's inequality to agent i's utility estimator $\hat{v}_i^t(j)$ for each other agent $i \neq j \in \pi^{\mathbf{x}}(i)$ to obtain a probabilistic upper bound on $|\hat{v}_i^t(j) - d_{i,j}|$. By summing those bounds and applying a union bound, we then derive (7) as $|\hat{v}_i^t(\mathbf{x}) - d_i(\mathbf{x})| \leq \sum_{i \neq j \in \pi^{\mathbf{x}}(i): N_{i,j}^t \geq 1} |\hat{v}_i^t(j) - d_{i,j}|$. □

We are now ready to prove the Nash regret bounds of Algorithm 1 under either semi-bandit or bandit feedback:

Theorem 2. *For any $\delta \in (0, 1]$, Algorithm 1 with $\varepsilon = \frac{1}{T}$ obtains the following Nash regret bound with probability at least $1 - \delta$ under* bandit *and* semi-bandit *feedback (resp.):*

$$\mathcal{R}^T \leq \mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)}(\sqrt{n^3} + 1)) \text{ [Bandit]} \quad (8)$$

$$\mathcal{R}^T \leq \mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)}) \text{ [Semi-Bandit]} \quad (9)$$

Proof. We supply the proof for semi-bandit feedback; the proof for bandit feedback is similar, and thus deferred to Appendix F. By Theorem 1, φ^t is an ε-NS strategy of the game $\hat{\mathcal{G}}^t$ from Lemma 3. Thus, $\hat{U}_i^t := \mathbb{E}_{\mathbf{x} \sim \varphi^t}[\hat{u}_i^t(\mathbf{x})]$ satisfies:

$$\hat{U}_i^t = \max_{\phi_i \in \Delta([n])} \mathbb{E}_{\mathbf{x} \sim (\varphi_{-i}^t, \phi_i)}[\hat{u}_i^t(\mathbf{x})] - \varepsilon \quad (10)$$

Recall that agent i's utility from her best response to the other agents' strategies at time t is then $V_i^\star(\varphi_{-i}^t) = \max_{\phi_i \in \Delta([n])} \mathbb{E}_{\mathbf{x} \sim (\varphi_{-i}^t, \phi_i)}[d_i(\mathbf{x})]$. Letting $\hat{w}_i^t(\mathbf{x}) := \hat{v}_i^t(\mathbf{x}) - \bar{b}_i^t(\mathbf{x})$ and $\hat{W}_i^t := \mathbb{E}_{\mathbf{x} \sim \varphi^t}[\hat{w}_i^t(\mathbf{x})]$, in Appendix D we combine (10) with Lemma 4 to show that $V_i^\star(\varphi_{-i}^t) \leq \hat{U}_i^t + \varepsilon$ and $V_i(\varphi^t) \geq \hat{W}_i^t$ with probability at least $1 - \delta$ concurrently for any time t and agent i, yielding:

$$V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t) \leq \hat{U}_i^t - \hat{W}_i^t + \varepsilon \quad (11)$$

Unlike the classic UCB analysis, (11) indicates that updating the agents' strategies via Algorithm 2 allows us to upper bound each agent's local gap in each round. Next, note that:

$$\max_{i \in N}[\hat{u}_i^t(\mathbf{x}) - \hat{w}_i^t(\mathbf{x})] \leq 2 \max_{i \in N} \hat{b}_i^t(\mathbf{x}) =: \hat{B}^t(\mathbf{x}) \quad (12)$$

from which we infer that $\max_{i \in N}[\hat{U}_i^t - \hat{W}_i^t] \leq \mathbb{E}_{\mathbf{x} \sim \varphi^t}[\max_{i \in N}[\hat{u}_i^t(\mathbf{x}) - \hat{w}_i^t(\mathbf{x})]]$, that is:

$$\max_{i \in N}[\hat{U}_i^t - \hat{W}_i^t] \leq \mathbb{E}_{\mathbf{x} \sim \varphi^t}[\hat{B}^t(\mathbf{x})] =: \hat{B}^t \quad (13)$$

Let $\mathcal{F}^t$ be the σ-algebra generated by the history up to time $t-1$. Denoting $M^t = \hat{B}^t - \hat{B}^t(\mathbf{x}^t)$, where $\mathbf{x}^t$ is the agents' joint assignment at time t, note that $(M^t)_{i=1}^T$ is a martingale difference sequence with respect to $\mathcal{F}^t$ since $\mathbb{E}[M^t | \mathcal{F}^t] = 0$. As $\hat{b}_i^t(\mathbf{x}) \leq \mathcal{O}(\sqrt{n^3 \log(n^2 T/\delta)})$ by (3), Azuma's inequality implies that, with probability at least $1 - \delta$, it holds that:

$$\sum_{t \in [T]} M^t = \mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)}) \quad (14)$$

We are now ready to bound our UCB-NS's Nash regret. As agent i's utility from any strategy is always at most n, note that $\mathcal{R}^T = \sum_{t=1}^T \max_{i \in N}(V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t)) = \sum_{t=1}^T \min\{\max_{i \in N}(V_i^\star(\varphi_{-i}^t) - V_i(\varphi^t)), n\}$. Thus, by (11), $\mathcal{R}^T \leq T\varepsilon + \sum_{t=1}^T \min\{\max_{i \in N}(\hat{U}_i^t - \hat{W}_i^t), n\}$ with probability at least $1 - \delta$. Due to (13) and $\varepsilon = 1/T$, we have $\mathcal{R}^T \leq 1 + \sum_{t=1}^T \min\{\hat{B}^t, n\}$. By the definition of M^t, then $\mathcal{R}^T \leq 1 + \sum_{t=1}^T [\min\{\hat{B}^t(\mathbf{x}^t), n\} + M^t]$. By (12) and (14), the Nash regret $\mathcal{R}^T$ is upper bounded by:

$$\mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)}) + 2 \sum_{t=1}^T \min\left\{ \max_{i \in N} \hat{b}_i^t(\mathbf{x}^t), n \right\} \quad (15)$$

In Appendix E, we show that the sum in (15) is at most $\mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)})$, giving our Nash regret bound. □

Remark 2. *Under either semi-bandit or bandit feedback, Theorem 2 indicates that Algorithm 1 achieves a Nash regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$. By standard online-to-batch conversion (see, e.g., (Jin et al. 2018, Section 3.1)), our Nash regret bound suggests a sample complexity bound of $\mathcal{O}(1/\varepsilon^2)$ for obtaining an ε-NS strategy. Surprisingly, though our method obtains less information under bandit feedback compared to semi-bandit feedback, our Nash regret and sample complexity bounds for both feedback models are* **optimal** *up to logarithmic factors (Hassani et al. 2020; Bai and Jin 2020).*

5 Conclusions and Future Work

In this paper, we presented a new algorithmic framework for studying online learning in coalition formation from a *non-cooperative* perspective and examined the impact of collective data utilization on online learning algorithms, where selfish agents exploit a shared data platform to improve their learning. Our goal was designing sample-efficient algorithms for self-interested agents that minimize their Nash regret under either semi-bandit or bandit feedback, yielding approximately Nash stable partitions. After proving that our model is a sequence of potential games, we presented a UCB-based method whose Nash regret and sample complexity bounds are *polynomial* in the number of agents under either semi-bandit or bandit feedback, and ***optimal*** up to logarithmic factors under both feedback models.

Our research opens the way for many future works. Immediate directions are exploring other classes of hedonic games and other solution concepts (e.g., fairness). In many realistic scenarios, agents cannot enhance their learning by leveraging a shared data platform. Thus, future research should also develop *distributed* algorithms with Nash regret and sample complexity bounds that are optimal up to logarithmic factors, despite the additional lack of knowledge. Further, future studies warrant studying other models of *partial* and possibly *noisy* information. Lastly, another intriguing future direction is devising algorithms with no-regret guarantees to any agent adopting them, i.e., each agent has diminishing regret, regardless of how others update their strategies.

Acknowledgments

This research was funded in part by ISF grant 1563/22.

References

Aziz, H.; Brandl, F.; Brandt, F.; Harrenstein, P.; Olsen, M.; and Peters, D. 2019. Fractional hedonic games. *ACM Transactions on Economics and Computation (TEAC)*, 7(2): 1–29.

Aziz, H.; Brandt, F.; and Seedig, H. G. 2011. Stable partitions in additively separable hedonic games. In *AAMAS*, volume 11, 183–190.

Aziz, H.; and Savani, R. 2016. Hedonic Games. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*, 356–376. Cambridge University Press.

Bai, Y.; and Jin, C. 2020. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 551–560. PMLR.

Ballester, C. 2004. NP-completeness in hedonic games. *Games and Economic Behavior*, 49(1): 1–30.

Banerjee, S.; Konishi, H.; and Sönmez, T. 2001. Core in a simple coalition formation game. *Social Choice and Welfare*, 18(1): 135–153.

Bilò, V.; Fanelli, A.; Flammini, M.; Monaco, G.; and Moscardelli, L. 2018. Nash stable outcomes in fractional hedonic games: Existence, efficiency and computation. *Journal of Artificial Intelligence Research*, 62: 315–371.

Boehmer, N.; Bullinger, M.; and Kerkmann, A. M. 2023. Causes of stability in dynamic coalition formation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 5499–5506.

Bogomolnaia, A.; and Jackson, M. O. 2002. The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2): 201–230.

Bouveret, S.; Endriss, U.; Lang, J.; et al. 2010. Fair division under ordinal preferences: Computing envy-free allocations of indivisible goods. In *ECAI*, 387–392.

Brandt, F.; Bullinger, M.; and Tappe, L. 2022. Single-agent dynamics in additively separable hedonic games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5): 4867–4874.

Brandt, F.; Bullinger, M.; and Tappe, L. 2024. Stability Based on Single-Agent Deviations in Additively Separable Hedonic Games. *Artificial Intelligence*, 334: 104160.

Bullinger, M.; and Romen, R. 2023. Online Coalition Formation Under Random Arrival or Coalition Dissolution. In *31st Annual European Symposium on Algorithms (ESA 2023)*, volume 274 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 27:1–27:18.

Bullinger, M.; and Romen, R. 2024. Stability in Online Coalition Formation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9): 9537–9545.

Carosi, R.; Monaco, G.; and Moscardelli, L. 2019. Local core stability in simple symmetric fractional hedonic games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 574–582.

Chen, X.; and Peng, B. 2020. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33: 18990–18999.

Cohen, S.; and Agmon, N. 2023a. Complexity of probabilistic inference in random dichotomous hedonic games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 5573–5581.

Cohen, S.; and Agmon, N. 2023b. Online Coalitional Skill Formation. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 494–503.

Cohen, S.; and Agmon, N. 2024a. Online Friends Partitioning Under Uncertainty. In *ECAI 2024*, 3332–3339. IOS Press.

Cohen, S.; and Agmon, N. 2024b. Online Learning of Partitions in Additively Separable Hedonic Games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2722–2730.

Cohen, S.; and Agmon, N. 2025. Online Learning of Coalition Structures by Selfish Agents – Supplementary Materials. https://u.cs.biu.ac.il/~agmon/CohenAAAI25Sup.pdf.

Cui, Q.; Xiong, Z.; Fazel, M.; and Du, S. S. 2022. Learning in congestion games with bandit feedback. *Advances in Neural Information Processing Systems*, 35: 11009–11022.

Daskalakis, C.; Fishelson, M.; and Golowich, N. 2021. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34: 27604–27616.

Daskalakis, C.; and Pan, Q. 2014. A counter-example to Karlin's strong conjecture for fictitious play. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 11–20. IEEE.

Ding, D.; Wei, C.-Y.; Zhang, K.; and Jovanovic, M. 2022. Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 5166–5220.

Dreze, J. H.; and Greenberg, J. 1980. Hedonic coalitions: Optimality and stability. *Econometrica: Journal of the Econometric Society*, 987–1003.

Eichhorn, M.; Banerjee, S.; and Kempe, D. 2022. Online team formation under different synergies. In *International Conference on Web and Internet Economics*, 78–95. Springer.

Elkind, E.; and Wooldridge, M. J. 2009. Hedonic coalition nets. In *AAMAS (1)*, 417–424. Citeseer.

Fanelli, A.; Monaco, G.; and Moscardelli, L. 2021. Relaxed Core Stability in Fractional Hedonic Games. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 182–188.

Fioravanti, S.; Flammini, M.; Kodric, B.; and Varricchio, G. 2023. PAC learning and stabilizing Hedonic Games: towards a unifying approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 5641–5648.

Flammini, M.; Kodric, B.; Monaco, G.; and Zhang, Q. 2021a. Strategyproof mechanisms for additively separable and fractional hedonic games. *Journal of Artificial Intelligence Research*, 70: 1253–1279.

Flammini, M.; Monaco, G.; Moscardelli, L.; Shalom, M.; and Zaks, S. 2021b. On the Online Coalition Structure Generation Problem. *Journal of Artificial Intelligence Research*, 72: 1215–1250.

Gairing, M.; and Savani, R. 2019. Computing stable outcomes in symmetric additively separable hedonic games. *Mathematics of Operations Research*, 44(3): 1101–1121.

Hassani, H.; Karbasi, A.; Mokhtari, A.; and Shen, Z. 2020. Stochastic conditional gradient++: (non)convex minimization and continuous submodular maximization. *SIAM Journal on Optimization*, 30(4): 3315–3344.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31: 4863–4873.

Johnson, D. S.; Papadimitriou, C. H.; and Yannakakis, M. 1988. How easy is local search? *Journal of computer and system sciences*, 37(1): 79–100.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Leonardos, S.; Overman, W.; Panageas, I.; and Piliouras, G. 2022. Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games. In *International Conference on Learning Representations*.

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

Liu, L. T.; Mania, H.; and Jordan, M. 2020. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, 1618–1628. PMLR.

Liu, L. T.; Ruan, F.; Mania, H.; and Jordan, M. I. 2021a. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211): 1–34.

Liu, Q.; Yu, T.; Bai, Y.; and Jin, C. 2021b. A Sharp Analysis of Model-based Reinforcement Learning with Self-Play. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7001–7010.

Maheshwari, C.; Sastry, S.; and Mazumdar, E. 2022. Decentralized, communication-and coordination-free learning in structured matching markets. *Advances in Neural Information Processing Systems*, 35: 15081–15092.

Panageas, I.; Skoulakis, S.; Viano, L.; Wang, X.; and Cevher, V. 2023. Semi Bandit Dynamics in Congestion Games: Convergence to Nash Equilibrium and No-Regret Guarantees. In *International Conference on Machine Learning*, 26904–26930. PMLR.

Rodet, M.; and Gaudel, R. 2022. Unimodal Mono-Partite Matching in a Bandit Setting. In *Complex Feedback in Online Learning Workshop at the 39th International Conference on Machine Learning*.

Sankararaman, A.; Basu, S.; and Sankararaman, K. A. 2021. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, 1252–1260. PMLR.

Sentenac, F.; Yi, J.; Calauzenes, C.; Perchet, V.; and Vojnovic, M. 2021. Pure exploration and regret minimization in matching bandits. In *International Conference on Machine Learning*, 9434–9442. PMLR.

Sliwinski, J.; and Zick, Y. 2017. Learning Hedonic Games. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17*, 2730–2736.

Sung, S.-C.; and Dimitrov, D. 2010. Computational complexity in additive hedonic games. *European Journal of Operational Research*, 203(3): 635–639.

Tekin, C.; and Van Der Schaar, M. 2015. Distributed online learning via cooperative contextual bandits. *IEEE transactions on signal processing*, 63(14): 3700–3714.

Woeginger, G. J. 2013. Core Stability in Hedonic Coalition Formation. In *SOFSEM 2013: Theory and Practice of Computer Science*, 33–50. Berlin, Heidelberg: Springer.

Zhang, Y.; Wang, S.; and Fang, Z. 2022. Matching in Multi-arm Bandit with Collision. *Advances in Neural Information Processing Systems*, 35: 9552–9563.